

# Concepts and Autonomous Agents

Paul Davidsson

LU-CS-TR: 94-124



Department of Computer Science, Lund University

# Preface

This is a thesis for the degree of Licenciat (a Swedish degree between M.Sc. and Ph.D.) The thesis consists of four parts, where the first part is a comprehensive introduction to the general topic under study. The first part can be considered as the main work whereas the other parts are smaller in-depth studies of particular topics.

Part I is essentially a survey paper and is only published in this thesis. It is an effort to pull together different lines of argumentation within cognitive science, philosophy, and artificial intelligence in order to establish a solid foundation for further research. Some of the material in this part originates from:

P. Davidsson. *Concept Acquisition by Autonomous Agents: Cognitive Modeling versus the Engineering Approach*. Lund University Cognitive Studies 12, ISSN 1101-8453, Lund University, Sweden, 1992.

Part II is to be published as:

P. Davidsson, E. Astor, and B. Ekdahl. *A Framework for Autonomous Agents Based on the Concept of Anticipatory Systems*. In *Twelfth European Meeting on Cybernetics and Systems Research*, Vienna, Austria, World Scientific Publishing Co., 1994.

Part III has been published as:

P. Davidsson. *Toward a General Solution to the Symbol Grounding Problem: Combining Machine Learning and Computer Vision*. In *AAAI Fall Symposium Series, Machine Learning in Computer Vision: What, Why and How?* (FS-93-04), pages 157-161, Raleigh, North Carolina, AAAI Press, 1993.

Part IV has been published as:

P. Davidsson. *A Framework for Organization and Representation of Concept Knowledge in Autonomous Agents*. In *Scandinavian Conference of Artificial Intelligence - 93*, pages 183-192, Stockholm, Sweden, IOS Press, 1993.

## Acknowledgements

First of all, I would like to thank my advisor Dr. Eric Astor for his invaluable help and encouragement during the work described in this thesis, Bertil Ekdahl and Robert Pallbo (who together with Eric and I constitute the AI-group) for their willingness to

discuss and criticize my work from different perspectives, and everyone else at the Department of Computer Science for providing a pleasant research environment.

I also want to thank everybody at the Department of Cognitive Science, led by Prof. Peter Gärdenfors, for always providing inspiring discussions on a seemingly unlimited range of topics, and especially Christian Balkenius for reviewing an earlier draft of this thesis.

Furthermore, I would like to thank Olof Tilly at the Department of Philosophy for his detailed comments on content and style of the last draft of this thesis, Prof. Lars Löfgren at the Department of Systems Theory for stimulating discussions in the beginning of my graduate studies, and Prof. Rune Gustavsson at SICS (Swedish Institute of Computer Science) and the Department of Economics and Computer Science, University of Karlskrona/Ronneby for showing interest in my graduate work.

Finally, I want to thank Tanja for love and patience, and my parents for their continuous support.

# Contents

<b>Part I Main Paper</b>	<b>1</b>
<b>1 Introduction</b>	<b>3</b>
1.1 Approach . . . . .	3
1.2 Outline . . . . .	5
<b>2 Autonomous Agents</b>	<b>7</b>
2.1 Introduction . . . . .	7
2.1.1 Different Kinds of Autonomous Agents . . . . .	8
2.1.2 Requirements . . . . .	9
2.1.3 Scenarios for an Autonomous Agent . . . . .	10
2.1.4 Possible Applications . . . . .	10
2.1.5 Two approaches . . . . .	11
2.2 Deliberative Agents . . . . .	11
2.2.1 Limitations of the Approach . . . . .	12
2.3 Reactive Agents . . . . .	12
2.3.1 The Work of Brooks . . . . .	13
2.3.2 Limitations of the Approach . . . . .	14
2.4 Combining the Approaches . . . . .	14
2.4.1 Anticipatory Autonomous Agents . . . . .	14
2.5 Conclusions . . . . .	15
<b>3 World Modeling</b>	<b>17</b>
3.1 What is a World Model? . . . . .	17
3.1.1 Terminology . . . . .	18
3.1.2 Explicit and Implicit Knowledge . . . . .	18
3.2 Is Explicit Knowledge about the World Necessary? . . . . .	18
3.2.1 On the Proper Level of Abstraction . . . . .	19
3.3 Is It Possible to Acquire Explicit Knowledge about the World? . . . . .	19
3.4 World Modeling within AI . . . . .	20
3.4.1 Computer Vision . . . . .	20
3.4.2 Machine Learning . . . . .	22
3.4.3 Computer Vision – Machine Learning Relation . . . . .	22
3.5 Conclusions . . . . .	24
<b>4 Concepts and Categories</b>	<b>27</b>
4.1 Terminology . . . . .	27
4.2 What does it mean to have a concept? . . . . .	29

<b>5</b>	<b>The Functions of Concepts</b>	<b>31</b>
5.1	The Functions of Human Concepts . . . . .	33
5.2	Functions of Concepts in Artificial Autonomous Agents . . . . .	35
<b>6</b>	<b>The Nature of Categories</b>	<b>37</b>
6.1	The Nature of Human Categories . . . . .	37
6.1.1	Properties . . . . .	38
6.1.2	Natural Kinds . . . . .	39
6.1.3	Similarity . . . . .	40
6.1.4	Derived Categories . . . . .	42
6.1.5	Artifact Categories . . . . .	42
6.1.6	Taxonomies . . . . .	43
6.2	The Nature of AI Categories . . . . .	45
6.2.1	Properties . . . . .	45
6.2.2	Similarity . . . . .	45
6.2.3	Taxonomies . . . . .	46
6.3	Conclusions . . . . .	46
6.3.1	Properties . . . . .	46
6.3.2	Similarity . . . . .	47
6.3.3	Taxonomies . . . . .	48
<b>7</b>	<b>Representation of Categories</b>	<b>49</b>
7.1	Human Representation of Categories . . . . .	49
7.1.1	The Classical View and It's Problems . . . . .	49
7.1.2	The Probabilistic View . . . . .	51
7.1.3	The Exemplar View . . . . .	52
7.1.4	Combining the Probabilistic and Exemplar View . . . . .	52
7.2	Representation of Categories in AI . . . . .	53
7.2.1	General AI Category Representations . . . . .	53
7.2.2	Machine Learning Category Representations . . . . .	54
7.2.3	Computer Vision Category Representations . . . . .	57
7.3	Conclusions . . . . .	58
7.3.1	Multiple Category Representations . . . . .	60
7.3.2	A Novel Framework for Composite Concepts . . . . .	61
7.3.3	Summary . . . . .	62
<b>8</b>	<b>Concept Acquisition</b>	<b>65</b>
8.1	Human Concept Acquisition . . . . .	65
8.1.1	Theories of Concept Acquisition . . . . .	66
8.2	AI Methods for Concept Acquisition . . . . .	67
8.2.1	ML Methods for Concept Acquisition . . . . .	68
8.2.2	Computer Vision Approaches . . . . .	74
8.2.3	Pattern Recognition . . . . .	75
8.3	What can be learned? . . . . .	77
8.3.1	Gold's Paradigm . . . . .	78
8.3.2	Valiant's Paradigm . . . . .	79
8.3.3	Critique of Current Formal Learning Theory . . . . .	79
8.4	Conclusions . . . . .	79

8.4.1	Incrementality . . . . .	80
8.4.2	Learning Multiple Concepts . . . . .	80
8.4.3	Fast Learning . . . . .	80
8.4.4	Integrating Different Learning Strategies . . . . .	81
8.4.5	Other Issues . . . . .	83
<b>9</b>	<b>Conclusions and Further Research</b>	<b>85</b>
9.1	Conclusions . . . . .	85
9.2	Suggestions for Further Research . . . . .	87
	<b>Bibliography</b>	<b>89</b>
<b>Part II A Framework for Autonomous Agents Based on the Concept of Anticipatory Systems</b>		<b>101</b>
1	Introduction . . . . .	103
1.1	Background . . . . .	104
2	Anticipatory Agents . . . . .	105
2.1	Computational Framework . . . . .	105
2.2	Implementational Issues . . . . .	106
2.3	Learning . . . . .	107
2.4	Perception . . . . .	108
3	Related Research . . . . .	108
3.1	Anticipation in Traditional Agents . . . . .	108
3.2	Anticipation in Control Systems . . . . .	108
3.3	Anticipation in Biological Systems . . . . .	109
4	Conclusions and Further Research . . . . .	109
<b>Part III Toward a General Solution to the Symbol Grounding Problem: Combining Machine Learning and Computer Vision</b>		<b>111</b>
1	The Symbol Grounding Problem . . . . .	113
1.1	Harnad's Solution . . . . .	113
1.2	A More General Solution . . . . .	114
1.3	Why Learning is Important . . . . .	114
2	What and How to Learn . . . . .	114
2.1	Relevant ML Paradigms . . . . .	115
2.2	Choice of Representation . . . . .	115
3	Discussion . . . . .	116
<b>Part IV A Framework for Organization and Representation of Concept Knowledge in Autonomous Agents</b>		<b>119</b>
1	Introduction . . . . .	121
2	The Functions of Human Concepts . . . . .	122
3	Functions of Concepts in Artificial Autonomous Agents . . . . .	123
4	Representation of Concepts . . . . .	124
4.1	The Classical View . . . . .	124
4.2	Non-classical Views . . . . .	125

4.3	Discussion . . . . .	126
5	The Idea of a Composite Structure . . . . .	126
6	How Should the Components be Represented? . . . . .	128
7	Conclusions . . . . .	129

**Part I**

**Main Paper**





# Chapter 1

## Introduction

The main topic of this thesis concerns the entities which “... seem to be the very stuff of which cognitions are made” [167] and “... are assumed to be the basic constituents of thought and belief” [188], namely, *concepts*.

Why study concepts in the first place? Is it not a well investigated topic within Artificial Intelligence (AI)? According to some leading researchers, it is not. Kirsh, for instance, writes [99]: “Evidence that the notion of concept is understudied in AI is easy to find.” As the present thesis examines concept formation in the context of autonomous agents<sup>1</sup> acting in real-world environments, this is even more pertinent. Further support for this belief comes from Subramanian [199] who states that: “The problem of building appropriate high-level descriptions of the world from sense-data, is an area that has received less attention.”

In the beginning of the work documented herein, the main objects of interest were the learning and formation, and to some extent the representation, of concepts. As the research proceeded, however, it became apparent that these topics could not, or at least should not, be studied without taking some more fundamental aspects of concepts into account. Examples of such aspects are the functions of concepts and the nature of the categories that the concepts represent. In contrast to most other studies of concept learning, these topics will here be given a detailed treatment.

### 1.1 Approach

There are in principle two approaches to the studying of computational intelligence (cognition):

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

Traditionally, cognitive modeling has mainly been studied within the different cognitive sciences. Most notable are *cognitive psychology*, which studies how humans deal with concepts in memory, perception, and reasoning, and *philosophy (of mind)*, where ontological and epistemological questions regarding the nature of concepts are

---

<sup>1</sup>For the time being, think of an autonomous agent as a mobile robot.

studied. However, some interesting work has also been carried out within the fields of *developmental psychology*, which deals with questions regarding how we learn and form concepts during childhood, *linguistics*, where the relation between concepts and language is studied, and *neurology*, which investigates the low-level processing of concepts in the brain. The engineering approach, on the other hand, has mainly been studied within the field of artificial intelligence. As concept learning and representation indeed are central parts of the general problem of computational intelligence, this distinction applies to these topics as well.

While the engineering approach has sometimes been successful in learning and forming artificial concepts in restricted domains, its success has been limited in more realistic scenarios. One such scenario, probably the most natural, general, and realistic, concerns a concept-learning *autonomous agent* acting in a real-world environment. In this part of the thesis, it is exactly in this scenario that the different aspects of concepts will be studied.

We will here basically adhere to the engineering approach. However, since humans obviously are autonomous agents capable of acquiring and using concepts when interacting with the real world in a far more successful way than current AI systems, it may be a good idea to become inspired by the research on cognitive modeling. When adopting such a mixed approach, there are some things one must keep in mind. The most important is perhaps that the task of creating an autonomous agent is not equivalent with cognitive modeling. For instance, if there are no advantages of mimicking a particular feature of human concept acquisition then we have no reason to do so (i.e., there are no reasons for assuming that humans are optimal agents). On the other hand, a cognitive model does not have to be a true model of human cognition to be useful in AI. As a consequence, we will not argue about the biological and psychological plausibility of the cognitive models presented. Moreover, it is important to remember that we only have very limited insight into the ways in which human cognition actually works. Another problem by adopting this approach is that in experimental psychology, the theories of cognitive processes are often not described in detail enough to be implemented on a computer.

The primary goal of this part of the thesis is to pull together different lines of argumentation within the cognitive sciences and artificial intelligence in order to establish a solid foundation for further research into representation, learning, and formation of concepts by autonomous agents. The ambition has been to do this in an unbiased fashion without any preconceptions concerning the actual implementation (i.e., without assuming a symbolic, connectionist, or hybrid system). In addition to being a survey of the research within these fields, this will also result in some original hypotheses concerning different aspects of concepts in the context of autonomous agents. Thus, the results, or conclusions, of the survey will be in terms of new insights and ideas rather than new algorithms or methods. Moreover, since no adequate formalism exists, it is not possible to prove formally any of these conclusions. Rather, it is the author's opinion that at the present stage of research any attempt at formalization would be premature.<sup>2</sup> This will hopefully become apparent to the reader as well

---

<sup>2</sup>In fact, Herbert Simon, in an invited talk at AAAI-93, pointed out that the apparent elegance of formalisms and theorems makes the AI field oversimplify and ignore complex problems and structures that need to be studied [41]. He argued that AI belongs to the *sciences of qualitative descriptions*, such as biology and chemistry, rather than being a science that is adequately describable by mathematical formulas, such as physics (and algorithm theory).

when the formalizations that do exist (which concern concept learning) are discussed. It will, for instance, be argued that they make too strong assumptions about both the representation and the learning situation to be of any practical interest in the context of autonomous agents. In fact, many of the hypotheses suggested here will concern this kind of assumptions, and are thus of interest for the development of novel, more interesting formalizations. Thus, the arguments will be informal, and moreover they will be of a qualitative nature rather than quantitative. We have not implemented any system to support particular hypotheses. This is, however, an important issue for future research.

## 1.2 Outline

In Chapter 2 and Chapter 3 the issues of autonomous agents and world modeling are discussed. We argue against both purely reactive agents, based on stimulus-response behavior, and purely traditional agents, based on the sense-model-plan-act cycle. Instead, support is presented for a hybrid approach that makes use of both explicit world models for deliberative reasoning and stimulus-response behavior for low-level control.

Chapter 4 and Chapter 5 address some fundamental questions regarding concepts, such as: What does it mean to have a concept? and: What functions do, or should, concepts serve? By analyzing the research in cognitive science and philosophy, these questions are first answered in the context of humans. This analysis then provides a basis for a discussion of the questions in the context of artificial agents. This approach, beginning with an analysis of the cognitive modeling research on a particular topic and then letting this analysis serve as a basis for the discussion on the corresponding topic in the engineering approach, is repeated in most of the remaining chapters of this part of the thesis as well.

Chapter 6 is devoted to a discussion on what can be said about categories without taking into account how they actually are represented internally within an agent. This includes analyses of some fundamental concepts such as similarity, property, and taxonomy. Moreover, an attempt to identify different kinds of categories is made.

Chapter 7, on the other hand, treats the issue of how an agent should represent categories internally. Different suggestions from different fields are presented and evaluated according to the desirable functions identified in Chapter 4. It is concluded that none of the existing approaches is able to serve all these functions. A new approach for representing categories in autonomous agents is then presented.

Chapter 8 then, concerns the acquisition of concepts. A number of theories from various fields are presented. Some requirements that a concept learning algorithm of an autonomous agent must meet are singled out. These provide the basis for the evaluation of the existing theories. This chapter also includes a brief discussion on the question of what actually can be learned.

Finally, Chapter 9 provides a summary of the conclusions of each chapter. Some pointers for further research are also suggested.



## Chapter 2

# Autonomous Agents

An autonomous agent can be seen as a system capable of interacting independently and effectively with its environment via its own sensors and effectors in order to accomplish some given or self-generated task(s). Thus, humans and most animals can in this sense also be regarded as autonomous agents. In the following, however, we will by autonomous agents refer to artificial ones. One of the ultimate goals for AI is to construct artificial intelligent autonomous agents capable of human level performance (or better). However, a look at the state of current research reveals that we are quite far from achieving this goal.

This chapter provides an introduction to autonomous agents in general, but with emphasis put on the two major approaches for designing autonomous agents. It is concluded that a combination of these may be the best solution, and a suggestion for such a solution is given.

### 2.1 Introduction

All autonomous agents have, more or less, the basic architecture illustrated in Figure 2.1. The arrows in the figure symbolize data flows. The sensors receive input from

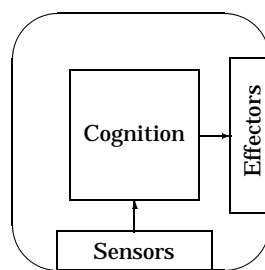


Figure 2.1: The basic architecture of an autonomous agent.

the environment and provide data for the cognitive component. The cognitive component then decides which actions to perform and commands the effectors to actually carry out these actions.

### 2.1.1 Different Kinds of Autonomous Agents

The term “autonomous agent” is, as most terms in AI, ambiguously used. What one researcher would consider an autonomous agent, another refers to as a simulation program. It is possible to divide the (most general) class of autonomous agents into categories on the basis of how, and to what degree, they actually interact with the real world. Two important features are, then, whether they are *situated* or not, and whether they are *embodied* or not. According to Brooks [36], situated agents are situated in the world in the sense that they do not only deal with abstract descriptions of it. The “here” and “now” of the environment directly influence the behavior of the agent. Embodied agents, on the other hand, “... have bodies and experience the world directly — their actions are part of a dynamic with the world, and the actions have immediate feedback on the robots’ own sensations” (p.1227). To make the distinction between situatedness and embodiment clearer, let us discuss some different kinds of autonomous agents in these terms.

Agents that are neither embodied nor situated are those that have least interaction with the real world; they are basically pure computer simulations of actual agents. A class of embodied agents which are not situated is, for instance, traditional industrial robots. They have physical bodies but do not use information about the current state of the environment to guide their behavior; they just execute a pre-programmed series of actions. A ticket reservation system, on the other hand, is situated, as the events in the environment (requests, database changes and so on) directly affect the system’s behavior. However, since the system is not physical (in some sense) and since the interaction with the environment only consists of sending and receiving messages, it cannot be regarded as embodied. Other agents belonging to this category are *softbots* (*software robots*), that is, intelligent agents in real-world software environments such as operating systems or databases. For instance, Etzioni [59] have implemented a UNIX softbot that accepts high-level user goals and dynamically synthesizes appropriate sequences of commands. In this case, the effectors are UNIX shell commands transmitted to the environment in order to change its state (e.g., `mv` or `compress`), whereas the sensors are commands that provide information to the agent (e.g., `pwd` or `ls`). Finally, we have agents that are both embodied and situated such as autonomous mobile robots (vehicles) and other physical robots that perceive the environment and use this information to guide their behavior. Table 2.1 summarizes the four possible cases.

	situated	not situated
embodied	mobile robots	traditional industrial robots
not embodied	softbots	computer simulations

Table 2.1: Categorization of autonomous agents.

It is common to regard only the members of the latter category as real autonomous agents (as I did earlier in this chapter). We will here follow this usage in the sense that we will concentrate on agents that are both embodied and situated. It is, however, not unusual to regard also agents that are not embodied (e.g., softbots) as autonomous agents.

Finally, let us make explicit the relation between traditional AI systems and autonomous agents and how they interact with the environment. In traditional AI-systems (see Figure 2.2) there is a human operator present who observes the

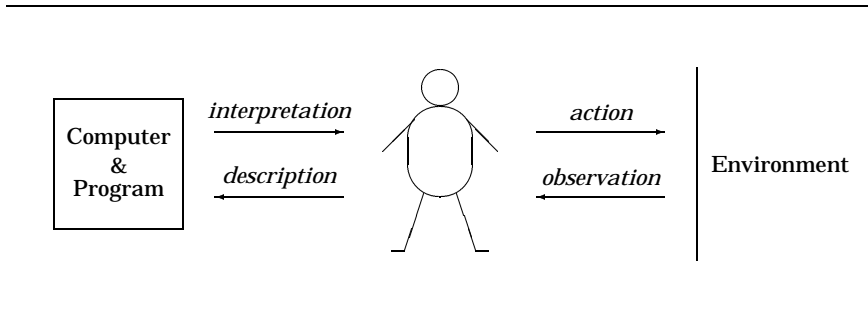


Figure 2.2: Traditional AI-system.

environment (i.e., the problem) and describes it to the computer. The results of the computer's computations are interpreted by the operator who then performs the required actions. An autonomous agent (see Figure 2.3), on the other hand, must

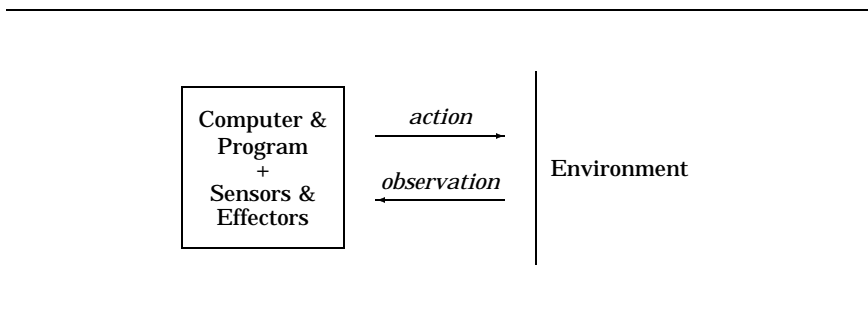


Figure 2.3: Autonomous agent.

observe the environment by itself and turn these observations into descriptions for further computations. Moreover, it must interpret the results of its computations and then perform the appropriate actions.

### 2.1.2 Requirements

What general requirements should we make on an autonomous agent? According to Brooks [35] and Hayes-Roth [84], for instance, an autonomous agent should be:



- Adaptive; it must cope appropriately and in a timely fashion with changes in the environment.
- Robust; minor changes in the properties of the environment should not lead to total collapse of the agent's behavior.
- Tactical; it should be able to maintain multiple goals and, depending on the circumstances it finds itself in, change which particular goals it is actively pursuing.
- Versatile; it must be able to perform a large variety of tasks (in contrast to being single-purposive).

At the moment, there certainly do not exist agents with all these features, the features should rather be seen as guidelines.

### 2.1.3 Scenarios for an Autonomous Agent

From the viewpoint adopted here, there are in principle two scenarios possible for an autonomous agent. Either the agent is alone in its environment, or there are other agents in the environment.

An agent on an exploration mission on Mars is a prototypical example of an agent being alone in its environment. In the other scenario where other agents exist they can be either humans or machines (or both), as on a factory floor. When it is said that other agents exist it is supposed that the agent is able to communicate with them<sup>1</sup>, otherwise it can be seen as being alone in its environment, for instance, as an agent on a factory floor that cannot communicate with the other workers. Another possible scenario is a combination of the 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 being trained here on earth and then sent away to Mars.

### 2.1.4 Possible Applications

There are numerous application areas for autonomous agents. Some of the already investigated are:

- maintenance activities in radiation-prone or toxic environments [204]
- means for the disabled [117, 200]
- deep-sea exploration and exploitation [27]
- servicing, management, and assembly tasks in space [144, 92]
- planetary exploration [18].

The objectives for developing these kinds of agents are mainly humane such as, replacing humans in hazardous, strenuous, or repetitive tasks. However, there may also be economic objectives, such as enhanced productivity, profitability, or quality. In

---

<sup>1</sup>To reduce the complexity I will in what follows suppose that communication is done, more or less, on the agent's terms. Thus, I will not bother about such difficult topics as natural language understanding.

addition, there is an enormous potential of unexplored applications which certainly will be investigated in the future, for example, personal household robots.<sup>2</sup>

### 2.1.5 Two approaches

There are two major approaches for designing autonomous agents: the traditional top-down approach and the recently emerged bottom-up approach. Agents constructed according to these approaches are often called *deliberative* and *reactive* respectively. Characteristic for the traditional approach is that the cognitive abilities (perception, world modeling, planning, and so on) are modularized. Thus, the cognitive component is functionally decomposed. In this way it is possible to begin with the design of the overall architecture of the agent and then develop the different components separately. According to the bottom-up approach on the other hand, one should start with implementing simple behavior, covering the complete range from perception to action, and then incrementally adding more sophisticated behaviors. Thus, behavioral modularization, instead of functional, is used. (This distinction is further elaborated by Brooks [35].)

## 2.2 Deliberative Agents

The traditional approach has a long tradition in several areas of research such as: AI, Cognitive Science, Philosophy and Robotics [5]. Several different agent architectures based on this approach has been suggested, for instance, SOAR [103] and Blackboard Architectures [83].

According to the traditional approach the cognitive component consists of essentially two parts; a planner and a world model. Figure 2.4 illustrates the basic architecture. The world model is an internal description of the agent's external environment

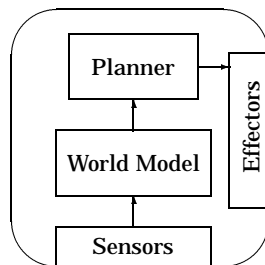


Figure 2.4: The basic architecture of a traditional agent.

(sometimes also including the agent itself). The planner uses this description to make

<sup>2</sup>In fact, this has already, at least partially, been subject for intensive studies; one symposium of the AAAI fall symposium series in 1993 [1] was devoted to autonomous vacuuming robots.

a plan of how to accomplish the agent's goal. These agents' way of working can be described as a sense-model-plan-act cycle. The sensors sense the environment and produce sensor-data that is used to update the world model. The world model is then used by the planner to decide which actions to take. These decisions serve as input to the effectors that actually carry out the actions.

There is an underlying assumption in this approach that it is possible to modularize cognition functionally, i.e., that it is possible to study the different cognitive functions (e.g., perception, learning, planning, and action) separately and then put these components together to form an intelligent autonomous agent. Moreover, this assumption seems to have influenced the division of the field of AI into sub-fields. For instance, perception (vision) is studied within the field of *computer vision*, *learning within machine learning*, planning within *planning*, and *action within robotics*. According to Minton [136] there are clearly good reasons for functional modularization from the engineering perspective as modularity reduces the apparent complexity of a system. In addition, some support for functional modularization comes from brain-research [74]: "An emerging view is that the brain is structurally and functionally organized into discrete units or "modules" and that these components interact to produce mental activities." Also within cognitive psychology, similar ideas concerning modularization has been suggested, for example, by Fodor [69].

### 2.2.1 Limitations of the Approach

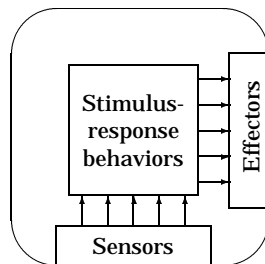
The main part of the research on deliberative agents have studied the cognitive component in isolation (i.e., as a disembodied agent). When actually embodying this kind of agents (e.g., "Shakey" [151]), however, it has been noticed that although the embodied agents are able to do some "sophisticated" cognitive tasks such as planning and problem solving, they have problems with "simpler" tasks such as routine reaction that require fast action but no extensive deliberation [36].

In addition, Brooks [35] argues that human-level intelligence is too complex and not well enough understood to be decomposed into the right components. Moreover, even if the decomposition should be right he thinks that we still do not know the right interfaces between the components.

## 2.3 Reactive Agents

The first reactive agents emerged in the mid-eighties. They were inspired by the idea that most of our every-day activities consist of routine action and not of abstract reasoning. So instead of doing world-modeling and planning, the agents should just have a collection of simple behavioral schemes which react to changes in the environment in a stimulus-response fashion. This results in the simple architecture shown in Figure 2.5 where the cognition is reduced to a mapping of perceptual stimuli onto primitive actions.

Some of the most influential agents of this kind are Brooks' robots based on the *subsumption architecture* [33], *Pengi* [3], and those based on *situated automata* [173]. In the following, however, we will concentrate on Brooks' work since it is the most influential and extreme, and since it relies on a few principles which can be explicitly stated. More comprehensive overviews of reactive agents are provided by Davidsson [49] and Lyons and Hendriks [115].



---

Figure 2.5: The basic architecture of a reactive agent.

### 2.3.1 The Work of Brooks

Some of the principles that guides Brooks' work are:

- behavioral modularization
- incremental construction
- no explicit representation of the world
- embodied agents
- purposive vision.

The reason he gives for choosing behavioral decomposition instead of functional is mainly, as described above, that we do not know which the functional components should be or, for that matter, the appropriate interfaces between them. Brooks adopts a pure engineering approach when he constructs his reactive agents and proceeds with the construction in an incremental fashion. He starts with the simplest behavior and makes it work before more advanced behaviors are added.

The most controversial of Brooks' principles is the one concerning representation. He argues that explicit representations of the world are not only unnecessary but do also get in the way when implementing actual agents. Instead the agent should use "...the world as its own model — continuously referring to its sensors rather than to an internal world model" [35]. According to him, this is possible only by having situated agents that act in direct contact with the real world, not by using abstract descriptions of it only.

Another principle is that the agents should be embodied. In this way one cannot "cheat"; one has to face all the real problems that often are disregarded otherwise. Moreover, Brooks argues that "... only through physical grounding can any internal symbolic or other system find a place to bottom out, and give "meaning" to the processing going on within the system." (In the next chapter we will further discuss the issue of symbol grounding.)

The last principle concerns perception. In AI and Computer Vision there is (or, at least, has been) an assumption that the goal of vision is to make a three-dimensional model of the real world. Brooks argues that this task is too difficult to perform when acting in the real world. Instead, his agents use *purpose vision*. (This issue also will be treated in the next chapter.)

### 2.3.2 Limitations of the Approach

Reactive agents have, at least in some experiments, been proved to be superior to traditional ones at doing a limited number of simple tasks in real-world domains. However, besides of not being particularly versatile, they have problems to handle tasks that require knowledge about the world that must be obtained by reasoning or from memory, rather than perception. According to Kirsh [100] some possible candidates for such tasks are activities which require: response to events beyond the agent's current sensory limits, some amount of problem solving, understanding a situation from an objective perspective, prediction of other agents' behavior, or creativity (i.e., stimulus free activities).

Moreover, reactive agents are often hard-wired (sometimes by hand) and do often not have any learning abilities. This and the fact that each behavior must be separately encoded in the agent, leads to complexity problems both at design time and at execution time (cf. Ginsberg [76]). (For more criticism of the reactive approach, see Kirsh [99, 100].)

However, as Kirsh points out [100], there is absolutely nothing wrong with, for instance, Brooks' approach from a scientific point of view. That is, if you interpret of his work as an attempt to see how far you can go using only this kind of simple architecture. He should also be acknowledged for having pointed out several weaknesses of the traditional paradigm. The problem is that he jumps to conclusions based on some initial success (much in the same way as early AI-scientists did).

## 2.4 Combining the Approaches

Recently several researchers [86, 31, 138, 10, 113, 116, 60, 105] have acknowledged that an intelligent agent must have both high-level reasoning and low-level reactive capabilities. In this way it is possible to utilize the reaction ability of reactive agents, which is necessary for routine tasks, and to still have the power of planning, which is necessary for more advanced tasks. Moreover, a combined approach seems to model human functioning closer than the purely reactive approach, which resembles that of more primitive animals.

### 2.4.1 Anticipatory Autonomous Agents

One way of combining reactive agents with deliberate agents is presented in Part II of this thesis. This approach is based on the idea of *anticipatory planning* [12] which, in turn, is based on the concept of *anticipatory systems* as described by Rosen [172].

## 2.5 Conclusions

There are at least two general conclusions to be drawn from this chapter. First, in order to be able to meet the requirements stated in the beginning of the chapter (i.e., that the agent should be adaptive, robust, tactical, and versatile), it seems necessary that the agent should be able to perform both reactive and high-level reasoning. Moreover, we have noticed that it is important to regard the agents as situated and embodied from the start, without making any premature modularization of the problem. That is, we should not study perception, learning, and planning independently, and when we modularize, it is important that the interfaces between the modules are well-designed.



## Chapter 3

# World Modeling

An autonomous agent based on the concept of anticipatory systems (like most other hybrid and traditional agents) will rely heavily on its world model. It is important that the model mirror the environment as close as possible so that the predictions of future states can be as probable as possible. However, since the agent is supposed to act in a highly dynamic environment, it is often not possible for the programmer to provide a complete world model to the agent at design time. As a consequence, the agent must be able to revise and update (i.e., maintain) the model autonomously as the environment changes.

In this chapter we will consider some fundamental questions concerning world models, including: what a world model is, if it is necessary to have one, and whether it is actually possible to construct adequate ones. The chapter also contains brief characterizations of the two AI-fields where the task of world modeling mainly has been studied, namely, *machine learning* and *computer vision*, together with an attempt to clarify the relation between these fields.

### 3.1 What is a World Model?

In general, a world model is an agent's internal representation of the external world. However, it is important to make a distinction between two types of world models: (1) those that only describe the current state of the agent's surroundings, and (2) those that include more general knowledge about other possible states and ways of achieving these states. I will here follow Roth-Tabak and Jain [174] and call models of the first kind *environment models*, and models of the second kind *world models*.

An environment model is typically some kind of spatial 3-D description of the physical objects in the environment. It contains dynamic and situation-dependent knowledge and can be used, for instance, in navigation tasks. Work on building environment models is described by, for example, Roth-Tabak and Jain [174] and Asada [11]. A world model, on the other hand, typically includes more stable and general knowledge about: objects, properties of objects, relationships between objects, events, processes, and so on.



### 3.1.1 Terminology

One should note that the word “model” has a different meaning in logic and related disciplines than in AI and the Cognitive Sciences. A model in logic is an *interpretation* of a sentence or theory (i.e., a description) that is stated in some formal language. In AI, on the other hand, the word “model” refers to a description, typically represented in some formal language. For reasons of convenience I will here follow the usage in AI and treat “model” as synonymous with “description”.

### 3.1.2 Explicit and Implicit Knowledge

An agent’s knowledge may be represented either implicitly or explicitly. Implicit ( $\approx$  procedural) knowledge is mainly embedded in the agent’s control and sensory processing algorithms. Knowledge is explicit ( $\approx$  declarative) when it is separated from the algorithm that uses the knowledge. Thus, we can say that reactive agents (such as Brooks’) only have implicit knowledge, whereas traditional agents mainly rely on explicit knowledge. Implicit knowledge has the advantage of simplicity, but at the expense of flexibility. On the other hand, explicit knowledge is complex, but flexible and general. Explicit knowledge is easily modified and generalizations can take place over entire classes of entities. In what follows, the term “knowledge” will refer only to explicit knowledge unless otherwise stated.

## 3.2 Is Explicit Knowledge about the World Necessary?

Does an agent need explicit knowledge about the world at all? Ten years ago this question would have been superfluous in the sense that everybody in the field believed that the answer was positive. However, as described in the previous chapter some researchers (Brooks and others) have begun to question this belief. To make the discussion more lucid let us formulate the following hypothesis: An autonomous agent needs explicit knowledge about the world to be able to act intelligible in it. By “acting intelligibly” we here mean ability to fulfill the requirements stated in Section 2.1.2 (adaptiveness, robustness, versatility and so on). Followers of the traditional view believe that this hypothesis is correct, whereas Brooks et al. believe that it is incorrect. The hypothesis can be falsified if somebody actually constructs an agent without explicit world knowledge that acts intelligible. This has not been done (yet).

In the previous chapter some support for the hypothesis was presented. We argued against purely reactive agents, mainly because of their lack of explicit knowledge about the world. One argument was that without a world model (i.e., environment model), it seems difficult to carry out tasks that demand knowledge about objects and other entities not currently perceivable. Moreover, problem solving activities are much harder to perform without explicit knowledge (i.e., world model).

Additional support for the hypothesis comes from biology. It seems reasonable to compare reactive agents to reptiles (and other lower animals) in the sense that their cognition is based on stimulus-response behavior. For instance, Sjölander writes [184]: “There is thus no true intermodality in the snake, just a number of separate systems, involving specific behavioral patterns connected to a specific sensory input.” Humans (and other higher animals), on the other hand, are equipped with central representations and Sjölander suggests that this could be an explanation of

why reptiles were superseded by mammals and birds. He concludes: “To go from monosensorially governed constructions of several internal representations to a centralized intermodal one must surely be one of the most important breakthroughs in the evolution of mind.”

### 3.2.1 On the Proper Level of Abstraction

A world or environment model, or any kind of representation, cannot perfectly describe a given subset of the real world. The only completely accurate representation of an entity is the entity itself, all other representations are only approximations (i.e., abstractions) (cf. Davis et al. [50]). Consequently, a representation must be on some level of abstraction.

On what level of abstraction should knowledge about the world (especially the environmental model) be represented? Gat [73] argues that the debate concerning deliberative agents versus reactive agents is really an argument about the proper use of world knowledge (or, as he calls it, internal state information). He argues that “... internal state should be maintained at a high level of abstraction and that it should be used to guide a robot’s action but not to control these actions directly.” (p.65) Thus, local sensor information is necessary for the immediate control.

He provides an example to show that this is also the way humans probably work. We are able to find our house and belongings because we have an environment model at a high level abstraction. We do not know the exact location of our house nor of our belongings, but we use sensor data to fill in the details that the model does not provide.

## 3.3 Is It Possible to Acquire Explicit Knowledge about the World?

As was pointed out in the beginning of this chapter, it seems unrealistic to assume that complete world and environment models can be pre-programmed into an agent. Thus, we have a situation where an agent must construct adequate models by itself. Is this really possible? The general opinion of AI researchers is that it probably is possible (but difficult). It is certainly possible in extremely simple worlds where light, reflection and other critical conditions can be precisely controlled (cf. “Shakey” [151]). However, no existing vision system is able to produce environment models of the desired quality in more realistic settings. (The task of learning more general knowledge for the world model is for some reason typically not addressed by the computer vision community.) There are two reactions to this: the optimistic, the most wide-held in AI, which says that it eventually will be possible produce the desired models, and there is the pessimistic, advocated by, for example, Brooks [34] who believes that “complete objective models of reality are unrealistic”. The reasons for his pessimism are mainly due to problems with sensors, such as noise, and the inherent complexity of the task [37].

However, here we will adopt a rather optimistic view partly based on the fact that humans indeed are able to produce sufficiently good models of the world (i.e., it is not impossible), and partly on the assumption made earlier that complete detailed models are not needed (i.e., descriptions on a rather high level of abstraction will suffice).

## 3.4 World Modeling within AI

The task of world modeling is mainly studied within two AI-fields: Computer Vision (CV) and Machine Learning (ML). In this section we will concentrate on the assumptions made and the tasks studied, rather than the methods (algorithms) used in these two fields.

### 3.4.1 Computer Vision

As humans (and most animals) rely to a large extent on vision when learning about the world, it seems reasonable to believe that visual sensors would be an important source for acquiring knowledge to autonomous agents as well. The ultimate purpose of a vision system is, according to Fischler and Firschein [61]: "... to provide the information that allows an organism to interact with its surrounding environment in order to achieve some set of goals."

In analogy with the research on autonomous agents, there exist within computer vision two competing paradigms: one traditional and one new. The traditional approach treats vision as a *reconstruction problem*; the goal is to construct a detailed symbolic representation of the world (i.e., an environment model) independent of the tasks under consideration. The new approach, on the other hand, studies vision from a *purposive* viewpoint; the goal is to solve particular visual tasks. Most traditional agents use reconstructionist vision, whereas some reactive agents, Brooks' for instance, use purposive vision. Related to purposive vision is the concept of *visual routines* [206] as used in Pengi [3].

#### Reconstructionist Vision

It is generally believed that one cannot proceed in a single step from raw camera images directly to a symbolic description. Almost all current reconstructionist vision systems successively transform the scene information through a series of representations.

The initial representation is often an intensity image (produced by for instance a video-camera). This image is typically processed numerically to produce a 2-D image description (or primal sketch, cf. Marr [118]) that makes explicit important information such as intensity changes. The 2-D image description is then used to produce a 3-D scene description that describes the spatial organization of the scene. The final stage is then to interpret this description in terms of classes of objects in order to form a symbolic description of the scene.

There are, of course, many vision systems that do not use exactly this set of representations. In general, one can say that at the lower levels of representation numerical computation is used, whereas symbolic computation is often used at the higher levels.

#### Purposive Vision

In contrast to the traditional vision paradigm where the goal is to make a complete description of the environment, the new purposive paradigm suggests that you should only describe the parts of the environment that are relevant for the tasks at hand. For instance, if an agent is looking for an object that can be used for a certain purpose,

it may only be necessary to recognize some of its qualitative features, not the exact shape. Similarly, if the agent needs to find a path from its current position to a desired position, it does not need to know the exact shapes of all the objects in the environment. From these examples it is clear that within this paradigm, vision is very task dependent (in contrast to the traditional paradigm where vision is task independent). In other words, the goal of purposive vision is to develop different “vision-guided behavior” that can be used to solve different tasks.

Since this paradigm is rather new, at least within computer-based vision, the terminology is somewhat confusing. The name *purposive vision* is adopted from Aloimonos and Rosenfeld [6], but *active vision* has also been used to label this paradigm. However, the latter is also the name of approaches to vision that use camera movements to facilitate reconstruction problems (often low- and middle-level vision) (cf. Pahlavan [154]). Yet another related concept is *animate vision* [17] that seems to cover both purposive vision and the latter interpretation of active vision.

A typical example of the use of purposive vision in the context of autonomous agents, is Herbert's [46] (one of the robots constructed at Brooks' lab) mechanism for the locating of soda cans.<sup>1</sup> Herbert does this by sweeping a plane of laser light up and down. By letting a camera, the scan lines of which are oriented vertically observe this, the depth can be computed by measuring the distance between the laser line and the bottom of the picture. Herbert considers every object that has “a flat top, and two fairly straight sides of equal length” as a soda can.

## Discussion

It is an undisputed fact that the reconstructionist approach to vision has not been completely successful. The reason is simply that the problem is inherently complex; any single object can be projected into an infinite number of 2-D images. The orientation of the object in relation to the viewer can vary continuously, giving rise to different 2-D projections, and the object can, in addition, be occluded by other objects. However, it seems that these problems can be solved, at least partially, by using stereo vision and active vision (i.e., making camera movements).

Purposive vision, on the other hand, also has some complexity problems, such as: a very large selection of “vision-guided behavior”, possibly one type of behavior for every task, seems to be needed; it probably will be hard to make decisions between competing behaviors. Additional problems (and advantages) of the two paradigms were presented at a panel discussion at IJCAI-93 [26]. The conclusion of this discussion was that extreme purposive and reconstructive views are both untenable and that a more pragmatic stance probably would be more fruitful.

In this section, we have only discussed visual sensors. However, since the task studied here is not 2-D image analysis, but actual perception of real objects, there are other kinds of sensors that can also be useful, for instance: range, proximity, force and touch sensors. For a study on multisensor integration refer to, for instance, Lou and Kay [114].

---

<sup>1</sup>Herbert's task is to wander around in office areas, going into peoples' offices and stealing empty soda cans from their desks.

### 3.4.2 Machine Learning

Machine Learning is the sub-field of AI that studies the automated acquisition of (domain-specific) knowledge. The aim is to construct systems that are able to learn, i.e., systems that improve their performance as the result of experience [179]. Learning in many different contexts has been investigated, two of the most predominant being classification and problem solving.

The most relevant kind of learning for world modeling is probably *concept learning*, which belongs to the classification domain. The goal in concept learning is typically to learn descriptions of categories that can be used for classifying instances. As input the learning system is typically given a set of *descriptions of instances*. We will discuss concept learning in greater detail in Chapter 8.

### 3.4.3 Computer Vision – Machine Learning Relation

In reconstructionist computer vision, learning is often restricted to the creation of environment models. This is typically done by using only object models already known (i.e., internal representations of the objects, or categories, to be recognized). Thus, the learning of new knowledge about categories is not performed, resulting in static systems that do not increase their performance in the face of experience. However, some examples of vision systems that actually learn this kind of knowledge will be presented in Section 8.2.2.

Within ML and within concept learning in particular, on the other hand, learning problems where the input is in symbolic form are almost exclusively studied. There is, however, an emerging understanding in the field that a change is necessary. For instance, Michalski and Kodratoff write [130]: “So far, this input information (examples, facts, descriptions, etc.) has been typically typed in by a human instructor. Future machine learning programs will undoubtedly be able to receive inputs directly from the environment through a variety of sensory devices.”

One might conclude that there is a natural relation between computer vision and machine learning; the vision system produces symbolic descriptions that the learning system can use as input. However, since these fields are often studied in isolation we cannot be sure that the vision system produces output that can be used by the learning system (regarding to both form and content). Moreover, it is not clear at which level learning should take place. It may be the case that the learning of different kinds of knowledge should be done on different levels. Thus, to integrate ML and computer vision we must in some way smoothen the transition from signals (analog, subsymbolic representations) to symbols.

#### From Signals to Symbols

Since most of an agent’s knowledge is about its environment, it must somehow extract this information from observations of the environment. Since we are dealing with autonomous agents that receive information directly from the environment and process it on different levels by different systems (for instance, vision and learning systems), we seem to need several notions of observation.

It might be useful to follow Gärdenfors [72], who distinguishes three levels of

describing observations.<sup>2</sup> The highest level is the *linguistic* level, where observations are described in some language (cf. ML). The second, intermediate, 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, like temperature, color and size, are closely related to what is produced by our sensory receptors while others, like time, are more abstract. At 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 (cf. computer vision). This input, however, is too rich and unstructured to be useful in any conceptual task. It must be transformed to suit the conceptual or the linguistic level. To do this the subconceptual information must be organized and its complexity reduced.

Whereas a situated and embodied agent probably has to deal with observations on the subconceptual level, disembodied agents such as softbots receive their information at the linguistic level. Harnad [81] makes a distinction between learning based on perceptual observation which he calls *learning by acquaintance*, and *learning by description* that bases the learning on observations at the linguistic level.<sup>3</sup> Thus, most ML systems perform learning by description, whereas computer vision-based systems, such as situated and embodied agents, are forced to learn by acquaintance.

This three-level perspective also raises several more or less philosophical questions. For instance, is it convenient or even possible, to perform any useful cognitive tasks (e.g., reasoning) on a sub-linguistic level, or do we need some kind of language? If we assume that some kind of language is necessary for reasoning, then observations must be described on some linguistic level in the reasoner. On the other hand, it seems clear that at some (early) stage in the sensors, observations must be described on the subconceptual level.

We can conclude that there are at least two levels of representation (observation) involved here: the linguistic and the subconceptual. Gärdenfors' conceptual level is only one suggestion of an intermediate level. As pointed out earlier, a series of representations are used in reconstructionist vision: the raw camera image (corresponding to the subconceptual level), 2-D image description, 3-D scene description (intermediate levels), and symbolic description (linguistic level). A problem with Gärdenfors' conceptual level is that, while the intermediate representations of CV-systems are designed for representing structural information, it is not clear if and how the conceptual spaces can handle this type information.

### The Symbol Grounding Problem

The symbol grounding problem as described by Harnad [82] concerns the meaning of the symbols in symbol systems. Traditional AI systems manipulate symbols that are systematically interpretable as meaning something. The problem is that the interpretation is made by the mind of an external interpreter. The system itself has no idea of what the symbols stand for. Since this is a problem that arises in all symbol systems, it concerns also the world model of an autonomous agent. However, as described in

---

<sup>2</sup>A similar distinction is made by Harnad [81].

<sup>3</sup>This distinction is probably inspired by Russell's [176], but is not equivalent to his.

Part III of this thesis, this problem disappears if we are able to integrate vision and learning in an adequate way and to use appropriate representation schemes.

### 3.5 Conclusions

As pointed out earlier, it is often assumed within AI that it is possible to modularize the cognition process without much thought; we solve the planning problem, you solve computer vision, and a third part solves the learning problem, and then we meet and build an autonomous agent. As we have noted, many of the underlying assumptions of ML and computer vision are not compatible. Most important is the difference in representation schemes. Whereas ML-systems typically use symbolic descriptions, often in the form of attribute-value pairs, to describe objects, CV-systems typically use non-symbolical, geometrical 3-D models. On the other hand, learning is essential for the development of vision systems, since the construction of object models by hand is very tedious and often does not produce the desired results. Thus, it is easy to agree with Moscatelli and Kodratoff [141] when they write "... that adaption through the use of machine learning is the key to autonomous outdoor vision systems." The general conclusion is that closer integration between machine learning and computer vision is necessary. The central problem is not only how to make the transition from the subsymbolic data that the sensors output to symbolic representation, but to decide on which levels different kinds of learning should take place.

Another fundamental problem is that in traditional computer vision there is a strong emphasis on building environment models and thereby ignoring the problem of building world models. It is typically assumed that there is a pre-programmed world model.

A problem that has not been mentioned earlier in this thesis, is how to make the agent focus its attention on the relevant aspects of the environment. It is not computationally tractable to process all the information available through the sensors. 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(s) 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? One suggestion, put forward by Sjölander [184], is that the focus of attention could be controlled by an anticipating higher level. For instance, a dog hunting a rabbit uses his internal world model to predict the rabbit's future positions. These predictions can then be used to focusing the attention on the relevant aspects of the situation.<sup>4</sup> Actually, these ideas are in line with the anticipatory agents approach described in Part II. Notice also that in this case, the focus of attention is closely coupled to the task at hand.

What if the hypothesis that an agent needs explicit knowledge about the world turns out to be false? Even if world and environment models, contrary to the author's belief, will turn out to be obsolete, it seems that an agent cannot manage without *concepts* (which actually may constitute a considerable part of a world model). Thus, we can formulate the weaker hypothesis that an autonomous agent needs concepts to

---

<sup>4</sup>This kind of anticipatory behavior can be of help in recognition tasks as well. For example, the dog needs not to see the whole rabbit all of the time. Since it can predict the rabbit's current position, it needs only glimpses of parts of the rabbit to confirm its predictions.

act intelligibly. This hypothesis is supported by, for instance, Kirsh [100] and Epstein [58]. The concepts must not be explicitly represented though. In Brooks' robots for instance, the concepts are present only implicitly in the circuits and the mechanical devices (cf. Herbert's representation of "soda cans"). However, Epstein argues that there are several advantages with having explicit representations of concepts. For example, it facilitates the organization of knowledge and the focus of attention.

In the rest of this part of thesis we will concentrate on the notions of concept and category. Starting (in the next chapter) with some fundamental issues, such as what it means to have a concept, and what possible purposes of concepts there are.





## Chapter 4

# Concepts and Categories

In the remaining chapters of this part of the thesis we will concentrate on concepts (and categories). Since this topic is rather complex, especially in an autonomous agent context, it is useful to divide it into the following four sub-topics:

- the functions of concepts
- the nature of categories
- the representation of concepts (i.e., categories)
- the acquisition of concepts.

Given the task of constructing an autonomous agent able to have and to acquire concepts, the only issues of direct interest would be the representation and the acquisition of concepts. However, as will become apparent, the representation is dependent on what functions the concepts should serve. Moreover, the choice of representation is constrained by the nature of the actual categories that they represent. As a consequence of this, and since representation and acquisition are clearly dependent on each other, it is obvious that we should not study either representation or acquisition of concepts without also examining these more fundamental sub-topics.

Each of the next four chapters will be devoted to one of the sub-topics listed above. In this chapter, we will after a terminological discussion concerning the words “concept” and “category”, discuss the basic question of what it actually means to have a concept.

### 4.1 Terminology

As with most other terms that are shared between several research fields, the term “concept” has been used in many different ways. In everyday language “concept” often refers just to the *name*, or designator, of a category (i.e., a word). However, the main concern here is not the linguistic task of learning the names of categories (although we will also discuss briefly this topic). As a matter of fact, the contents, and scope, of this thesis have very weak connections to traditional linguistics, if any at all. Rather, the assumption is made that concepts are independent of language, or at least that they can be studied independently of language.<sup>1</sup> The rejection of any

---

<sup>1</sup>By language we mean here a natural language, not an internal language (i.e., language of thought, mentalese).

assumption of language as a prerequisite for concepts has been made by a number of scientists. Edelman [57], for instance, cites chimpanzees as an example of animals which lack linguistic abilities but can have, and are able to acquire, concepts.

Instead, we will in what follows use the term in a different way, more in line with uses within cognitive science and AI. Smith [188] provides a typical cognitive science definition: "... a concept is a mental representation of a class or individual ...". However, we will here not deal with concepts representing individuals. A typical AI, or rather ML, definition is Rendell's [164]: "The term concept is usually meant as a description of a class (its intension)."<sup>2</sup> What in these definitions are referred to as classes we will call categories, and by category we will mean a set 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.<sup>3</sup>

In contrast to these mainstream definitions we have, for instance, more general ones such as Epstein's [58]: "Concept is defined here as some recognized set of regularities detected in some observed world." There are also some rather odd ones like Matlin's [121]: "A concept is a way of categorizing items and demonstrating which items are related to one another." In this definition the term concept refers to a process rather than a representation.

In the light of these definitions, and several others not mentioned, it is not difficult to agree with Heath [85] who suggests that: "the term "concept" is thus essentially a dummy expression or variable, whose meaning is assignable only in the context of a theory, and cannot be independently ascertained." Thus, before we continue we should make explicit the intended interpretation of the term "concept". In this thesis we will use, and have used, these definitions:

**Definition 1** *A category is a class of entities<sup>4</sup> in the world that are united by some principle(s).<sup>5</sup>*

**Definition 2** *A concept is an agent's internal representation of a category.*

Although these definitions are consistent with the most common uses in cognitive science and AI other good suggestions exist. One is provided by Matheus [120] who suggests that: "a concept is a purposeful description of a possibly fuzzy class of objects." This definition is more specific than Definition 2 in the sense that it contains a constraint, i.e., that the concept should be purposeful. As will become apparent, the author completely agrees with Matheus that the purpose, or function, of a concept is important. However, since it is possible to think of a concept without an explicit purpose, we will hold on to the purer definition. The other addition made, as compared to our definition, "possibly fuzzy", seems redundant (i.e., the class of objects may be fuzzy, not fuzzy, or whatever) and is thus unnecessary. On the other hand, in our

<sup>2</sup>Thus, the commonly used term "concept description" is tautologous since a concept actually is a description.

<sup>3</sup>Unfortunately the terms "category" and "concept" are sometimes used in a rather arbitrary manner in the literature. For instance, Michalski writes [127]: "... concepts, that is, classes of entities united by some principle."

<sup>4</sup>Examples of entities are objects, events, and situations.

<sup>5</sup>As mentioned above, 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.

definition we emphasize that the representation is internal to an agent. This is due to the approach of this thesis: studying concepts in the context of autonomous agents.

When discussing concepts it is important to clarify which fundamental view one has on universals. There are basically two camps: realists and non-realists. Realists believe that universals are non-mental, mind-independent entities that exist in themselves. Plato was a typical realist. Non-realists, on the other hand, argue that universals are mental, mind-dependent entities that would not exist if there were no minds. Conceptualism, suggested by the classical British empiricists (e.g., Locke and Berkeley), and nominalism, suggested by Hobbes among others, are some of the non-realist theories. A more detailed discussion of these theories is provided by Woozley [216].

As may have been understood by earlier statements, we are here adopting a non-realist stance. Thus, rather than being a priori entities, it is supposed that categories are the inventions of an agent or a collective of agents, used to structure the environment in order to facilitate cognition. Examples of collectively formed categories are “chair” and “ostrich”. More personal categories invented by a particular individual are, for example, “the-things-that-are-mine” and “articles-relevant-for-my-thesis”. In ML, both the learning of concepts representing categories invented by humans (i.e., learning from examples) and categories formed by the learning system itself (i.e., learning by observation, or concept formation) are studied.

## 4.2 What does it mean to have a concept?

Kirsh [99] has suggested that in AI “the worry about what it is to have a concept is seldom articulated.” When considering AI in isolation this is certainly true, but in related fields there have been made several attempts to state explicitly what it means to have a concept. For instance, the philosopher Heath [85] suggests that to have a concept “x” is:

- to know the meaning of the word ‘x’
- to be able to pick out or recognize a presented x, or to be able to think of x (or x’s) when they are not present
- to know the nature of x, to have grasped or apprehended the properties which characterize x’s and make them what they are.

As we can notice, these conditions are rather vague, especially if we try to apply them to artificial agents. Several questions remain open: What is it for a computer system to know the meaning of, to think of, and to apprehend something? What is the nature of a concept? It is only the first part of the second condition, to be able to recognize instances of the category “x”, that seems reasonably straightforward.

A proposal that is easier to comprehend, comes from Smith [187], a cognitive psychologist, who suggests that: “To have a concept of *X* is to know something about the properties of *X*’s instances.” However, it seems that this condition is too weak and underspecified; it seems not to capture the full meaning of “having a concept”. Kirsh, on the other hand, gives the following view (in AI terms) on the problem:

We cannot just assume that a machine which has a structure in memory that corresponds in name to a structure in the designer’s conceptualization

is sufficient for grasping the concept. The structure must play a role in a network of abilities; it must confer on the agent certain causal powers [25]. Some of these powers involve reasoning: being able to use the structure *appropriately* in deduction, induction and perhaps abduction. But other powers involve perception and action—hooking up the structure via causal mechanisms to the outside world. [99] (p.10)

From this, and the above discussion, it seems appropriate to draw two conclusions. (1) It is perhaps not adequate to treat “having-a-concept” as a two-valued predicate (i.e., either you do have the concept or you do not). We should instead think in terms of a continuum of degrees of having the concept. (2) Instead of trying to state a number of conditions that should be satisfied for having the concept, it seems that a more fruitful approach would be to ask which *functions* (cf. causal powers) a concept should have. The more functions it can serve and the better it can serve these functions, the higher is the degree to which one has the concept. In the next chapter we will discuss the functions of concepts.

## Chapter 5

# The Functions of Concepts

This chapter is devoted to the functions of concepts. Here, and in most of the following chapters, we will start with a survey of the research in the cognitive sciences on the current sub-topic of human concepts. This is then followed by a survey of the corresponding work in AI and a discussion that compares the research and tries to draw some conclusions concerning this aspect of concepts in an autonomous agent context.

The importance of investigating the functions of concepts and how they affect the representation and acquisition, becomes even more apparent if we study the following example. It is taken from a paper written by Michalski and some of his colleagues [22] that describes their two-tiered approach for the learning of concepts from examples. The example in the paper concerns the category “chair”. One part of the representation suggested by the authors is shown in Figure 5.1. A two-tiered representation consists of two parts. The Basic Concept Representation (BCR) describes the most relevant properties of the category whereas the Inferential Concept Interpretation (ICI) handles, for instance, special cases. Since the ICI is not very relevant to the point I wish to make, it is omitted here. (We will, however, discuss the two-tiered approach later.) The BCR part of the representation says that: a chair is a kind of

---

*Superclass:* A piece of furniture.

*Function:* To seat one person.

*Structure:* A seat supported by legs and a backrest attached from the side.

*Physical properties:* The number of legs is usually four. Often made of wood. The height of the seat is usually about 14–18 inches from the end of the legs, etc.

(BCR may also include a picture or a 3D model of typical chairs)

---

Figure 5.1: The BCR part of a two-tiered representation of the category “chair”.

furniture, its function is to seat one person, it consists of a seat that is supported by legs and a backrest, it often has four legs and is often made of wood and so on.

This seems to be a rather powerful representation that probably can be used for several purposes. It is probably something like this that we want our agent to have. In

any case, it seems more appropriate than most other approaches to the representation of categories suggested in the ML literature. But if we study the text in the article closer, we find that this description is *not* something that their system has learned. They write: "... (for an example of a two-tiered chair description actually learned, see [21])".

Before we take a look at the learned description, let me describe the learning situation. The system was given a number of positive and negative examples of chairs. To the left in Figure 5.2 a positive example is depicted. However, the input to the system was symbolic descriptions as illustrated to the right in Figure 5.2. We should

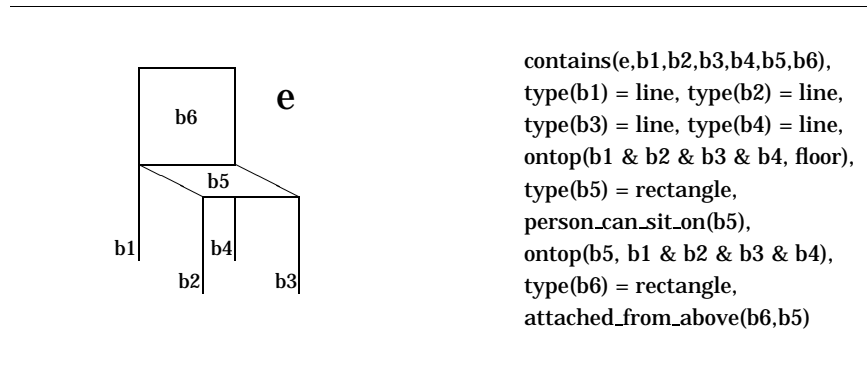


Figure 5.2: A 2-D line drawing and a symbolic description of a chair.

notice at this point that the transition from real objects to symbolic descriptions is made in two steps (using two different kinds of representation). To the left in Figure 5.2 we have a 2-D line drawing (a sub-symbolic representation of the actual object) and to the right we have a symbolic representation of this drawing. Both these transitions, which in this case are made by hand, are of course very difficult to implement in a computer.

From a number of such symbolic descriptions, the representation in Figure 5.3 was learned.<sup>1</sup> It says that a chair is something that has something that a person can

---


$$\exists x \exists z \exists (y \geq 3) [ \text{person\_can\_sit\_on}(x) ] \ \& \ [ \text{type}(y) = \text{leg} ] \ \& \\ [ \text{ontop}(x,y) ] \ \& \ [ \text{attached\_from\_above}(z,x) ]$$


---

Figure 5.3: The actually learned BCR part of a two-tiered representation.

sit on that is on top of at least three legs. Moreover, there should be something that

<sup>1</sup>It was also given some background knowledge such as the following rule:  $[\text{type}(x) = \text{line}] \ \& \ [\text{ontop}(x,\text{floor})] \Rightarrow [\text{type}(x) = \text{leg}]$ . (It seems somewhat inconsistent, however, to use a high-level predicate such as "person\_can\_sit\_on(x)" directly in the description of the example. It would have been nicer if a similar rule had been used to infer it from more basic predicates.)

is attached to the “sit-thing” from above.

This seems to be a much less powerful description than the first one. In any case, it contains less information. Why does not the system learn something like the first description, which seems better? The reason is, in fact, rather obvious. The second representation is learned for a certain purpose; it is meant to serve a certain function. Namely, to discriminate between (symbolic descriptions of) chairs and non-chairs. The first description, on the other hand, is probably meant to be more general in the sense that it should be able to serve many functions.

It should be stressed that this does not cause any problems in most traditional AI settings where the learning system is used as a tool (used to improve human performance). Because in this case, discriminating between members and non-members of a category might be just what we want the system to do; it might be precisely the function we want the learned concept to serve. The other functions can be taken care of by the human operator. In an autonomous agent setting, on the other hand, there is no human operator available. Thus, a more powerful way of representing categories, able to support all the desired functions is needed.

Which are these functions that the concepts should serve? Unfortunately, this question has almost never been discussed in the AI literature. So, to answer this question we have to turn to literature in cognitive science and philosophy to find out what functions human concepts serve. A survey of this work will then provide a basis for a discussion on what functions the concepts of artificial agents should serve.

## 5.1 The Functions of Human Concepts

To begin with, we can restate some of the very first words of this thesis. Namely, that concepts seem to be the very stuff on which reasoning and other cognitive processes are based. Actually, it is difficult to think of a mental activity that does not make use of concepts in one way or another. However, it is possible to distinguish several functions of human concepts, some of them are:

- stability functions
- cognitive economical functions
- linguistic functions
- metaphysical functions
- epistemological functions
- inferential functions.

This list is inspired by different work in cognitive science and philosophy, in particular by Rey [167] and Smith [187]. However, we will not always use the terms in exactly the same ways as in these articles.

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 chair, we can compare this situation with other situations where we have encountered chairs. 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 set of objects in 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 “chair” instead of having a representation for every chair we have ever experienced, we do not have to remember that the chair we saw in the furniture-shop yesterday can be used to rest on.

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 “chair” and the Swedish word “stol” have the same meaning enables us to translate “chair” into “stol” and vice versa. Furthermore, it seems that it is the linguistic function together with the interpersonal stability function that makes 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 can say that something actually is a chair if it has been made with the purpose of seating one person (or something like that).<sup>2</sup> 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 chair by size, material, form, and so on. A better example for illustrating this distinction is the category “gold”. Something is actually a piece of gold if it has a particular atomic structure. However, when we recognize a piece of gold, we use other features such as: color, weight, and so on.<sup>3</sup> We should note that both these functions are related to categorization: the metaphysical considers what actually makes an entity an instance of a particular category, whereas the epistemological considers how an agent decides whether the entity is of a particular category.

Finally, concepts allow us to *infer* non-perceptual information from the perceptual information we get from perceiving an entity, and to make predictions concerning it. In this sense, we can say that concepts enable us to go beyond the information given. For instance, by perceptually recognizing a chair we can infer that it can be used to rest on, or by recognizing a scorpion we can infer that it is able to hurt us. This is maybe the most powerful function of concepts; it emphasizes the role of concepts as the central element of cognition. As Smith [187] writes: “Concepts are our means of linking perceptual and non-perceptual information ... they serve as entry points into our knowledge stores and provide us with expectations that we can use to guide our actions.” In addition to prediction, concepts allow us to explain relationships, situations, and events.

---

<sup>2</sup>We use the word “metaphysic” in a more pragmatic way than in philosophy. In our 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.

<sup>3</sup>For human concepts this distinction is maybe not as clean-cut and unproblematic as described here (cf. Lakoff [104]) but nevertheless it suits our purposes very well.

## 5.2 Functions of Concepts in Artificial Autonomous Agents

As mentioned earlier, the functions of concepts have almost never really been subject to discussion in AI-literature. The only treatment of this topic known to the author is made by two AI-researchers, Matheus and Rendell, and two psychologists, Medin and Goldstone, [119]. They consider five functions: *classification*, *prediction*, *explanation*, *communication*, and *learning*. The classification function corresponds to our two categorization functions, the epistemological and the metaphysical. Prediction, on the other hand, can be seen as a special case of what we have called the inferential function. This can to some extent also be said about the explanation function.<sup>4</sup> The communication function can be divided in two kinds of functions. One kind corresponds approximately to a combination of what we have called the (external) linguistic function and the interpersonal stability, providing the basis for a shared understanding. The other kind of communication function is that of transmitting the description of the concept to some other representational system (e.g., a human). This function is more relevant for traditional ML-systems in which the learned concepts are intended for human comprehension than for autonomous agents where the concepts are mainly created for internal use. The last function on their list, learning, seems hard to regard as a *function* of concepts.<sup>5</sup> It should, of course, be possible to learn the concept, and the easier this, the better. (To facilitate learning, on the other hand, could possibly be seen as a function of concepts.) These topics are of great importance and will be discussed in later chapters. However, let us now concentrate on the actual functions of concepts.

In ML there is often an implicit assumption made that the concepts acquired are to be used for some classification task (cf. the example in the beginning of this chapter). Thus, the function of the concepts learned by ML-systems is mainly of an epistemological (or metaphysical, depending on attitude) nature. To see if this is also sufficient for autonomous agents, we will now go through the functions in the previous section one by one, discussing whether they are desirable (or necessary) or not, for an artificial autonomous agent.

The function of intrapersonal stability is of course important, but it is trivial in the sense that it emerges more or less automatically for the agent just by having concepts, independently of the choice of representation. This can also be said about the function of cognitive economy, that is, cognitive economy will emerge as long as we do not memorize every instance of the category (i.e., an extensional description).

By analogy to the stability functions, we can say that an agent's concepts can serve both intrapersonal and interpersonal linguistic functions. However, the intrapersonal function is a rather weak one, implied only by the fact that the categories have names internal to the agent (that may be different to the external names). 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 also must be able to serve these functions.

---

<sup>4</sup>The explanation function was probably included as a separate function because of the then recently emerged learning paradigm explanation-based-learning (EBL) [139, 52] that was very influential at the time.

<sup>5</sup>They mainly discuss whether the concept description is adequate for incremental learning. As we shall see later, however, this is a very important question in the context of autonomous agents.

It is, however, the remaining three functions, the metaphysical, the epistemological and the inferential, that are the most interesting, and the ones we will further concentrate on in the remaining part of this thesis. 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. In fact, the relevance of this function even for psychology is not fully understood, for contrasting opinions compare Smith et al. [190] (less relevant) and Rey [168] (more relevant).

Finally, if the agent is to be able to reason and plan about objects it is necessary for it to have at least some inferential functions. This is certainly the central function of concepts corresponding to both the prediction and the explanation function of Matheus et al. [119]. For an autonomous agent, however, the ability to predict future states seems more relevant than the ability to explain how the present state has arisen. For this reason, we will in the following mainly discuss the inferential function in terms of predictions.

## Chapter 6

# The Nature of Categories

What can be said about *categories* in general, without taking into account the issue of representation and acquisition of concepts? (Remember that by categories we mean classes of entities in the world and that concepts are an agent's internal representations of these.) As with the functions of concepts, this issue, which we will refer to as the nature of categories, is almost never discussed in the AI literature. In analogy with the last chapter, we will begin with a survey of the psychological and philosophical research on the nature of categories that humans use and then discuss the nature of the categories used by an artificial agent.

In this chapter we will try to identify the most important categories of categories. Moreover, one section will be devoted to a discussion on the nature of properties. In the definition of category we stated that it was a class of entities united by some principle(s). The most important of these principles, similarity, will be treated in detail. Finally, some important issues regarding taxonomies will be discussed.

### 6.1 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 equivalence classes that are constructed for a particular psychological experiment [189],<sup>1</sup> 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 often not 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 able to learn concepts efficiently, find this state of affairs rather unfortunate, since (1) machines learn definitions that specifies artificial concepts relatively easily, and (2) an autonomous agent in a real-world environment will have to deal with natural categories, not artificial. However, during the last decades it has become apparent that the study of artificial categories will not significantly increase our understanding of how humans really acquire real concepts (i.e.,

---

<sup>1</sup>Typical examples of artificial categories can be found in Bruner et al. [38]. The problem-domain is cards with different symbols. On each card there are one, two, or three symbols of some kind: circle, square or cross. These are colored red, green, or black. Finally, the cards have one, two, or three borders. A category in this domain is then, for instance, "cards that have red crosses".

representations of natural categories). Therefore, some researchers have begun to use natural categories for their experiments. This movement toward a more sound approach is further elaborated by Neisser [148].

Members of natural categories are either concrete, such as physical objects, or abstract, such as events or emotions. In what follows we will concentrate on concrete object categories. While dealing with autonomous agents trying to learn about their environment, this is a quite natural initial assumption; the environment consists of physical objects. However, at some point in the future it will certainly be necessary to introduce also event categories.

As we will see later, humans use other types of categories that cannot be classified as natural, namely *derived* [186], or *ad-hoc* [20], categories. Moreover, natural categories can be divided into *natural kinds* and *artifacts* [186]. However, let us begin by examining a more fundamental topic: the properties of objects.

### 6.1.1 Properties

In both cognitive science and AI, it is generally assumed that the basis for the representation and categorization of an object should be a set of properties that characterize the object. A general opinion is, however, that there are fundamental differences between different kinds of properties. Several ways of dividing properties into classes have been suggested.

Sometimes a distinction is made between *quantitative* properties and *qualitative* properties. Quantitative properties are often called dimensions and are exemplified by, for instance, temperature in °C. Qualitative properties, on the other hand, are nominal features that can take on only a finite number of discrete values (binary features that are either true or false being a special case) and are exemplified by, for instance, temperature measured in terms of hot and cold. It is, however, in principle possible to transform any qualitative property into a quantitative and vice versa.<sup>2</sup> As a consequence, we will in most cases not hold on to this distinction, and just speak of properties (sometimes, however, referring to them as features or attributes).

Moreover, some properties are called *perceptual*, in the sense that they (in some sense) are directly available from the perceptual system, while others are more *abstract*, such as functional properties (cf. Smith and Medin [189] p.18). Furthermore, some features are *global* (e.g., a chair is made of wood) whereas others describe *parts* (e.g., a chair has legs).

What is considered a feature is relative in the sense that some features correspond to categories themselves, or, at least, can be regarded as principles that unite categories (cf. Definition 1). As a matter of fact, we have a kind of tree-hierarchy of categories. An example of such a hierarchy is shown in Figure 6.1. (Observe that only a small part of the hierarchy is shown. Apples obviously have more than three properties and so on.)

Smith and Medin [189] make a distinction between the *identification procedure* and the *core* of a concept. They argue that some of the leaves (i.e., the branch endpoints, in this case, “edible” and “round”) of the resulting tree structure are perceptual

<sup>2</sup>Transformation from quantitative into qualitative is straight-forward. The values for any dimension can be expressed as a set of nested features [13]. When we go in the opposite direction, however, we may lose some aspects that we usually ascribe to quantitative properties. For example, that a dimension should have the property of betweenness (cf. Smith and Medin [189] p.13).

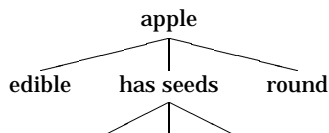


Figure 6.1: Part of the property-hierarchy of the category “apple”.

features (“round”)<sup>3</sup> are those that we normally use to determine which category an object belongs to. In their terminology, they constitute a considerable part of the identification procedure. The core of the concept is the features on the second level (edible, has seeds, round). Thus, the identification procedure is closely related to what we have called the epistemological function whereas the core together with the rest of the features is more connected to the metaphysical and inferential functions. Or like Smith [187] puts it: “When reasoning we use the cores, when categorizing we use the identification procedures.” (p.29)

The structure in Figure 6.1 is not entirely consistent, though. It is primarily the features that represent parts (i.e., has seeds) that can be thought of as categories.<sup>4</sup> The functional features (i.e., edible) and perceptual features are best thought of as just features. In Figure 6.2 we have a tree-structure that describes the different parts of an instance of a category.<sup>5</sup>

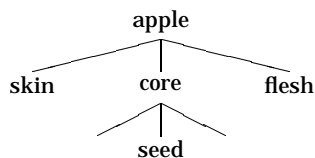


Figure 6.2: Part of the part-hierarchy of the category “apple”.

### 6.1.2 Natural Kinds

It seems natural to assume that categories emerge as a consequence of the correlational structure of the environment, i.e., where the perceived 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 encounter an elephant for

<sup>3</sup>Whether “round” is a perceptual feature or not may be open for discussion, but let us for a moment suppose that it is.

<sup>4</sup>Actually, it is “seed” that is the category.

<sup>5</sup>Note that this hierarchy does not reveal how the different parts are fitted together. To do that we need a structural description. Moreover, do not confuse “core” (i.e., the part of the apple where the seeds are located) with the core of the concept.

the first time (supposed that you have not read, or been taught, anything about elephants). Then, 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 [161] has termed this type of categories *natural kinds*. Rosch and her colleagues [171] also emphasized that natural categories emerge in this way, assuming 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, and inconsistent with our non-realist stance, to argue that there exist only 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. Thus, a more sensible and somewhat less strong, rather epistemological, claim would be that some categories “through human perception” stand out as natural categories.

### 6.1.3 Similarity

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 category. 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*. In these models two objects are regarded as similar if the difference between the properties used to characterize them is small. A fundamentally different approach is to regard objects similar if they are related by some kind of transformation taken from a set of specified transformations (cf. Kanal and Tsao [95]). In this case, objects that do not share many properties may be regarded as similar and vice versa.

Geometric models (cf. Shepard [182]) tend to treat all properties as quantitative. An object is represented as a point in the coordinate space that is defined by the dimensions used to describe the object. The similarity between two objects is then measured (defined) by the metric distance between them, i.e., the closer they are, the more similar they are. 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 [132].
- All selected attributes are given equal weight [132].
- It is more appropriate to represent some features as qualitative [205].

The geometric models seem related to Gärdenfors’ conceptual spaces. Let us try to make this relation explicit. It is possible to interpret 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 above have been taken care of, corresponding to a “refined” geometric model.

In set-theoretical models, on the other hand, objects are represented as collections of qualitative features. The most well-known set-theoretical model is Tversky’s [205] *contrast model*. It expresses the similarity between two objects as a linear combination of the measures of their common and distinctive features. However, pure set-theoretical models such as Tversky’s 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 features must be weighted seems to be implied by the theorem of the ugly duckling provided by Watanabe [210].<sup>6</sup> 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 [178] 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:

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. (p. 640)

This suggests that the categories that humans normally use not always arise in the purely *bottom-up* fashion [88] described above. Thus, 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 [169] argues (taking back her earlier strong claim) that some types of attributes present a 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 seem 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 in such a way that the categories, once given, would appear maximally distinct from one another.

Similarly, Murphy and Medin [143] have claimed that people’s intuitive theories about the world guide the representational process. They placed the demand on 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.”

To sum up, 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, certain kinds of categories. However, an actual perceptual system of an embodied

---

<sup>6</sup>This theorem, which is formally proved, shows that whenever objects are described in terms of logical predicates, 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.



autonomous agent will have some built-in constraints that determine what will count as an attribute and the salience (weight) an attribute will have. This topic seems closely related to what Harnad [81] has labeled *categorical perception*. He writes:

For certain perceptual categories, within-category differences look much smaller than between-category differences even when they are of the same size physically. For example, in color perception, differences between reds and between yellows look much smaller than equal-sized differences that cross the red/yellow boundary ... Indeed, the effect of the category boundary is not merely quantitative, but qualitative. (p. 535)

Thus, in an autonomous agent the perceptual constraints, or categorical perception, are determined by the physical properties of its sensors.

#### 6.1.4 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 perceptual features such as structure and color. Thus, as Corter [48] 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 (i.e., they are “similar”), whereas top-down categories often consist of disjunctive groupings of different types of objects that may not share many properties (i.e., 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 [20] 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,<sup>7</sup> 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 behavior.” As an example of an ad-hoc category he takes “activities to do in Mexico with one’s grandmother”. However, there also are some permanent top-down categories such as “food”.

#### 6.1.5 Artifact Categories

Not all natural categories are natural kinds. A natural division can be made between *species* (i.e., natural kinds) and *artifacts*. Rosch’s examples above, “chair” and “table”, which certainly are natural categories, are typical artifacts. Characteristic for artifacts is 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 (i.e., most chairs look like each other). Moreover, some objects made for one purpose may be used for another purpose, it is for instance possible to use a chair

---

<sup>7</sup>The representations can, of course, be modified but they are permanent in the sense that there always exists a representation of the category.

as a table. Thus, we can say that artifact categories differ from natural kinds in that they seem to have the potential to arise both in a bottom-up and a top-down fashion.

We can now summarize the discussion by suggesting a classification of categories. Figure 6.3 illustrates the hierarchical relationships between the different types of categories as they have been described above. First we have the ordinary natural

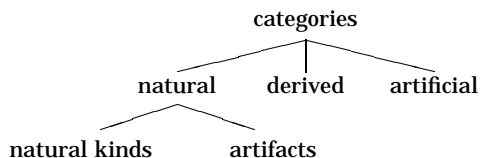


Figure 6.3: The relationships between the different kinds of categories.

categories that have evolved through everyday use. These are either natural kinds or artifacts. In contrast to natural categories there are the artificial categories, often constructed for a particular scientific experiment. In addition to these we have the derived, or ad-hoc, categories that are typically formed during problem solving.

### 6.1.6 Taxonomies

Categories can also be hierarchically organized in a different way than by their properties or parts, namely, in taxonomies (i.e., a hierarchical classification scheme that shows how categories are divided into sub-categories).<sup>8</sup> A part of a taxonomy is illustrated in Figure 6.4. (Figure 6.3 is, in fact, also a taxonomy.)

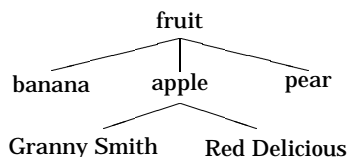


Figure 6.4: Part of a taxonomy of fruits.

Apart from organizing knowledge, taxonomies also serve an important function by promoting cognitive economy. How this is possible is demonstrated by Figure 6.5 and Figure 6.6. In Figure 6.5 we have a part of the fruit-taxonomy augmented with some features of the categories.

<sup>8</sup>This is a rather strong idealization since some categories may not belong to any taxonomy at all while others belong to several.

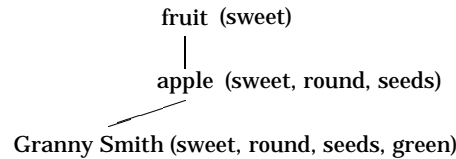


Figure 6.5: Part of a part of a taxonomy of fruits augmented with features.

By noticing that categories on one level inherit the features from its parent category (i.e., the category on the level above) we can reduce the amount of information that we must store on each level. This is illustrated in Figure 6.6.

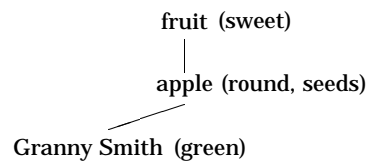


Figure 6.6: Part of a part of a taxonomy of fruits augmented with features (optimized).

Rosch et al. [171] argue that there exists a “basic level” in these taxonomies. They write: “Basic categories are those which carry the most information, possess the highest cue validity<sup>9</sup> and are thus, the most differentiated from one another ... Basic-level categories 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 6.4) 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 humans prefer for categorization, the epistemological function is probably, in some respect, maximized at this level. In addition, 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 (e.g., fruit) have relatively few properties and hence cannot enable us to make that many inferences. Although subordinate categories (e.g., Granny Smith) have many properties they have so few distinctive properties that

<sup>9</sup>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.

they are more difficult to categorize perceptually. Finally, we should note that the question of which level is actually the basic level is context dependent in the sense that the normal basic level is the most appropriate in most situations but not all.

## 6.2 The Nature of AI Categories

In traditional AI, categories are often presumed to be artificial; they are often constructed for a particular experiment and it is assumed that all relevant aspects of the category can be summarized by a short and simple definition in terms of necessary and sufficient conditions. However, it is becoming more common to test and evaluate learning algorithms on real-world data, forcing the learning systems to deal also with natural categories.

### 6.2.1 Properties

In most existing learning systems, objects are described by a list of attribute-value pairs. Typically, these attributes represent only global properties. Some systems, however, make use of *structural descriptions* (cf. [55, 202]) that in addition to global properties also involve properties of the object's parts and relationships among these parts.

An assumption often made when constructing concept learning systems, is that attributes are atomic units of description that are given to the system and are to be used as building blocks of concepts without further processing. Some systems, however, do create new attributes not present in the input data. They try to generate high-level features from the lower level features originally used to characterize the instances. This task, or problem, is often labelled *constructive induction* or the *new-term problem*. According to Rendell [165], one of the purposes of constructive induction is to transform the instance space to diminish disjuncts.<sup>10</sup> An overview of approaches to constructive induction is provided by Rendell [166].

### 6.2.2 Similarity

Most work on concept formation in ML is concerned with bottom-up concept formation, often called similarity-based learning (SBL). However, exceptions such as explanation-based learning (EBL) [52, 139] exist. In EBL the categories are formed beforehand and a high-level description of them is given as input to the learner. The task is to transform the (abstract) high-level characterization into a low-level characterization (often in terms of perceptual features). Thus, no categories are actually *formed*.

In similarity-based learning, both geometric and set-theoretic models of similarity are frequently used. Geometric models are often used by *conceptual clustering* systems, such as CLUSTER/2 [132, 197], whereas systems that learn from examples, such

---

<sup>10</sup>The instance space consists of all possible examples and counterexamples of concepts to be learned. (The description space, on the other hand, is the set of all descriptions of instances or classes of instances that are possible using the description language of the learner [127].) Disjuncts are separate clusters of category members.

as version spaces [137], often use set-theoretic models. Models based on transformations, on the other hand, are rare. Nagel's [147] approach to learning from examples is one of the few.

Michalski and Stepp [132] propose an approach for measuring similarity in geometric models that, besides the two object descriptions, takes into account other objects and the set of concepts (in this case, features and background knowledge) that is available for describing categories. They call this measure *conceptual cohesiveness*. The background knowledge may include: definitions of property range and scale, specificity hierarchies over property values, implicative rules of constraints of one property on others, rules for the construction of new properties, suggestions or derivational rules for ranking properties by potential relevancy [196].

In AI the problem of selecting relevant features is often solved by letting the user select them. This choice of features is one kind of *bias*<sup>11</sup> of a learning system. In some systems, however, the learning system itself 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 [102] and neural networks [72].<sup>12</sup>

In the spirit of Schank and his colleagues [178], some experiments have been conducted that use explanations to select relevant attributes when doing top-down concept learning (EBL). However, the success has been limited, probably due to the difficulties in specifying the appropriate background knowledge.

### 6.2.3 Taxonomies

Taxonomies and their properties are rather well studied both in AI and in computer science in general. Take, for instance, object-oriented languages, such as Smalltalk and Simula, where the classes are members of taxonomies and where features are inherited from super-classes. The topic of taxonomies in AI and computer science is further elaborated by Jansson [94]. However, among the existing concept learning systems it is only the conceptual clustering systems (cf. [62]) that actually construct taxonomies. Some of these systems (cf. [80, 64]) also try to include basic-level aspects.

## 6.3 Conclusions

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

### 6.3.1 Properties

It seems that some features are represented more naturally as qualitative and some as quantitative. Thus, it would be desirable to have agents that could handle both types of features.

<sup>11</sup>Actually, this is one of several types of bias. Other types are, for instance, 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 (cf. Utgoff [207]).

<sup>12</sup>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.

The assumption that features are atomic entities readily available for the cognitive module is not compatible with our discussion in the chapter concerning world modeling on different levels of observation. In particular, the notion of perceptual features seems anomalous. This problem has also been acknowledged by Wrobel [217] who writes:

... we believe that any concept formation process relies on the filtered perception/interpretation of the world that the observer imposes ... this means that any concept formation model relying on features must include an account of their creation. Otherwise, we would have only replaced the concept formation problem by the equally hard feature formation problem. (p.713)

It seems that a natural candidate for solving this problem would be constructive induction. However, as Wrobel points out, all existing approaches construct their new features in terms of features already known. Thus, the new features are not more powerful than the original ones in the sense that they cannot distinguish objects that could not be distinguished with the original set of features, they are just abbreviations that allow more concise concept descriptions. Instead, he suggests that the more primitive features might be innate structures that has developed through evolution. In an autonomous agent context this would correspond to hard-wired, or pre-programmed, structures. This idea seems closely related to the concept of categorical perception.

Let us, however, suppose that some of the properties of an object are known to the learning system. How should these properties be used? First, we can note that there seems to exist properties of different types. Some properties, common to all objects of the category,<sup>13</sup> are characteristic or discriminant, these can be used for metaphysical and epistemological classification (e.g., for the category “human” we have, for instance, the genetic code and “walking upright” respectively). Other 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 heart). The more or less useless properties, sometimes called irrelevant properties, that are not common to all objects of the category (e.g., hair color) are then left over. These can be used for the representation of individual entities, but this is beyond the scope of this thesis.

### 6.3.2 Similarity

As we have seen, the formation of bottom-up categories has been rather well studied in AI. However, some problems remain to be solved, such as finding an appropriate similarity measure (and whether such a measure actually is necessary) and the origin of features. 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

---

<sup>13</sup>One should not take “all” too literally. It may be the case that universal regularities do not exist, implying that reasoning about categories must be probabilistic in nature.

the emphasis on the function of the artifact objects. The recognition of the possible functions of an object from perceptual observations seems like a very hard problem.<sup>14</sup> However, some ideas of how to transform structural knowledge about objects into functional knowledge are presented by Vaina and Jaulent [208]. On the other hand, it would require a very large amount of background knowledge 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 then constructing temporal top-down (derived) categories when convenient in problem solving tasks.

In sum, one might somewhat carelessly 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 categories comes from more abstract observations. Thus, a passive agent, just trying to make a description of its environment, could manage with only bottom-up categories, whereas a problem solving agent will probably also need top-down categories.

### 6.3.3 Taxonomies

Finally, we need to structure the categories into taxonomies to promote cognitive economy and inferential functions. Since this is a topic that has been extensively studied within computer science, it is probably better to concentrate on other less studied problems. However, the phenomena of basic levels needs further studies.

---

<sup>14</sup>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." [189] (p.19). This problem is to some extent studied within explanation-based learning. Moreover, the functions must, of course, be known in advance, pre-programmed 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 given explicitly, this problem with functional properties disappears. However, assuming that the agent has access to such observations is too generous for most applications.

## Chapter 7

# Representation of Categories

We are now ready to study how concepts, an agent's internal representations of categories, should be represented. In this section we will discuss the following questions: How do humans represent categories? How do present AI systems represent categories? How should AI systems represent categories?

### 7.1 Human Representation of Categories

Medin and Smith [125] present three views of human concepts: the *classical*, the *probabilistic* and the *exemplar*. These views are to a great extent theories about representation. We will begin with a discussion of the classical view and the problems it has to explain some empirical findings concerning human behavior.

#### 7.1.1 The Classical View and It's Problems

According to the classical view, all instances of a category share common features that are singly necessary and jointly sufficient for defining the category. Moreover, it says that it would suffice to represent a category by these features, thus generalizing the details of the instances of the category into a single, summary description covering all members of the category. Categorization would then be a matter of straightforward application of this "definition".<sup>1</sup> For instance, a classical representation of the category "chair" could be: can be used by a person to sit on, has at least three legs, and has a back. Thus, by this definition, an object is a chair *if and only if* it can be used by a person to sit on, has at least three legs, and has a back.

However, there are some problems with this view (cf. Smith et al. [189, 187]):

- For some *natural* categories it seems not possible to find necessary and sufficient features.
- Even if a category can be defined as above we tend not to use this definition.
- There are unclear cases of category membership.
- Some instances of a category are regarded as more typical than others.

---

<sup>1</sup>Although this chapter mainly is concerned with representational issues, we will in most cases describe how a particular representation is intended to be used.



- 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 [7]. A nice and famous example, mentioned by Wittgenstein, is the category “game”. He argued that instances of “game” share many common features, but that no subset of these features can be found to be both necessary and sufficient.

Assuming that a classical definition exists for a category, it is interesting to notice that instead of using this definition we often (and are sometimes forced to) use non-necessary features to characterize a category or to categorize objects of the category. For instance, in recognizing a piece of gold, we generally cannot perceive the atomic structure of the material directly.<sup>2</sup> Instead, we use such features as color and weight.

Not only is it in some cases unclear which particular category an object belongs to; the same person may even categorize an object differently as the context changes [123]. For example, it is sometimes hard to decide for some objects whether they are bowls or cups. In this example there is a relation between an object and a category but the same problem can arise between two levels in a taxonomy (i.e., subcategory-category relations). For instance, is a tomato a fruit or a vegetable, or is a rug a piece of furniture?

The observation that people regard some instances of a category as more typical than others inspired the invention of the notion of *prototypes*. However, this term has been used ambiguously, often in one of the following meanings:

1. the best representative(s) or most typical instance(s) of a category
2. a description of the best representative(s) or most typical instance(s) of a category
3. a description of a category that is more appropriate for some members than it is for others.

The first of these thus refers to actual objects, whereas the other two refer to representations. Moreover, the second refers to a representation of a singular object, whereas the third refers to a representation of a class of objects. It is possible to make a further distinction regarding the second meaning: the described instance may be either (a) an actual instance, or (b) a constructed ideal, or average, instance that does not have to correspond to an object in the world.

The introduction of prototypes is in opposition to the treatment of categories as equivalence classes. It has, for instance, been shown that (at least for the experiment subjects) robins and bluebirds, in contrast to penguins and bats, are prototypical birds [187]. However, the existence of prototypes does not have any clear implications for the construction of models of human category representation, processing and learning. Thus, prototypes do not specify such models, only impose constraints on them.

The last problem on the list concerns the fact that it seems that we often think about specific objects when we actually refer to a category. For example, if someone says that he had to see a dentist (without specifying which dentist), it is hard not to think of a specific dentist.

---

<sup>2</sup>We are here assuming that it is possible to provide a classical definition of gold in terms of its atomic structure.

From these five objections, it seems clear that the classical view cannot explain all aspects of human concepts. 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* [215, 170]. A common measure of family resemblance is the number of features that are shared by members of a category. Thus, it can be viewed as a measure of typicality since typical members of a category share many attributes with other members of the category (and few with members of other categories). Conforming to these considerations, the probabilistic and the exemplar view have been presented as theories being more realistic and consistent with empirical findings.

### 7.1.2 The Probabilistic View

According to the probabilistic view, a category is represented by a summary representation in terms of properties that may be only probable, or characteristic, of its members. Membership in a category is graded rather than all-or-none. Better members have more characteristic properties than the poorer ones. Thus, this kind of representation corresponds to a prototype in the third meaning as described above.

Several approaches to probabilistic category representation have been proposed [189]. They differ mainly in the assumptions made regarding the nature of the properties used to describe the categories. The *featural approach* assumes that the properties used to characterize categories and objects are qualitative (i.e., features) and that the probability for them to occur in instances of the category is high.<sup>3</sup> A category is represented by a list of weighted features, where the weight corresponds to the probability and/or salience of the feature occurring in an instance of the category. A probabilistic representation of the category “chair” could be, for example: (1.0) can be used of a person to sit on, (0.9) has at least three legs, (0.9) has a back, (0.7) is made of wood. An object will then, for example, be categorized as an instance of a category if it possesses some critical sum of the weighted properties included in the representation of that category.

The *dimensional approach* assumes, in contrast to the featural approach, that the properties are quantitative (i.e., dimensions). These dimensions, which ought to be relevant (salient), provide a multidimensional space (cf. geometric models of similarity). A category is then represented by a point in this space, a prototype, which is typically the “average” instance. For instance, “sit-ability” = 1.0, number of legs = 4, ... An object will then be categorized as an instance of a category if it is within some threshold distance of the prototype. Thus, rather than applying a definition, categorization is in this case a matter of assessing similarity. The dimensional approach as described here have, in fact, many similarities with the exemplar view (and can be seen as a hybrid between the probabilistic and the exemplar views). For instance, the point that represents the category is a prototype in the second meaning rather than the third. However, while the prototypes of the exemplar view are descriptions of actual instances (2a), the prototypes of the dimensional approach are descriptions of constructed entities (2b).

---

<sup>3</sup>In addition, they should be salient in some respect (perceptually or conceptually).

### 7.1.3 The Exemplar View

Those in favor of the exemplar view argue that categories should be represented by (some of) their individual exemplars, and that concepts should correspond to representations of these exemplars.<sup>4</sup> Thus, such a representation corresponds to a prototype in the second meaning as described above. 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 that simply stores all instances. An instance is categorized as a member of the category that contains its most similar stored exemplar. Another model is the *best examples* model that stores only 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. Yet another approach is the *context model* [124] that in addition to individual instances also allows exemplars to be summary descriptions and thereby reduces the number of exemplars to be memorized. The context model also differs from other exemplar models in that the categorization is to some extent context dependent.

### 7.1.4 Combining the Probabilistic and Exemplar View

Another possibility is that the representation of a category contains both a probabilistic summary representation and exemplars [189]. 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* [75] and *belief revision* [70]. Consider a point in time where a person has both a summary and an exemplar representation of the category "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 such a combined representation.<sup>5</sup>

---

<sup>4</sup>In contrast to the classical view that seems to try to capture the intension of concepts, the exemplar view (at least partially) describes their extension.

<sup>5</sup>The possible connection between prototype-based representations and non-monotonic reasoning has been pointed out by Gärdenfors [71]. 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.

## 7.2 Representation of Categories 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. In this section we will, after a brief review of general AI approaches for representing categories, concentrate on the different types of concepts used in the sub-fields of machine learning and computer vision.

### 7.2.1 General AI Category Representations

As pointed out earlier, the concept of concepts has not been a main subject for research in AI. *Semantic networks* [160], however, is one approach for representing categories that has gained some attention. One of the main ideas behind semantic networks is that the meaning of a concept comes from the ways it is connected to other concepts. The information is represented as a set of nodes, which represent categories (or objects), connected to each other by labeled arcs, which express relationships among the nodes. Figure 7.1 shows an example of a partial semantic network representing the category “chair”.

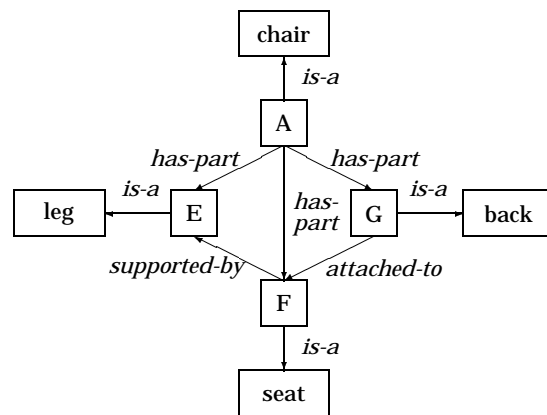


Figure 7.1: Part of a semantic network representing the category “chair”. Only one leg (E) is included. Leg B, C and D are supposed to be connected in the same way as E.

A *frame* [134] is a general structure that has been successfully used for representing different kinds of knowledge. Essentially, a frame is a collection (of representations) of facts that, for instance, can be used to represent a category or an instance of a category (see Figure 7.2). The contents of the frame is a list of slots that define relationships to other frames that have various functions in the definition. For instance, the definition of the slot “made of” states that the frame MATERIAL has the function of being the stuff of which a chair is made. Moreover, a slot can contain a default value that is used in the absence of other information. Thus, unless told

---

```

frame CHAIR
  number of legs: NUMBER
  made of: MATERIAL
  back: BOOLEAN default = true

```

---

Figure 7.2: A frame representing the category “chair”.

otherwise a system using the frame in the example will infer that a chair has a back. An approach, similar to frames, for representing event categories is *scripts* [177].

These representation schemes, frames in particular, are assumed to be used mainly for inferential functions. This, in contrast to the other representations that has been, and will be, presented that mainly are used for categorization.

### 7.2.2 Machine Learning Category Representations

In ML, it is mainly three kinds of representation languages that have been used to represent classical concept definitions: *logic-based notations*, *decision trees*, and *semantic networks*. In one of the first concept learning programs ever made, Winston [213] employed semantic network representations of structural descriptions (both of instances and categories). The network illustrated in Figure 7.1 is an example of what such a description might look like.

Logic-based notations have been used in, for instance, the AQ-programs [126]. We have, in fact, already seen an example of a logic-based representation, namely the BCR part of a two-tiered representation of the category “chair” from Chapter 5 (repeated in Figure 7.3). A disadvantage with the classical view that has not been

---


$$\exists x \exists z \exists (y \geq 3) [ \text{person\_can\_sit\_on}(x) ] \ \& \ [ \text{type}(y) = \text{leg} ] \ \& \\ [ \text{ontop}(x,y) ] \ \& \ [ \text{attached\_from\_above}(z,x) ]$$


---

Figure 7.3: Logic-based category representation (conjunctive).

mentioned, is that classical definitions, being limited conjunctive descriptions, are not able to represent disjunctive category descriptions. For example, if we want our description of chairs also to cover wheel-chairs, we need augment it with a disjunction as illustrated in Figure 7.4.<sup>6</sup> In what follows, however, we will regard disjunctive descriptions of this kind as belonging to the classical view as well.

The most popular way of representing classical concept definitions is probably by using decision trees. The first attempt to learn induction trees resulted in the CLS program by Hunt and his colleagues [91]. It is, however, the ID3 program by Quinlan

---

<sup>6</sup>This is, of course, dependent on the description language. If we have term that covers both legs and wheels, a conjunctive description would be sufficient.

---


$$\exists x \exists z \exists y (y \geq 3) [ \text{person\_can\_sit\_on}(x) ] \& [ [ \text{type}(y) = \text{leg} ] \vee [ \text{type}(y) = \text{wheel} ] ] \& [ \text{ontop}(x,y) ] \& [ \text{attached\_from\_above}(z,x) ]$$


---

Figure 7.4: Disjunctive logic-based representation.

[162] that can be regarded as the source for the present popularity of induction of decision trees. Figure 7.5 shows an example of a decision tree describing the category “chair”. At each node in the tree there is a test for sorting instances down the

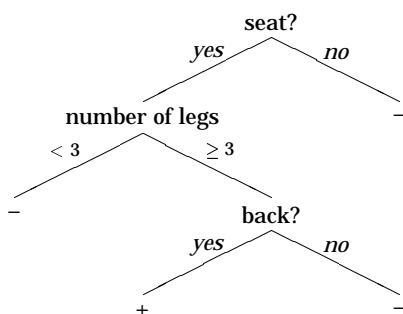


Figure 7.5: Decision tree representation of the category “chair”.

alternative branches. The terminal nodes are marked with either a “+”, to indicate category membership, or a “-”, to indicate non-membership. In this example, the tree corresponds to a conjunctive description since only one terminal node is marked with a “+”. It is, however, possible to achieve a disjunctive representation by having several such nodes. A disadvantage with decision trees is their not being suited for representing structural properties.

### Non-traditional Representations

Within the last few years, several ML experiments with non-classical representations have been carried out. While some researchers are inspired by the exemplar view, others are influenced by the probabilistic view.

Beginning with those who are influenced by the exemplar view, Kibler and Aha [97] have experimented both with the proximity model where all instances are stored, and with selected examples models where only a subset of the instances is stored. Systems using this kind of representation often employ 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. In contrast to most of the approaches

inspired by the exemplar view which use geometric or set models of similarity, Nagel [147] suggests a best examples model that employs a transformation-based similarity model. In addition to the prototype(s), transformations are stored that transform less typical instances to a prototype. Learning systems that use specific instances rather than abstractions to represent categories have by Aha and his colleagues [4] been labeled *instance-based*. They also provide a theoretical analysis of such algorithms.

A follower of the probabilistic view is, for instance, de la Maza [51] who calls his type of representation *augmented prototypes*. These are created from prototypes which consist of one vector for every attribute that is used to describe the examples. If the attribute is nominal, the vector contains the frequency of each value that the attribute can take on. If the attribute is continuous the vector contains the mean and standard deviation of the attribute. Thus, this is a hybrid approach that combines the featural and the dimensional approach. The augmented prototypes are created by adding a weight to each attribute that are computed by comparing the prototypes.

Another probabilistic approach is described by Musgrove and Phelps [145]. They have adopted the dimensional approach where the prototype reflects the average member of the category. Moreover, they use multidimensional scaling to reduce the number of dimensions. Yet another approach is Fisher's [64] *probabilistic concept tree* that represents a taxonomy of probabilistic concepts.

### Subsymbolic Representations

Recall Gärdenfors' three levels of observation from Section 3.4.3. 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, or symbolic, level. A method of representing (and acquiring) concepts on a lower level is using neural networks. A neural network basically consists of a number of nodes connected by weighted links. Neural networks were initially meant to be cognitive models of the brain at the level of neurons.

Pylyshyn [159] 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 [150] refers to this level as the symbol manipulation level. Thus, whereas the representations discussed earlier have all belonged to the middle level, neural networks belong to the lowest of these levels.

During the last years there has been a growing optimism about the capacities of neural networks, both as cognitive models (e.g., the works of Grossberg and Carpenter [40] and of Edelman [57]) and as tools for pattern recognition (e.g., backpropagation networks [175]). However, one must keep in mind that neural networks that can be simulated on a computer (i.e., most current neural networks), are of course at the most Turing-machine-equivalent. They might be better suited (e.g., more efficient or easier to program) than symbolic systems for some problems, but are not a more powerful tool in general.

In addition, there are some problems with neural networks. For example, neural networks do not represent knowledge explicitly, something which seems crucial for the implementation of the metaphysical and inferential functions. The functions that

the subsymbolic methods will be able to handle seem, at least for the moment, limited to tasks like perceptual categorization.<sup>7</sup> Thus, there is a possibility that they might be able to implement the epistemological function. However, most neural network approaches assume that the input instances are represented by a set of binary, or real-value, features, and are thus not suited for structural descriptions.<sup>8</sup> Moreover, it is difficult to introduce background (a priori) knowledge.<sup>9</sup> Furthermore, it is difficult to exploit and reason about both the learned knowledge and the learning process. A more detailed discussion about possibilities and limitations of connectionist models in general is provided by Smolensky [191].

### 7.2.3 Computer Vision Category Representations

The typical goal of model-based object-recognition systems for robot vision (i.e., the subfield of computer vision that will be regarded here) is to “recognize the identity, position, and orientation of randomly oriented industrial parts” [45].<sup>10</sup> However, there is a problem with comparing the representations that such systems use when recognizing objects with the category representations we have discussed earlier. Since the objects to be recognized typically are industrial parts, they have almost the same shape, whereby the representations (or models, as they often are called) used in this task often are not representing categories, but rather can be seen as representations of specific objects. That is, they are categories, but on a much lower level than, for instance, the basic level (e.g., the category might be wrenches of a particular type and brand, rather than wrenches in general). Thus, almost no generalization takes place and representations are mainly of a conjunctive nature.

According to Chin and Dyer [45], there have been three types of representations used in model-based vision systems: 2-D,  $2\frac{1}{2}$ -D, and 3-D object models. A 2-D model consists typically of shape features derived from the silhouette (i.e., the set of boundaries) formed by the gray-scale (or binary) image of an object. 2-D models can be used when the objects to be recognized have a simple structure and are presented against a high-contrast background. Moreover, systems based on such models are often only able to recognize the objects from a few fixed viewpoints and typically demand that they are not occluded by other objects. Thus, this class of representation is only appropriate for tasks where the environment can be completely controlled. The advantage with 2-D models is that they are relatively easy to construct automatically.

In conformity with 2-D models,  $2\frac{1}{2}$ -D models are viewer-centered representations. In addition, however, they try to capture surface properties of the object such as range (depth) and surface orientation. Thus,  $2\frac{1}{2}$ -D models are descriptions of surfaces rather than boundaries. Disadvantages with such models are that they, just as 2-D models, are viewpoint-specific and that the additional step of deriving the surface description makes them harder to construct automatically.

<sup>7</sup>Even though the opposite opinion is sometimes held (cf. Balkenius and Gärdenfors [16]).

<sup>8</sup>However, some initial experiments also trying to incorporate structural knowledge have been carried out (cf. [15, 192]).

<sup>9</sup>There have been experiments introducing symbolic knowledge into “knowledge-based” neural networks, see for instance Towell et al. [203]. However, in the author’s opinion these networks are rather symbolic than subsymbolic representations since every node explicitly represents something. This implies moreover that the knowledge in these kinds of nets is not distributed, which is one of the characteristic features of neural networks.

<sup>10</sup>Cf. the “bin-picking” problem in which the parts to be identified are disorderly placed in a bin.



In contrast to 2-D and  $2\frac{1}{2}$ -D models, 3-D models are viewpoint-independent representation. They are often volumetric representations that allow a complete object description from an unconstrained viewpoint using either sweep models, surface models, or volume primitives. Representations that have been used are for instance: generalized cylinders (“sweep representations”) [32], surface patches [183] (see [29] for a recent review of surface-based representations), superquadrics (superellipsoids) [19], geons (a subset of the generalized cylinders) [24, 54].

In a (3-D) model-based vision system, the recognition of objects consists in matching the input image with a set of models that are typically preprogrammed using a CAD system. Some critical assumptions often made, are (1) that the objects (categories) to be recognized are exactly specified, with known tolerances on dimensions and features, (2) that number of objects (categories) is usually small (less than 50). Thus, most existing approaches are expected to function only in very controlled environments.

Another kind of 3-D model is multi-view feature representations in which a set of 2-D or  $2\frac{1}{2}$ -D descriptions (one for each relevant view), also called *characteristic views*, are combined into a single composite model. Thus, rather than trying to capture the *shape* of objects, they try to describe the *appearance* of objects. An interesting approach called *Appearance models* is described by Murase and Nayar [142]. The authors argue that since “the appearance of an object is the combined effect of its shape, reflectance properties, pose in the scene, and the illumination conditions”, recognizing an object from a brightness image (such as the output from a video-camera) is more a problem of appearance matching rather than shape matching. The appearance model is constructed from a large set of images of the object with varying pose and illumination. This set is then compressed into an eigenspace (a low-dimensional subspace) in which the object is represented as a hypersurface. To recognize an unknown object, you only need to check which hypersurface the image of the object is projected onto (the exact position can, in addition, be used to determine pose and illumination). As we shall see in the next chapter, however, there are some assumptions made regarding the learning of these models that makes it hard to employ this approach directly.

### 7.3 Conclusions

So, how should autonomous agents represent categories? Let us analyze this question in terms of the functions that the concepts should be able to serve. From the above review of existing approaches, it should be clear that most of them concern representations to be used in some categorization task. Thus, they can be said to serve the epistemological function.<sup>11</sup>

Although there may be some categories that can be characterized by a classical definition, such a definition is often based on features that under normal circumstances are impossible, or at least difficult, to detect by perception, such as atomic structure, genetic code or functionality. Thus, these definitions are not adequate for perceptual classification,<sup>12</sup> and consequently not appropriate representations for supporting the

<sup>11</sup>Even if some researchers probably would argue that their systems perform the metaphysical function.

<sup>12</sup>However, in traditional AI it is very common to try to make a classical definition of a category based

epistemological function. Instead, the implementation of the epistemological function seems to demand some kind of prototype-based, or possibly subsymbolic, representation. Moreover, since structural relations (between parts) are important when perceptually categorizing objects, a representation supporting the epistemological function should also be able to represent structural relationships. As we have seen, only a few of the representations used in machine learning have the ability to describe structural relationships, whereas in computer vision this is compulsory. On the other hand, object models used in computer vision often lack the ability to represent global features such as color, texture, and function that can be useful in object recognition.

Smith, Medin and Rips [190] suggest that there are two kinds of epistemological categorization. One that makes use of the properties in the core of the concept and one that uses the identification procedure. They write:

Specifically, identification properties are often useful for a “quick and dirty” categorization of objects, and such properties tend to be salient and easy to compute though not perfectly diagnostic of category membership: core properties, on the other hand, are more diagnostic of category membership, but they tend to be relatively hidden and hence less accessible for rapid categorization. (p.267)

They illustrate this by the problem of identifying gender of a person (i.e., categorizing instances of the categories “males” and “females” respectively). Categorization by identification properties may then take into account style of clothing, hair, voice and so on, whereas categorization by core properties may involve what kind of sexual organ the instance has. This latter kind of categorization bears a strong resemblance to what we have called metaphysical categorization.<sup>13</sup> Thus, the distinction between categorization by core properties and metaphysical categorization is clearly unclear. However, since it is required that the rules that determine category membership are explicitly represented, the implementation of the metaphysical function demands by definition a classical definition.

To implement the inferential function, on the other hand, it seems that we must have some “encyclopedic” knowledge about the category and its members. This knowledge is naturally seen as a collection of universal or probabilistic rules. Kirsh [98] has called this collection “a package of associated glop”. Closest at hand is, of course, the representing of this kind of knowledge in some logic-based notation and/or by frames. However, another possibility would be to use diagrammatic representations (cf. [42]).

The probably most important reflection on the review above is that almost all work on category 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. This opinion seems to be shared by the majority of the members of the cognitive science community.<sup>14</sup> However, the above discussion should make clear that this is not possible except in very restricted domains. This implies that a richer, composite representation is needed that is structured according to the desired functions of the concept. The insight that multiple

---

directly on the perceptual data.

<sup>13</sup>Although one might suggest that metaphysical categorization would take into account the genes of the object.

<sup>14</sup>Surprisingly, even in the debate in Cognition [167, 190, 168] where the need for concepts to serve several functions was stressed, it was assumed that categories should be represented by a single structure.

category representations sometimes are required, has only on very few occasions been explicitly expressed in the AI-literature. As an example, Flann and Dietterich [67] write:

The performance task for which a concept definition is to be learned may require a structural representation (e.g., for efficient recognition), a functional representation (e.g., for planning), or a behavioral representation (e.g., for simulation or prediction). (p.461)

In a similar vein, Matheus [120] draws the conclusion that “... there are purposes for which a single representation simply cannot satisfy all requirements.” (p.42) In more general terms, Sloman [185] points out that different purposes require different kinds of knowledge representations. In the next section we will present some AI approaches that make use of multiple representations.

### 7.3.1 Multiple Category Representations

There are at least two senses in which a category representation can be composite (multiple):

- the components merely use different vocabularies
- the components are represented in fundamentally different ways.

Flann and Dietterich [67] present an approach of the first kind where the components use the same fundamental representation but with different vocabularies. A concept consists of two parts: the *learning representation*, which is used to facilitate effective (i.e., incremental) learning, and the *performance representation*, which is used to facilitate task performance. This approach has been applied to the learning of concepts in board games (e.g., “skewer” and “knight-fork” in chess). In this case the performance task was the recognition of board positions that are instances of such categories. Since these kinds of concepts are naturally functional (in the sense that the similarity between the instances is mainly of a functional nature whereas they may be very dissimilar according to structure, cf. derived categories), the learning representation used a functional vocabulary. The performance representation, on the other hand, was represented in a structural vocabulary that permits efficient matching against board positions.

An approach to multiple category representation where the components are represented in fundamentally different ways<sup>15</sup> is, as we have seen earlier, taken by Michalski and his colleagues [128, 22]. 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 is a set of inference rules that 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.

Yet another approach is Rendell’s PLS [163] which makes use of three different kinds of representation: an exemplar-based table of counts that is used for learning, a

<sup>15</sup>However, they use the same representation *language* (a variation of predicate logic).

probabilistic region representation that is used for generalization, and an evaluation function that is used for prediction.


However, common to these approaches is the fact that they do not support all the desired functions presented in Section 5. In fact, they support only one single performance task (categorization or prediction). The supporting of all the desired functions, on the other hand, implies that the concept must be used for several performance tasks including, for instance, categorization, communication, and prediction. Thus, none of the presented multiple category representations is powerful enough to meet the demands required by an autonomous agent.

### 7.3.2 A Novel Framework for Composite Concepts

In Part IV of this thesis, a new approach for representation of categories by autonomous agents is presented. It consists of the following five components: the external designator, the internal designator, the epistemological representation, the metaphysical representation, and the inferential representation.<sup>16</sup>

The idea of this composite representation is illustrated in Figure 7.6 by using the category “chair”. For the external designator it is natural to choose “chair” (in an

---

<i>External designator:</i>	chair
<i>Internal designator:</i>	thing.artifact.gfj65
<i>Epistemological component:</i>	
<i>Metaphysical component:</i>	can seat one person
<i>Inferential component:</i>	Number of legs is usually four, often made of wood have a back, can be used to sit on, ...

---

Figure 7.6: Composite representation of the category “chair”.

environment where communication is based on the English language, that is). The choice of the internal designator, on the other hand, is entirely up to the system, it should be as convenient and effective as possible for the system. In the figure, the epistemological representation is a 3-D model of a prototypical chair, but any representation that can be used by the perceptual system to successfully identify members of the category would be adequate. For the metaphysical representation I have chosen that something is a chair if it could seat one person. Finally, the encyclopedic knowledge in the inferential representation includes the facts that a chair usually has four legs, and is often made of wood and so on.<sup>17</sup>

<sup>16</sup>However, 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. Moreover, the external designators are only necessary for communicating agents.

<sup>17</sup>Note that in the metaphysical and inferential representations in the figure, it is only the *content* of the representations that are described (in natural language).

Of course there are no sharp distinctions between what types of information is included in these representations. They may even contain redundant information. For example, besides being a part of the epistemological representation, the fact that chairs have legs is a rather natural part of the encyclopedic knowledge represented in the inferential representation. However, the fact is not represented in the same way in these representations. For instance, it may be implicitly represented in a prototype-based representation for the epistemological representation and explicitly represented in a logic-based representation for the inferential representation.

This composite structure enables concepts to serve all the functions listed earlier. The epistemological, metaphysical and inferential representations support the epistemological, metaphysical and inferential functions respectively. 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 category 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 stimulus 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 when the agent is performing (symbolic) reasoning, the concept is accessed via the internal designator.

### 7.3.3 Summary

The main conclusion of this chapter is that it is necessary for an autonomous agent to represent categories by composite representations where the different components are chosen according to the function they should serve. However, since representation of categories to be used for categorization is the thing most often discussed in the literature, it also has been the main topic here. In particular, representations for epistemological categorization have been investigated. Since their relevance is uncertain, the metaphysical component will, in fact, not be discussed in any detail in the following chapters.

As we have seen, the AI community has already studied all of the well developed psychological models of category representation. The only approach that has not been implemented yet (at least to the author’s knowledge) 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 categories. 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 seems 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.

We have argued that it is necessary, at least for the epistemological component, to also be able to represent structural knowledge. However, as there are many different kinds of structural representations, the epistemological representation must be adapted to the agent’s perceptual system as well (i.e., it must be in a form that the perceptual system can use).

An issue that has not been dealt with, is the problem that arises when the number of concepts grows. Having thousands (or more) of concepts will turn the speed issue into an acute problem for the performance of the epistemological function. An efficient way of indexing the concepts is necessary. This seems to require a hierarchical organization of the concepts of the kind we have mentioned earlier. Some interesting work has been carried out regarding this issue (cf. [181]).



## Chapter 8

# Concept Acquisition

Finally, we have reached the stage where we are able to discuss how concepts can and should be acquired. We will consider theories of human concept acquisition and approaches to concept acquisition taken within AI (i.e., machine learning, computer vision, and pattern recognition) as well as the theoretical studies of what can actually be learned. In this chapter, as in the earlier, the different approaches will be evaluated qualitatively rather than quantitatively.

### 8.1 Human Concept Acquisition

According to Atkinson et al. [14], humans learn about categories in two different ways:

- by being explicitly taught
- by learning through experience.

Unfortunately, they do not elaborate this distinction any further, and it has been hard to find any other discussions concerning this topic.<sup>1</sup> However, it seems reasonable to believe that it is possible to be explicitly taught about categories both on the linguistic level (i.e., learning by description) and on a sublinguistic (perceptual) level (i.e., learning by acquaintance). Examples of learning on the linguistic level are when you learn something reading a book or by being told something by some kind of teacher. It seems likely that we learn metaphysical, and to some extent inferential, knowledge 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 actual exemplar of a category (cf. ostensive definitions).<sup>2</sup> Thus, it is primarily epistemological, but also inferential, knowledge that is learned in this way.

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

---

<sup>1</sup>In fact, there is not much written (lately, at least) about human concept acquisition. In the last decades the researchers in the field seem to have focused on representation and classification processes. Therefore, the main part of this section is the author's own interpretations and elaborations of the few notes on the topic that have been found.

<sup>2</sup>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 main categories above.



you know rather well, but this instance is different in some respect from the ones you have previously been acquainted with, 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 in advance. 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. Note that the input when learning through experience often is 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, perhaps more commonly, by asking questions (i.e., “Is this an elephant?” or “What features characterize an elephant?”). This type of learning seems to bear some 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. On the contrary, it is the author’s opinion that we use them all alternatively. Which kind of learning that is the appropriate one in a particular situation is, of course, to a great extent determined by the situation (e.g., whether there is a teacher present or not).

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 instances 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. In more technical terms, the learning system must be able to switch modes between learning and classification without destroying any previous learning. Moreover, we do not learn one concept at a time, concepts are rather acquired in *parallel*.<sup>3</sup> Yet another constraint is that the learning must be accomplished relatively *fast* in the sense that we are able to learn a fairly useful category representation just by encountering instances of the category on one or a few occasions. For instance, if we are hurt once by an animal of a (to us) unknown species, we are often able to recognize other animals of this species later, and infer that it is able to hurt us.

As Schank et al. [178] 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.

Theories of learning by being explicitly taught, however, do not have to specify the first process since it is assumed that this is done by the teacher (i.e., the category is already formed).

### 8.1.1 Theories of Concept Acquisition

As pointed out earlier, all our knowledge about categories cannot be innate. However, it is possible, and even plausible, that some knowledge about categories is innate.

---

<sup>3</sup>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.

Different researchers emphasize this to different degrees. Fodor's [68] theories of cognition, for instance, rely heavily on innate knowledge.

If it is not the case that all concepts are 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 by Solso [195] seems rather outdated, with its roots in stimulus-response psychology. It holds that the learning of a category representation is the result of (1) reinforcing the correct pairing of a stimulus with the response of identifying it as a category, and (2) non-reinforcing (punishment) the incorrect pairing of a stimulus with a response of identifying it as a category. This theory seems to cover only 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. For instance, it does not specify what part of the concept is being changed.

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." [14] 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 learns only one concept at a time. In fact, the hypothesis testing theory seems more like a model of learning by experimentation, like 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. Thus, this theory is consistent with the exemplar view of representation. However, since it seems implausible that we remember every instance we ever encountered, this simple version of the theory has to be modified in some way. Thus, several questions remains open, for example: How many, and which, instances should be memorized? Moreover, the strategy is only specified for learning by being explicitly taught. 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 its incremental nature and that it accounts for the acquisition of many concepts at the time. Thus, the exemplar strategy is the only theory of the three that has at least a chance of being adequate.

## 8.2 AI Methods for Concept Acquisition

The concept acquisition process of an autonomous agent is from a general point of view restricted by the environment in the same way as that of a human. Thus,

from the earlier discussion we can conclude that for artificial autonomous agents the concept acquisition process must be incremental and relatively fast, concepts must be acquired in parallel, and several methods must be employed simultaneously.

### 8.2.1 ML Methods for Concept Acquisition

In AI, several ways of learning about categories have been studied, the most predominant being:

- direct implanting of knowledge
- learning from examples
- learning by observation
- learning by discovery
- learning by deduction.

The following five sections will present these concept acquisition paradigms in greater detail.

#### Direct Implanting of Knowledge

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 category descriptions and the case when the descriptions are programmed directly into the system. 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. However, one should be careful when doing this since the category representation probably will to some extent reflect the programmer's conception of the category which may not necessarily coincide with what would be optimal for the system (cf. Part III of this thesis).

*Learning by instruction*, or learning by being told, is similar to direct implanting of knowledge in that the learner acquires concepts (explicitly described on the linguistic level) from a teacher, database, textbook or some other organized source. However, this form of learning, in contrast to direct implanting of knowledge, requires selecting the relevant information and/or transforming this information to a usable form.

#### Learning from examples

Learning from examples is by far the most studied type of learning in AI and can be seen as a parallel to learning by being explicitly taught. In this kind of learning the learner induces a category description from preclassified examples and 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 an instance of *supervised learning*. Thus, it is the teacher who decides when to create a new concept. From the classical viewpoint, the task for this type of learning can be seen as finding a concept definition (i.e., description) consistent with all positive examples but no negative examples in the training set.

Some of the systems learning from examples can be viewed as carrying out a search through a space of possible category 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 descriptions have all features specified, corresponding to individual instances. There are basically two strategies for searching the space of category descriptions. In the general-to-specific, or model-driven, strategy, one begins with the most general description as the hypothesis of the correct category description, and as new instances are encountered, more specific descriptions (i.e., hypotheses) are produced. The specific-to-general, or data-driven, strategy, on the other hand, begins with a very specific description, typically a description of the first instance encountered, 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 those based on *version spaces* [137], combine the two strategies. Since there is no inherent non-incrementality in this approach, it seems possible to make systems that learn incrementally based on this approach.<sup>4</sup>

A different kind of learning-from-examples system is the so called top-down induction of decision trees (TDIDT) systems [162]. These systems accept instances represented as lists of attribute-value pairs as input and produce a decision tree as output. TDIDT systems, which are model-driven, 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. A serious limitation is, however, their non-incremental nature. To incorporate new instances, the tree often has to be recomputed from scratch. Moreover, since they require feature-vector representations of the training instances, they are not suited for dealing with structural information.

The learning-from-examples systems typically learn just one concept at a time, without considering other known category descriptions. An exception to this is AQ11 [131, 129] by Michalski and his colleagues, which learns multiple concepts. Another exception is a system by Gross [79] that incrementally learns multiple concepts where the category description currently learned is constrained by the descriptions of the other categories. However, this system can be interpreted as learning by experimentation, since it is the system itself that selects the next instance to be analyzed from a given description space. This instance is then classified by an oracle. The introduction of an oracle being able to classify every possible instance (as is done also in some other kinds of systems) makes the learning easier but is an unrealistic assumption when regarding autonomous agents contexts. This algorithm, as well as some of the other search-based algorithms, seems to be consistent with the hypothesis testing theory of human concept acquisition.

While all systems described above have employed classical representations, we will now look at some systems that use non-classical representations. Kibler and Aha

---

<sup>4</sup>However, some systems, version spaces for instance, have several competing hypotheses at some stages in the learning process. Having several hypotheses makes it difficult to use the concept and requires more memory space. Nevertheless, the memory requirements of such systems are substantially less than for systems that must memorize all instances, such as Winston's [213].

[97] describe three algorithms that learn from examples using an exemplar representation of categories. The *proximity* algorithm simply stores all training instances. The *growth*, or additive, algorithm stores only those training instances that would not be correctly classified if they were not stored. These two algorithms are incremental in contrast to the third, the *shrink*, or subtractive, algorithm. The shrink algorithm begins by placing all the training instances into the category representation, and then continues by testing each instance in turn to see if it would be correctly classified using only the remaining instances. Nagel [147] 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 category description to be used for assimilating new instances.<sup>5</sup> De la Maza's PROTO-TO system [51] also learns incrementally from examples but uses a probabilistic representation. It groups the instances according to their categories and then builds a prototype for each category. The prototypes are then augmented, weighting each attribute in order to form a probabilistic representation.

There are also subsymbolic approaches to learning from example. The simplest type of neural network for this task is the (single layer) *perceptron*. It consists of a number of input nodes, each corresponding to a feature in a feature vector, which are connected to an output node. The perceptron is able to classify the input vector into either of two categories and can be used for both continuously valued (dimensions) and binary (features) inputs. A detailed analysis of the capabilities and limitations of perceptrons is provided by Minsky and Papert [135]. It is, for instance, proved that if a perceptron can classify a series of inputs, then it is also able to learn that classification. A severe limitation is, however, that perceptrons can only classify linearly separable categories (i.e., categories that can be separated by a straight line in a 2-dimensional feature space).

The multilayer perceptron is a generalization of the single layer perceptron that is able to learn classifications of categories that are not linearly separable. It has one or more additional layer(s) of nodes, called hidden layer(s), between the input and the output nodes. A two-layer perceptron can form any convex region in the feature space and a three-layer perceptron can form arbitrarily complex regions. The algorithm most often used for learning multilayer perceptrons is called the *back-propagation algorithm* [175]. Although it has not been proven that this algorithm can learn every classification the network can represent, it has been shown to be successful for many problems. A disadvantage is, however, that the algorithm is non-incremental.<sup>6</sup> Like most other neural network approaches the back-propagation algorithm has scaling problems, which means that, when the number of training examples and/or the number of categories to be learned gets larger, the algorithm will be too slow and use too much memory.

### Learning by observation

In contrast to learning from examples, a system performing learning by observation has to form categories by itself. Thus, it is the learner that decides when to create

---

<sup>5</sup>How the prototypes are learned in the first place is not described in the material that, for the moment, is available to me (i.e., [146]).

<sup>6</sup>However, some variants of the algorithm are claimed to behave incrementally.

new concepts. In learning from examples the categories were formed beforehand and there was a teacher present who provided examples of them (where the category membership was explicitly stated) to the learning system. Thus, the learning process was supervised. In analogy, learning by observation is an instance of *unsupervised learning*. Typically, the learner is given a number of descriptions of entities (with unknown category membership). It groups the entities into categories based on their features, a process often referred to as *aggregation*. When this is done, the system creates descriptions of the categories, a process often called *characterization*.

Tasks similar to the aggregation task has been studied for a long time within statistics under the labels *cluster analysis* and *numerical taxonomy*. In these contexts the observations, or instances, are typically described by a number,  $n$ , of numerical features and treated as points in an  $n$ -dimensional space. Some clustering algorithms build hierarchies of categories, i.e., hierarchical methods, whereas other algorithms cluster the observations only at one level, i.e., optimization methods. It is common to further divide the hierarchical methods into agglomerative and divisive methods. Agglomerative methods work in a bottom-up fashion, beginning with joining similar separate observations together to form small clusters. By recursively forming larger and larger clusters, the process eventually halts when all observations belong to a single universal cluster. In this way the algorithm builds, step by step, a hierarchy of clusters. Divisive methods work in the opposite way, beginning with a single universal cluster and then repeatedly breaking it into smaller and smaller clusters until only single observations exist in each cluster. Optimization methods, on the other hand, form clusters at a singular level by optimizing the clustering according to user-provided information such as the number of clusters to form and desired cluster size. More details on clustering algorithms are provided by Jain and Dubes [93].

The clustering algorithms described above all use some kind of similarity measure for the aggregation task which, in turn, depends on some kind of distance metric. As pointed out earlier, such a metric has several disadvantages, for instance that 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 may take into account totally irrelevant features. An interesting optimization clustering technique that does not use a distance metric is suggested by Matthews and Hearne [122]. The clusterings are instead optimized on the intended function of the clustering, which, according to the authors, is the prediction of unknown feature values. Thus, this approach aims at maximizing the utility of the clustering.

The characterization task, creating descriptions of the categories, is in principle equivalent to the task of learning from examples as described above. This suggests that one way to perform learning from observation would be to employ a statistical clustering algorithm for the aggregation task and then use one of the ML-algorithms presented in the last section to create descriptions of the categories. There are, however, some problems associated with such an approach. Besides being limited to numerical feature values, the aggregation step would be totally independent from the characterization step, not taking into account the language used to describe the resulting concepts. This would result in clusters that may not be well characterized (i.e., comprehensible by humans) in the chosen description language. Rather than relying on a pure metric similarity measure Michalski and his colleagues [132] introduced the notion of conceptual cohesiveness described in Section 6.2.2.

Systems which integrate the aggregation and characterization steps are commonly

called *conceptual clustering* systems. Some of the most influential are CLUSTER/2 [132, 197] and RUMMAGE [62] which both characterize the formed categories by classical descriptions. Although these systems learn in a non-incremental fashion, incremental systems that use classical concept definitions, like UNIMEM [108], exist. As should be clear from above, conceptual clustering systems form concepts in parallel (i.e., many at the time). Moreover, while other types of systems usually learn concepts at a single level, most conceptual clustering systems structure the created concepts into taxonomies (cf. hierarchical clustering methods). Whereas conceptual clustering systems typically do not form structural concepts, CLUSTER/S [198], which is version of CLUSTER/2, do.

An example of an approach that uses non-classical representations is the PLANC system by Musgrove and Phelps [145] that 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 a prototype (i.e., a hypothetical average member). Whereas this system is non-incremental, Fisher's COBWEB [63, 64] acquires concepts in an incremental fashion. COBWEB builds a probabilistic concept tree. As an evaluation measure of clusterings in the aggregation task, it uses *category utility* instead of a distance metric. This measure was originally developed by Gluck and Corter [77] as a means of predicting the basic level in human taxonomies. Thus, one can interpret COBWEB's strategy as trying to form basic levels on every level, beginning at the top of the tree. It is similar to Matthews and Hearne's approach in that it tries to maximize the predictive ability of the clustering. LABYRINTH [201, 202] is an adaption of COBWEB which is also able to handle structural descriptions.

Kohonen's [101] self-organizing feature maps is a neural network approach to conceptual clustering. These networks consist of a number of input nodes that are connected to each of the nodes in a (typically) two-dimensional array of output nodes. The result after the learning phase, which is based of a winner-takes-all principle, is a mapping where similar inputs are mapped to nearby output nodes where each output node, or rather neighborhood of nodes, represents a category. These networks are able to handle both continuously valued and binary inputs. In contrast to the other conceptual clustering systems presented here, the self-organizing feature maps do not form a hierarchy of categories.<sup>7</sup> The network decides by itself the number of categories to be formed.

Another neural network approach to unsupervised learning is Grossberg and Carpenter's [40] adaptive resonance theory (ART) networks. The structure and function of these networks are to a higher degree than most other neural networks based on biological and behavioral data. The desire to replicate learning in humans, who actually are autonomous agents, has resulted in a network with some interesting features. For instance, the ART networks learn in an incremental fashion. Another positive aspect is that when fast-learning<sup>8</sup> is used, it requires, in contrast to most other networks, only one pass (at most) through the training set to learn the category representation. Furthermore, the network has proven to be stable and does not suffer from any convergence problems (e.g., local minima). While ART-1 only takes binary input, ART-2 can deal also with continuous input. However, it does only form categories at one level and to decide the number of categories to be formed (or, rather,

---

<sup>7</sup>At least not the kind of hierarchies that have been described here.

<sup>8</sup>The ART networks have two training schemes often referred to as fast and slow learning.

the metrical size of the categories) , one must choose a vigilance threshold.

### Learning by Discovery

Learning by discovery is, just like learning by observation, a kind 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 accepting the descriptions of instances given to it.

The most famous system of this kind is Lenat's AM system [110, 111]. 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 worked (or appeared to work) 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 for this 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, if such a system is to be more dynamic, it must also be able to reason and manipulate with its heuristics. For a more comprehensive discussion of this topic, see Lenat and Brown [112]. Another well-known system that learns from discovery is GLAUBER [107].

### Deductive Learning

In deductive learning, the learner acquires a category description by deducing it from the knowledge given and/or already possessed (i.e., background knowledge). The most investigated kind of deductive learning is *explanation-based learning* (EBL) [139, 52] that transforms a given abstract category description (often based on non-perceptual features) to an operational description (often based on perceptual features) using a category example (described by operational, or perceptual, features) and background knowledge for guidance.

The standard example of EBL concerns the category "cup". In this example the abstract category description is a classical definition that says that a cup is an open, stable and liftable vessel. 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 perceptual features (e.g., light, has a handle) and the operability criterion that the category description must be expressed in terms of the perceptual features used in the example, the EBL-system produces a description of the category "cup" that includes the facts that a cup is light, has a handle, has a flat bottom, and so on. Seen in the light of the core versus identification procedure view of concepts described in Section 6.1.1, we can interpret the function of an EBL system as taking the core of the concept as



input and producing an identification procedure as output.

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 nothing but a reformulation of category descriptions, since an abstract description of the category is given to the system. Thus, no new categories are created.

### 8.2.2 Computer Vision Approaches

The models learned by vision systems are intended to be used for perceptual recognition tasks and can thus be interpreted as being epistemological representations. However, as mentioned earlier, they are often interpreted as models of particular objects rather than categories.

In the past, there has only been a few studies of the learning of concepts (i.e., object models) by direct perception of objects. However, the AAAI Fall Symposium on Machine Learning in Computer Vision in 1993 [2] showed that this state of affairs is changing. In this symposium a number of approaches were presented covering both the more traditional 2-D and 3-D models as well as characteristic views models. However, most of the research was still in progress and had not yet been applied to realistic situations (i.e., receiving data directly from video cameras, or other sensors, that perceive non-laboratory scenes). Typically, the systems were given line-drawings as input<sup>9</sup> and/or could not cope with occluded objects.

#### 2-3-D Model Approaches

One of the earliest studies of the learning of object models from real images was made by Connell and Brady [47]. Their system, the learning component of which is a modified version of Winston's ANALOGY [214] program, learns semantic network representations from examples. Input to the system is a set of grey-scale images of real objects such as hammers and airplanes. However, the system is limited to learning 2-D shapes and only works well with objects composed of elongated pieces.

Another approach using semantic network representation is suggested by Dey et al. [53]. It is an unsupervised system that incrementally acquires a hierarchy of concepts, but seems to have the same weaknesses as Connell and Brady's system. In general, it is interesting to note that so few, if any, of the learning vision systems use 3-D volumetric representations.

An interesting idea within this paradigm is, however, the adaptive feature extraction framework presented by Bhandaru, Draper, and Lesser [23]. It integrates feature extraction and learning object models, instead of treating these as separate processes.

#### Characteristic View Approaches

These systems are based on the principle that having enough 2-D views of an object is equivalent to having its 3-D structure specified. As an example, Pope and Lowe [158] present an approach that learns a set of characteristic views from pre-classified examples. An incremental conceptual clustering, similar to COBWEB, is used to construct

---

<sup>9</sup>This applies more to the 2-D and 3-D than the characteristic view approaches which were often given real images.

(identify) the characteristic views from the examples. When recognizing (classifying) new images a probabilistic similarity metric is used to compare the image with the characteristic views. Another approach is suggested by Poggio and Edelman [157], who have implemented a neural network that learns to transform an image from any viewpoint into a standard view. From this standard view it is easy recognize which category the object belongs to. In the article, however, they only tested the network on simple line-drawings.

An approach of learning appearance models, a continuous version of characteristic views, from examples was presented in Section 7.2.3. While it seemed to have some interesting features, there are some problems with the approach that makes it problematic to apply directly. For instance, it is probably difficult to obtain all the images necessary for learning an adequate representation. Moreover, it is assumed that objects are not occluded by other objects and that they can be segmented from the background of the scene. Furthermore, the learning is batch-oriented and it is not clear whether it could be performed in an incremental fashion.

In general, it seems very hard to introduce symbolic knowledge (i.e., direct implanting of knowledge) to this kind of systems. The representation of objects, or categories, is on a sub-symbolic level and, moreover, heavily dependent on the system's sensors.

### 8.2.3 Pattern Recognition

Pattern recognition methods are often divided into *statistical* and *syntactical*, or structural, methods. Whereas we have already described some statistical methods, the syntactical approach will provide a new view on the problem of concept representation and acquisition.

A fundamental question that should be answered before we continue, is whether a pattern can be interpreted as a concept (or, rather, as a description of an instance). Actually, we made such an interpretation when we discussed neural networks. Neural networks (as well as some other systems described above) are in fact a kind of pattern recognizers. The features that are used to characterize an object constitutes a pattern. Thus, if a pattern recognition system can recognize a pattern corresponding to an object, it can ideally also recognize the object, and consequently the system would implement the epistemological (or metaphysical) function.

#### Statistical Pattern Recognition

Statistical methods [56] are based on statistical studies of the input (i.e., feature vectors) in order to recognize patterns. There are two major types of methods, parametric, or Bayesian, and non-parametric.

The aim in *parametric* methods is to decide, on the basis of a model for the distribution of the instances of each category, to which category an unknown instance has the greatest probability of belonging. The decisions are based on statistical decision theory. In the basic versions of the methods, it is required that the statistical distributions,  $p(v|c)$  (where  $v$  is the input vector, the instance, and  $c$  is the category), are known. Since this assumption is too strong in most cases, a parametric system often has to *estimate* the distributions from the training instances. If the general form of the distribution is known (e.g., Gaussian), this task reduces to the estimation

of a finite number of parameters. The disadvantages with this approach are that the distribution of instances within the category often is not sufficiently regular, and that it requires a large number of training instances [133].

The *non-parametric* methods, on the other hand, do not assume that the form of the underlying statistical distribution is known, rather they try to estimate it using the given training instances only. Their aim is to find the boundaries of the different categories in the description space, so that a series of simple tests will suffice to categorize an unknown instance into one of the known categories. One way of doing this is to learn a *linear classifier*, i.e., to find a hyperplane that divides the description space into two parts (in the two-category case). The perceptron, described in the last chapter, is an example of a linear classifier. This approach demands, of course, that the categories be linearly separable. Another popular method is that of *nearest neighbor* in which the unknown instance is categorized into the category of its nearest neighbor. In fact, this is exactly the same algorithm as the instance-based proximity algorithm described above.

The methods described above are all supervised (cf. learning from examples) algorithms. However, statistical clustering algorithms as described on page 71, are sometimes also regarded as pattern recognition methods. A more sophisticated algorithm, called AUTOCLASS is provided by Cheeseman et al. [43]. It is based on the parametric (Bayesian) approach and “automatically determines the most probable number of classes, their probabilistic descriptions, and the probability that each object is a member of each class.” Thus, rather than assigning the instances to specific categories, it derives probabilities of the instances belonging to a category. This is not to be confused with probabilistic representations, as described in Chapter 7, which vary in their degree of membership, not their probability of membership. The algorithm need not be told the number of clusters to form,<sup>10</sup> and need, according to Cheeseman et al. [44], no ad hoc similarity measure or clustering quality criterion. Another interesting feature of AUTOCLASS is that it can easily be adapted to perform supervised learning. On the other hand, a clear disadvantage is that the algorithm is inherently non-incremental. However, an incremental algorithm that also relies on the Bayesian method, is presented by Anderson and Matessa [8]. In contrast to AUTOCLASS, this algorithm assigns each instance to a specific category.

A disadvantage with statistical pattern recognition methods in general, is that they assume feature vector representations, and are thus not suitable for structural information. Syntactical methods, on the other hand, emphasize this kind of information rather than the metric properties emphasized by the statistical methods.

### Syntactical Pattern Recognition

In syntactical methods patterns (i.e., instances) are typically viewed as sentences in some formal language. Consequently, categories are represented by grammars in that language. The recognition of an instance as a member of a category is accomplished by parsing the sentence corresponding to the instance to determine whether or not it is syntactically derivable using the grammar corresponding to the category. To learn a concept then corresponds to inferring a grammar from a number of sentences from which all sentences can be derived.

---

<sup>10</sup>It will arise naturally because of an optimization of a trade-off situation between forming small and large clusters

There exist many different kinds of grammar inference algorithms (cf. [133]). However, most of these learn only from examples (supervised learning) and just one concept at a time. Moreover, since traditional formal languages typically are used, they learn concepts that correspond to classical definitions. Despite these negative aspects, it may be fruitful to try to apply, or combine, the theories developed within this paradigm to (with) the ML framework. As pointed out by Honavar [90], a particularly interesting approach would be to use *template matching* wherein an instance is matched to one or more stored instance(s) for each of the categories. The instance is then assigned to the category with the best match. This approach is very similar to the instance-based algorithms in ML (and to the non-parametric statistical method, called nearest neighbor), but is able to deal also with structural information. As with other algorithms of this kind, some kind of similarity measure is needed when matching two instances.

A serious limitation of syntactical methods is that representation of instances by normal sentences (i.e., linear strings) permits the encoding of only a single structural relation between the representational primitives; a primitive can only precede or succeed another [90]. To represent, for instance, complex structural relationships of 2- or 3-dimensional objects, more powerful grammars are needed, such as web grammars and tree grammars. A generalized similarity measure, applicable to arbitrary structured patterns, is presented by Honavar [89].

### 8.3 What can be learned?

In the field of *formal learning theory* the term *inductive inference* has come to denote the process of hypothesizing a general rule (e.g., a classical concept definition) from positive examples of the rule. Thus, inductive inference addresses problems similar to that of concept learning from examples.

Although there exist several paradigms, or frameworks, within formal learning theory we will here concentrate on the two most dominant. Until the mid eighties most of the research was carried out within the paradigm formulated by Solomonoff [194] and established by Gold in his fundamental paper [78]. This paradigm will here be referred to as Gold's paradigm. In 1984 Valiant [209] formulated a paradigm that was more closely related to the ongoing research on concept learning in ML. In short, one can say that Gold's approach addresses the question of whether there exists an algorithm for learning the concept definition (i.e., computability), whereas Valiant's approach addresses the question of whether there exists an *efficient* algorithm (i.e., computational complexity).

To define an inductive inference problem, the following items must be specified [9]:

- the class of rules being considered
- the hypothesis space
- for each rule, its set of examples, and the sequences of examples that constitute admissible presentations of the rule
- the class of inference methods under consideration
- the criteria for a successful inference.

The class of rules is usually a class of functions, boolean expressions, or formal languages. The hypothesis space is a set of descriptions such that each rule in the class has at least one description in the hypothesis space.

### 8.3.1 Gold's Paradigm

Gold's paper presents an investigation, motivated by the psycholinguists' study of the acquisition of grammar by children, concerning to what extent different classes of languages can be learned from positive examples only.<sup>11</sup> Other names for this paradigm are *the identification paradigm* [152] and *grammatical inference* [106]. The last name suggests that it is related to learning in syntactical (structural) pattern recognition. In fact, this is to a great extent a theoretical treatment of the syntactic approach. However, Blum and Blum [28] later transferred the posing of the problem to identification of functions instead of grammars.

In his paper Gold introduces *identification in the limit* as a criteria of success, which can be formulated as follows. Suppose that  $M$  is an inductive inference method attempting to describe some unknown rule  $R$ . If  $M$  is run repeatedly on larger and larger collections of examples of  $R$ , an infinite sequence of  $M$ 's conjectures is generated. If there exists a number,  $n$ , such that the  $n$ th conjecture is a correct description of  $R$  and that all the conjectures that follow are the same as the  $n$ th, then  $M$  is said to identify  $R$  (correctly) in the limit on this sequence of examples. Note that Gold here views inductive inference as an infinite process, and that  $M$  cannot determine whether it has converged to a correct hypothesis (conclusion).

*Identification by enumeration* is an example of an inference method. It systematically searches through the space of possible rules until one is found that agrees with all the data so far. Suppose that a particular domain of rules (category descriptions) is specified, and that there is an enumeration of descriptions  $(d_1, d_2, \dots)$  such that each rule in the domain has one or more descriptions in this enumeration. Given any collection of examples, we just have to work through the list of descriptions until we find the first description that is compatible with the given examples and then conjecture it. The method is guaranteed to identify in the limit all the rules in the domain if the following conditions are satisfied:

- A correct hypothesis is always compatible with the examples given.
- Any incorrect hypothesis is incompatible with some sufficiently large collection of examples (and with all larger collections).

The method is computable if the enumeration  $(d_1, d_2, \dots)$  is computable and if it is possible to compute whether a given description and a given collection of examples are compatible.

However, since the issue of computational feasibility has not been central to this paradigm, many of the positive results (i.e., that something is learnable) have relied on algorithms, such as identification by enumeration, that are intractable with respect to time or/and space. Moreover, many of the negative results have been due to the

---

<sup>11</sup>If one wants to study the actual learning process of natural language grammar by children this model seems to be a poor one, since it only takes into account the syntactic component and does not bother about the semantic and pragmatic issues. Furthermore, it does not seem plausible that this type of learning is based only on positive examples. (Although Chomsky and others believe it is.)

fact that the learning domains have been too general to allow any algorithm to (ever) distinguish the target concept from among other possible hypotheses [155].

### 8.3.2 Valiant's Paradigm

Valiant's paradigm, often referred to as *Probably Approximately Correct (PAC) learning*, has become more popular than Gold's since it addresses more of the requirements that are typically placed on a learning algorithm in practice. Since its criteria of success is that a good approximation of the target concept should to be found with high probability (rather than exact identification), it allows greater emphasis on computational efficiency.

Valiant's framework considers the learning of category descriptions in the form of Boolean vectors of features. The learning system is given positive and negative examples of the target concept. The learner must produce, with at least the probability  $1 - \theta$ , a category description that incorrectly classifies future examples with probability lesser than or equal to  $\varepsilon$ . If there exists an algorithm that can accomplish this task for any target concept in the concept class in time polynomial in the size of the target category description,  $\frac{1}{\theta}$ , and  $\frac{1}{\varepsilon}$ , then the class is said to be learnable.<sup>12</sup> An example of a concept class that has proven to be learnable is the class of conjunctions of boolean variables. For more results PAC-learnability see, for instance, Kearns et al. [96].

### 8.3.3 Critique of Current Formal Learning Theory

The obvious critique of current formal learning theory is that it only regards classical concept definitions. Moreover, most approaches do only regard learning from examples. Pitt and Reinke [156], however, have considered unsupervised learning. They have developed a formalism for analyzing conceptual clustering algorithms. As criteria for success they have chosen the ability to efficiently produce a clustering that maximizes a given objective function based on individual cluster tightness and overall distance between clusters. They show that under a wide variety of conditions, the agglomerative-hierarchical algorithm can be used to find an optimal solution in polynomial time.

Although Valiant's paradigm is more closely related to the current research in ML than Gold's, it has been criticized for not being so to a sufficient degree. For instance, Buntine [39] claims that it can produce overly-conservative estimates of errors and that it fails to match the induction process as it is often implemented. Thus, while being possibly interesting from a theoretical viewpoint, the current approaches to formal learning theory are only of limited interest in the context of autonomous agents.

## 8.4 Conclusions

The issue of concept acquisition, in contrast to functional and representational issues, has been more intensively studied in AI than in the Cognitive Sciences. In fact, for all the psychological models presented in this section, there exist corresponding AI

---

<sup>12</sup>This is a simplification of Valiant's original formalization. My hope is, however, that it probably approximately resembles his central ideas.

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 most recognized of the existing 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, the insight that approaches to concept acquisition by autonomous agents are constrained by several demands. For instance, they must be incremental, be able to learn multiple concepts, be able to learn relatively fast, and apply several methods simultaneously. In the following sections we will discuss to what extent these demands have been met by existing AI systems and what could be done to meet them.

#### 8.4.1 Incrementality

In early machine learning research, most systems were non-incremental in the sense that they had two separate processes, an initial training phase followed by a classification phase. An autonomous agent, however, should be able to learn new concepts throughout its life-cycle (i.e., it should never stop learning). It would, of course, be possible to let the agent use a non-incremental learning system, memorizing all previously encountered objects, and recomputing the whole learning procedure every time a new object is encountered. For practical reasons, such as time and memory requirements, this is not an adequate solution to the problem. Instead, the learning system must be able to switch modes between learning and classification without destroying any previous learning. In recent years, however, many incremental systems have been constructed. For each approach to concept acquisition there often also exist an incremental version, unless the approach is inherently non-incremental.

Incremental learning is also the basic feature of the *active agent* paradigm suggested by Pachowicz [153]. In addition, he suggests a closer integration between vision and learning than just finding bridges between stand alone vision and learning modules. In the article he points out some important problems along with some hints for the solution of these.

#### 8.4.2 Learning Multiple Concepts

There exist many systems that learn more than one concept 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 is typically not used to constrain the hypothesis space. How this should be done seems like an important area of research (especially if one is interested in the metaphysical functions of concepts).

#### 8.4.3 Fast Learning

The learning speed of algorithms has been a topic that for obvious reasons has received considerable attention. Examples of empirical comparisons of different algorithms are, for instance, [212, 140, 211, 65]. The general conclusion of these studies is that symbolic methods (e.g., ID3) learn faster than subsymbolic (e.g., back-propagation).

However, the number of training examples used in these comparisons is typically very large (100 – 10000 instances). From an autonomous agent point of view, it

is equally interesting to study learning from just a few training instances of each particular category. It seems that an empirical comparison of this kind has never been made, but a qualitative statement is provided by Fisher and Pazzani [66] who write that:

... specific-to-general learners can more quickly exploit relevant information for purposes of prediction than can their general-to-specific counterparts. For example, a system can ideally distinguish mice from men with high, but not perfect, accuracy after a single example of each.” (p.30).

Thus, most neural networks (with the exception of ART-networks) learn too slowly in the sense that they need many instances (and epochs) to learn a fairly good representation. The reason is that since the weights in neural networks are often randomly chosen, the behavior of the net is unpredictable in early stages leading to misclassifications.

#### 8.4.4 Integrating Different Learning Strategies

Still, it is the requirement that several methods of learning must be applied simultaneously that 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 (cf. Lebowitz [109]).

In particular, learning a composite category representation as outlined in the last chapter seems to demand a large variety of learning mechanisms. Let us discuss for each of the components how they could be learned by an autonomous agent. While it is clear that the agent does not learn all the components at exactly the same time, the concept must be created by initially learning one, or possibly two, of the components.<sup>13</sup> It seems natural to assume that the agent either learns the name or the epistemological component first. Which of these is actually learned first has been, and probably is, a controversial question within *developmental psychology*. In most early research it was hypothesized that linguistic input was the fundamental source of information in the concept formation process (i.e., first the words are learned, then the other components). In contrast, Nelson [149] has argued that concepts originate from interaction with the physical world. Then, the words that fit to these concepts are learned. A more neutral position is taken by Bowerman [30] who suggests that: “there appears to be a complex interaction in word acquisition between children’s own predispositions to categorize things in certain ways and their attention to the words of adult language as guides to concept formation.” This view is in line with the approach that will be described below, which allows either of the components to be learned first. In the same article, Bowerman also argues that children’s concept formation processes are quite similar to adults’. This is one of the reasons that there has not been put much weight on developmental psychology in this thesis.

Let us, however, begin with the epistemological component necessary for recognizing instances of the category. There seem to be two different ways of learning this component: (1) by observation (i.e., through experience), or (2) from examples (i.e., by being explicitly taught). The first of these is probably the most basic where the agent

<sup>13</sup>It would have been meaningless to have a concept if all the components were “empty”.



through perceptual interaction with the environment autonomously identifies a new category and creates an epistemological representation.<sup>14</sup>

Learning from examples, on the other hand, requires that another agent, a teacher, already knows the name of the category and how to recognize its instances, and is willing to communicate this knowledge. Since this kind of learning also involves another component, the external designator (i.e., the name of the category), there are two possible situations: (a) the agent already knows the name, or (b) it does not. If the name is known, the task will be to associate this already learned component with the new epistemological representation and make them parts of the same concept. If the name of the category is not known to the agent, it must learn both the external designator and the epistemological component. Whereas (1) and (2b) correspond to the traditional ML-paradigms of learning by observation and learning from examples respectively, (2a) seems not to have received any attention within the ML-community.

When learning the external designator we seem to be restricted to learning from examples (by being explicitly taught). There are two possibilities in this case also: (1) the agent already has an epistemological component (learned, for instance, by observation), or (2) it has not. The second case corresponds to (2b) above, whereas the first case bears some resemblance to (2a) in that what is actually to be learned is a connection between the epistemological component and the external designator. Some experiments in this direction have been carried out by Schyns [180]. He has constructed a modular neural network that uses an unsupervised network (i.e., a self-organizing map) to form categories (and representations of them) and a supervised (i.e., an auto-associator) to learn their names. A third way of learning the name is by direct implanting of knowledge (i.e., pre-programming) or learning by being told. Regarding the epistemological representation, however, this approach may not be a very good idea for reasons explained in Part III.

The internal designator, on the other hand, need not to be learned, since it is the agent itself that decides what the category will be called internally. It is expected that the agent will invent this name when the concept is originally created.

It seems that in order to learn inferential knowledge through experience, the agent must already have an adequate epistemological component. That is, it must be able to recognize instances in order to detect more general rules regarding the category's members. On the other hand, to learn at the linguistic level, for instance by being told, it would suffice to have the external designator. It seems that neither of the traditional ML-paradigms can be applied to learn the inferential component. However, it seems possible to develop novel approaches to both supervised and unsupervised learning of inferential knowledge. (We will not waste much space discussing the metaphysical representation. In realistic settings, it seems that only direct implanting of knowledge or learning by being told are applicable.)

As indicated above, which types of learning that are adequate to integrate also depends heavily on the scenario in which the agent works. As mentioned in Chapter 2, 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 categories. Apart from this, it seems to be limited to unsupervised learning, such as learning by observation. In the case where other agents exist, there is the possibility of supervised learning,

---

<sup>14</sup>It is, of course, impossible to learn the name of the category just by observing instances.

such as learning from examples, in addition to learning by observation and direct implanting of knowledge. Thus, for this kind of agent, an integration of learning from examples and learning from observation, possibly in ways suggested above, would be fruitful.

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 implement such systems able to function in 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 categories. Thus, a problem-solving agent may benefit from using EBL. However, as pointed out earlier, EBL systems do not form any new categories. The actual category formation step is when the high-level description is created during problem solving. This is not a very well studied topic, deserving more attention. In addition, it may be the case that a new kind of high-level representation component (instead of the metaphysical) is needed for this task. EBL may then be used to infer a representation from this component that can support the epistemological function. However, as Lebowitz has pointed out [108], it is questionable whether there exist real-world situations where EBL can be applied, i.e., where the agent possesses all the background knowledge required to make the transformation into a low-level description. The problem is that the EBL-algorithm requires that the background knowledge is complete and consistent (in the more general senses of these terms). As pointed out by Honavar [90], an interesting approach to solving this problem would be to treat the background knowledge as though it was non-monotonic.

#### 8.4.5 Other Issues

One of the key problems for a learning system (which, for instance, integrates learning from examples and learning by observation) is to decide when to create a new concept. The learning system needs to know when it encounters an instance of an unknown category. As described in Part III of this thesis, the only way of doing this is, according to Smyth and Mellstrom [193], to learn *generative*, or *characteristic*, category descriptions that try to capture the similarities between the members of the category instead of learning *discriminative* descriptions that are representations of the boundaries between categories. The difference between these kinds of descriptions is illustrated in Figure 8.1. It shows some instances of three known categories ( $\star$ ,  $\bullet$ , and  $\diamond$ ), and the resulting category boundaries of the concepts learned by a system using discriminative descriptions (to the left) and by a system using characteristic descriptions (to the right). In this case, a member of an unknown category ( $\otimes$ ) will be categorized wrongly by the discriminative descriptions, whereas it will be regarded as a member of a novel category by the characteristic descriptions. ID3, back-propagation, and nearest neighbor algorithms, which are examples of systems learning discriminative representations, would then not be adequate for this kind of learning. Systems that learn characteristic category descriptions are, for instance, some logic-based and prototype-based algorithms. Note that this is not just a repre-

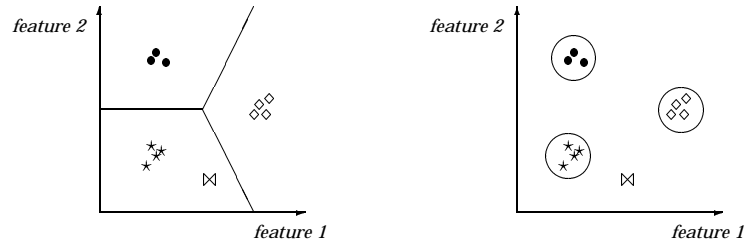


Figure 8.1: Discriminative vs. characteristic category descriptions.

sentational issue since it is also dependent on the learning and use of the category description.

A problem that has not been addressed in this chapter 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 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 (cf. the problem of small disjuncts [87]).

## Chapter 9

# Conclusions and Further Research

The goal of this part of the thesis has been, besides that of reviewing the research within the cognitive sciences and AI on different aspects of concepts, to investigate the issue whether psychological and philosophical 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 in the previous chapters, it is from more fundamental aspects than acquisition that influence from the cognitive sciences has the potential of being most fruitful.

In this chapter we will summarize the conclusions from the previous chapters and make some suggestions for further research based on these conclusions. Since this part of the thesis should be seen as an effort to establish a foundation for further research, the results, or conclusions, are formulated in terms of new insights and ideas rather than new algorithms.

### 9.1 Conclusions

Based on the initial general discussion about autonomous agents, a categorization scheme was suggested that was based on the distinctions regarding whether the agents were situated or not, and whether they were embodied or not. It was concluded that when developing new approaches for physical autonomous agents that are supposed to work in real-world domains, it is important to regard the agent as being situated and embodied from start. The traditional (i.e., deliberative) and the behavior-based (i.e., reactive) approaches were then compared with the result that a hybrid approach is preferable since both high-level deliberative reasoning and low-level reaction on perceptual stimuli seems necessary. A novel framework for this kind of hybrid agents is presented in Part II of this thesis.

In the following chapter we discussed world modeling and concluded that, although they are very difficult to acquire because of the necessary signal to symbol transformation, world and environment models (i.e., explicit knowledge about the world) are vital for deliberative, and hybrid, agents. To bridge the gap between signals and symbols, closer integration of learning and vision systems is necessary. However, the environment model does not need to be at the detailed level often assumed by the proponents of the traditional view. Rather, it should be at a high level of abstraction, used only to guide the agent's behavior. We also identified concepts as one of the most important primitive entities of which world and environment models

are built. Moreover, we concluded that although reactive agents do not have explicit knowledge about the world, they need to have concepts in one form or another.

In Chapter 4, we tried to make explicit what it actually means to have a concept, but suggested that a more appropriate question would be to ask which functions a concept should serve. While some cognitive science literature discusses this topic, it has hardly ever been discussed in the AI literature. In Chapter 5, six classes of functions of human concepts were identified: stability, cognitive economical, linguistic, epistemological, metaphysical and inferential functions. All of these proved to be desirable also for the concepts of artificial autonomous agents (with a possible exception for the metaphysical function).

We then discussed the nature of categories, which are the entities that the concepts are supposed to represent. AI researchers have, or at least have had, 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 as most previous AI-systems have done. It is also important to make a distinction between natural and derived categories since they must be acquired in different ways. Natural categories, natural kinds in particular, are typically formed by observing the external world and grouping similar objects together, whereas derived categories arise during internal problem solving activities. As a consequence, concepts corresponding to natural categories are probably best learned by similarity-based algorithms, whereas derived categories need top-down algorithms. EBL is, in a sense, a top-down approach, but does not address the problem of *formation* of concepts. In this chapter we also discussed where the features used to describe objects actually come from; are they innate (cf. categorical perception) or are they learned? That is, should an autonomous agent's feature detectors be preprogrammed or learned? At the present stage of AI research, it is difficult answer this question univocally. A related, and equally complicated issue, is the problem of how to deal with the concept of similarity.

As for the representation of categories, we concluded that a single and simple representation does not suffice to account for all the functions that we want concepts to serve. Thus, an autonomous agent must have a complex, or composite, category representation structured according to the desired conceptual functions. A suggestion for such a representation scheme is presented in Part IV of this thesis. The suggested conceptual structure has an epistemological component for perceptual (i.e., normal) categorization and an optional metaphysical component for more "scientific" categorization. As we have seen, it seems that some kind of prototype-based representation also able to represent structural knowledge probably will be the best alternative for the epistemological component, whereas a logic-based classical representation seems to be 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 inferential component. How this should be represented has not been discussed in detail, but in the light of past (and most current) AI-research some kind of logic-based, possibly probabilistic, representation language seems a natural choice. Finally, to support stability and linguistic functions, the concept structure should also include an internal and an external designator.

Regarding the actual acquisition of the different parts of this structure, it seems that the agent has to rely on learning from examples (if there is some kind of teacher available), learning by observation and some method for forming derived (top-down) categories. Learning from discovery seems too difficult for an agent in a real-world

domain. Moreover, the learning must be incremental, reasonably fast, and it must not concern only one concept at a time. However, the most urgent topic for research seems to be to develop systems that integrate different acquisition methods. 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, not discriminative, category representations. This demand disqualifies several popular learning methods such as TDIDT and the backpropagation algorithm. Furthermore, 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 on the perceptual level as well. In particular, agents that are alone rely heavily on such observations, whereas communicating agents also make observations on the linguistic level. As we have seen, however, there is a growing interest in developing vision systems able to learn category representations. Finally, we noted that concept learning has been limited to the learning of epistemological (and metaphysical) components, ignoring the inferential component.

## 9.2 Suggestions for Further Research

This thesis is certainly not going to be the last word in the research on concepts for autonomous agents. Some examples of interesting topics for further research are:

- Trying out the framework of composite category representation in realistic settings.
- Investigating whether it is necessary to expand the framework, which is now limited to the perception of and reasoning about objects, making it to cover actions related to the objects as well.
- Studying concepts representing non-object categories, such as event and situation categories.
- Examining the actual relevance of the metaphysical function and its relation to categorization by core (i.e., whether there are two fundamentally different ways of categorizing objects).

While these topics are closely related to my own research, there are several more general issues that need to be addressed in the future. For instance:

- The relationship between features, similarity, and categorical perception.
- The combination of exemplar-based and probabilistic representations.
- Top-down category formation.
- Integration of supervised and unsupervised learning.
- Learning inferential knowledge.

However, although it has begun to receive considerable attention, the most fundamental problem to be solved is how to integrate learning algorithms with perception systems successfully.



# Bibliography

- [1] *AAAI Fall Symposium Series, Instantiating Real-World Agents, (FS-93-03)*. AAAI Press, 1993.
- [2] *AAAI Fall Symposium Series, Machine Learning in Computer Vision: What, Why and How?, (FS-93-04)*. AAAI Press, 1993.
- [3] P.E. Agre and D. Chapman. Pengi: An implementation of a theory of activity. In *AAAI-87*, pages 268–272, 1987.
- [4] D.W. Aha, D. Kibler, and M.K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.
- [5] J.S. Albus. Outline for a theory of intelligence. *IEEE Trans. on Systems, Man, and Cybernetics*, 21(3):473–509, 1991.
- [6] Y. Aloimonos and A. Rosenfeld. Computer vision. *Science*, 253:1249–1254, 1991.
- [7] J. Amsterdam. Some philosophical problems with formal learning theory. In *AAAI-88*, pages 580–584, St. Paul, MN, 1988.
- [8] J.R. Anderson and M. Matessa. An incremental bayesian algorithm for categorization. In D.H. Fischer, M.J. Pazzani, and P. Langley, editors, *Concept Formation: Knowledge and Experience in Unsupervised Learning*, pages 45–70. Morgan Kaufmann, 1991.
- [9] D. Angluin and C.H. Smith. Inductive inference: Theory and methods. *Computing Surveys*, 15:237–268, 1983.
- [10] R.C. Arkin. Integrating behavioural, perceptual and world knowledge in reactive navigation. In P. Maes, editor, *Designing Autonomous Agents*, pages 105–122. MIT Press, 1990.
- [11] M. Asada. Map building for a mobile robot from sensory data. *IEEE Trans. on Systems, Man, and Cybernetics*, 20(6):1326–1336, 1990.
- [12] E. Astor, P. Davidsson, B. Ekdahl, and R. Gustavsson. Anticipatory planning. Technical Report LU-CS-TR: 90-69, Dept. of Computer Science, Lund University, Lund, Sweden, 1990. Also in Advance Proceedings of the European Workshop on Planning, 1991.
- [13] R.C. Atkinson and W.K. Estes. Stimulus sampling theory. In R.D. Luce, R.R. Bush, and E. Galanter, editors, *Handbook of Mathematical Psychology*. Wiley, 1963.
- [14] R.L. Atkinson, R.C. Atkinson, E.E. Smith, and E.R. Hilgard. *Introduction to Psychology, ninth edition*. Harcourt Brace Jovanovic Publishers, 1987.
- [15] C. Balkenius. Neural mechanisms for self-organization of emergent schemata, dynamical schema processing, and semantic constraint satisfaction. Lund University Cognitive Studies 14, ISSN 1101-8453, Lund University, Sweden, 1993.
- [16] C. Balkenius and P. Gärdenfors. Nonmonotonic inferences in neural networks. Lund University Cognitive Studies 3, ISSN 1101-8453, Lund University, Sweden, 1991.



- [17] D. H. Ballard. Animate vision. *Artificial Intelligence*, 48:57–86, 1991.
- [18] J. Bares, M. Hebert, T. Kanade, E. Krotkov, T. Mitchell, R. Simmons, and W. Whittaker. Ambler: An autonomous rover for planetary exploration. *IEEE Computer*, 22(6):18–26, 1989.
- [19] A.H. Barr. Superquadrics and angle-preserving transformations. *IEEE Comput. Graph. Appl.*, 1(1):11–23, 1981.
- [20] L.W. Barsalou. Are there static category representations in long-term memory? *Behavioral and Brain Sciences*, 9:651–652, 1986.
- [21] F. Bergadano, S. Matwin, R.S. Michalski, and J. Zhang. Learning two-tiered descriptions of flexible concepts, part I: Principles and methodology. Technical Report MLI 88–6 TR–14–88, Machine Learning & Inference Laboratory, Center for Artificial Intelligence, George Mason University, 1988.
- [22] F. Bergadano, S. Matwin, R.S. Michalski, and J. Zhang. Learning two-tiered descriptions of flexible concepts: The POSEIDON system. *Machine Learning*, 8(1):5–43, 1992.
- [23] M.K. Bhandaru, B.A. Draper, and V.R. Lesser. Learning image to symbol conversion. In *AAAI Fall Symposium on Machine Learning and Computer Vision: What, Why, and How?*, (FS-93-04). AAAI Press, 1993.
- [24] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987.
- [25] L. Birnbaum. Rigor mortis: A response to Nilsson’s “Logic and artificial intelligence”. *Artificial Intelligence*, 47(1):57–77, 1991.
- [26] M.J. Black et al. Action, representation, and purpose: Re-evaluating the foundations of computational vision. In *IJCAI-93*, pages 1661–1666, 1993.
- [27] D.R. Blidberg. Autonomous underwater vehicles: Current activities and research opportunities. In T. Kanade, F.C.A. Groen, and L.O. Hertzberger, editors, *Intelligent Autonomous Systems 2*, 1989.
- [28] L. Blum and M. Blum. Toward a mathematical theory of inductive inference. *Information and Control*, 28:125–155, 1975.
- [29] R.M. Bolle and B.C. Vemuri. On three-dimensional surface reconstruction methods. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(1):1–13, 1991.
- [30] M. Bowerman. The structure and origin of semantic categories in the language learning child. In M. Foster, editor, *Symbol as Sense: New Approaches to the Analysis of Meaning*, pages 277–299. Academic Press, 1980.
- [31] J. Bresina and M. Drummond. Integrating planning and reaction: A preliminary report. In J. Hendler, editor, *AAAI Spring Symposium on Planning in Uncertain, Unpredictable or Changing Environments*, 1990.
- [32] R.A. Brooks. Symbolic reasoning among 3-d models and 2-d images. *Artificial Intelligence*, 17:285–348, 1981.
- [33] R.A. Brooks. Elephants don’t play chess. In P. Maes, editor, *Designing Autonomous Agents*, pages 3–15. MIT Press, 1990.
- [34] R.A. Brooks. Intelligence without reason. In *IJCAI-91*, pages 569–595, Sidney, Australia, 1991.
- [35] R.A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47(1):139–159, 1991.
- [36] R.A. Brooks. New approaches to robotics. *Science*, 253:1227–1232, 1991.

- [37] R.A. Brooks. The role of learning in autonomous robots. In *Proceedings of the Fourth Annual Workshop on Computational Learning Theory*, pages 5–10. Morgan Kaufmann, 1991.
- [38] J.S. Bruner, J.J. Goodnow, and G.A. Austin. *A Study of Thinking*. Wiley, 1956.
- [39] W. Buntine. A critique of the Valiant model. In *IJCAI-89*, pages 837–842, 1989.
- [40] G.A. Carpenter and S. Grossberg. Neural dynamics of category learning and recognition: Attention, memory consolidation, and amnesia. In J. Davis, R. Newburgh, and E. Wegman, editors, *Brain Structure, Learning, and Memory, AAAS Symposium Series*, pages 239–286, 1986.
- [41] B. Chandrasekaran. Coming out from under a cloud: AAAI and IAAI '93. *IEEE Expert*, 8(5):78–81, 1993.
- [42] B. Chandrasekaran, N.H. Narayanan, and Y. Iwasaki. Reasoning with diagrammatic representation. *AI Magazine*, 14(2):49–56, 1993.
- [43] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman. AutoClass: A bayesian classification system. In *Fifth International Conference on Machine Learning*, pages 54–64, Ann Arbor, MI, 1988.
- [44] P. Cheeseman, M. Self, J. Kelly, W. Taylor, D. Freeman, and J. Stutz. Bayesian classification. In *AAAI-88*, pages 607–611, St. Paul, MN, 1988.
- [45] R.T. Chin and C.R. Dyer. Model-based recognition in robot vision. *ACM Computing Surveys*, 18(1):67–108, 1986.
- [46] J.H. Connell. *Minimalist Mobile Robots: A Colony-style Architecture for an Artificial Creature*. Academic Press, 1990.
- [47] J.H. Connell and M. Brady. Generating and generalizing models of visual objects. *Artificial Intelligence*, 31(2):159–183, 1987.
- [48] J.E. Corter. Relevant features and statistical models of generalization. *Behavioral and Brain Sciences*, 9:653–654, 1986.
- [49] P. Davidsson. On reactive planning. LU–CS–IR: 90–5, Dept. of Computer Science, Lund University, Lund, Sweden, 1990.
- [50] R. Davis, H. Shrobe, and P. Szolovits. What is a knowledge representation? *AI Magazine*, 14(1):17–33, 1993.
- [51] M. de la Maza. A prototype based symbolic concept learning system. In *Eighth International Workshop on Machine Learning*, pages 41–45, Evanston, IL, 1991.
- [52] G. DeJong and R. Mooney. Explanation-based learning: An alternative view. *Machine Learning*, 1(2):145–176, 1986.
- [53] L. Dey, P.P. Das, and S. Chaudhury. Recognition and learning of unknown objects in a hierarchical knowledge-base. In *AAAI Fall Symposium on Machine Learning and Computer Vision: What, Why, and How?, (FS-93-04)*. AAAI Press, 1993.
- [54] S.J. Dickinson et al. The use of geons for generic 3-d object recognition. In *IJCAI-93*, pages 1693–1699, 1993.
- [55] T.G. Dietterich and R.S. Michalski. Inductive learning of structural descriptions. *Artificial Intelligence*, 16(3):257–294, 1981.
- [56] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.
- [57] G.M. Edelman. *The Remembered Present: A Biological Theory of Consciousness*. Basic Books, 1989.

- [58] S.L. Epstein. The role of memory and concepts in learning. *Minds and Machines*, 2(3):239–265, 1992.
- [59] O. Etzioni. Intelligence without robots (a reply to Brooks). To appear in *AI Magazine*, 1993.
- [60] I.A. Ferguson. Touring machines: Autonomous agents with attitudes. *IEEE Computer*, 25(5):51–55, 1992.
- [61] M.A. Fischler and O. Firschein, editors. *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms*. Morgan Kaufmann Publishers, 1987.
- [62] D. Fisher and P. Langley. Approaches to conceptual clustering. In *IJCAI-85*, pages 691–697, Los Angeles, CA, 1985.
- [63] D.H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2:139–172, 1987.
- [64] D.H. Fisher. A computational account of basic level and typicality effects. In *AAAI-88*, pages 233–238, St. Paul, MN, 1988.
- [65] D.H. Fisher and K.B. McKusick. An empirical comparison of ID3 and back-propagation. In *IJCAI-89*, pages 788–793, 1989.
- [66] D.H. Fisher and M.J. Pazzani. Computational models of concept learning. In D.H. Fischer, M.J. Pazzani, and P. Langley, editors, *Concept Formation: Knowledge and Experience in Unsupervised Learning*, pages 3–43. Morgan Kaufmann, 1991.
- [67] N.S. Flann and T.G. Dietterich. Selecting appropriate representations for learning from examples. In *AAAI-86*, pages 460–466, 1986.
- [68] J.A. Fodor. *The Language of Thought*. Thomas Y. Crowell, 1975.
- [69] J.A. Fodor. *The Modularity of Mind*. Bradford Books, MIT Press, 1983.
- [70] P. Gärdenfors. *Knowledge in Flux*. MIT Press, 1988.
- [71] P. Gärdenfors. Frameworks for properties: Possible worlds vs. conceptual spaces. *Language, Knowledge, and Intentionality, Acta Philosophica Fennica*, 49:383–407, 1990.
- [72] P. Gärdenfors. Three levels of inductive inference. *Lund University Cognitive Studies* 9, ISSN 1101–8453, Lund University, Sweden, 1992.
- [73] E. Gat. On the role of stored internal state in the control of autonomous mobile robots. *AI Magazine*, 14(1):64–73, 1993.
- [74] M.S. Gazzaniga. Organization of the human brain. *Science*, 245:947–952, 1989.
- [75] M. Ginsberg. *Readings in Nonmonotonic Reasoning*. Morgan Kaufmann Publishers, 1987.
- [76] M. Ginsberg. Universal planning: An (almost) universally bad idea. *AI Magazine*, 10(4):40–44, 1989.
- [77] M.A. Gluck and J.E. Corter. Information, uncertainty, and the utility of categories. In *Seventh Annual Conference of the Cognitive Science Society*, pages 283–287. Lawrence Erlbaum Associates, 1985.
- [78] M. Gold. Language identification in the limit. *Information and Control*, 10(5):447–474, 1967.
- [79] K.P. Gross. Incremental multiple concept learning using experiments. In *Fifth International Conference on Machine Learning*, pages 65–72, Ann Arbor, MI, 1988.
- [80] S.J. Hanson and M. Bauer. Conceptual clustering, categorization, and polymorphy. *Machine Learning*, 3:343–372, 1989.

- [81] S. Harnad. Category induction and representation. In S. Harnad, editor, *Categorical Perception*, pages 535–565. Cambridge University Press, 1987.
- [82] S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.
- [83] B. Hayes-Roth. A blackboard architecture for control. *Artificial Intelligence*, 26(3):251–322, 1985.
- [84] B. Hayes-Roth. An integrated architecture for intelligent agents. *SIGART*, 2(4):79–81, 1991.
- [85] P.L. Heath. Concept. In P. Edwards, editor, *Encyclopedia of Philosophy*. Macmillan, 1967.
- [86] J. Hendler. Abstraction and reaction. In J. Hendler, editor, *AAAI Spring Workshop on Planning in Uncertain, Unpredictable or Changing Environments*, Stanford, 1990.
- [87] R.C. Holte, L.E. Acker, and B.W. Porter. Concept learning and the problem of small disjuncts. In *IJCAI-89*, pages 813–818, 1989.
- [88] K.J. Holyoak and R.E. Nisbett. Induction. In R.J. Sternberg and E.E. Smith, editors, *The Psychology of Human Thought*. Cambridge University Press, 1988.
- [89] V. Honavar. Inductive learning using generalized distance measures. In *SIPE Conference on Adaptive and Learning Systems*, Orlando, FL, 1992.
- [90] V. Honavar. Toward learning systems that integrate different strategies and representations. Technical Report TR: 93–22, Dept. of Computer Science, Iowa State University of Science and Technology, IA, 1993. A draft to be included in: *Symbol Processors and Connectionist Networks for Artificial Intelligence and Cognitive Modelling: Steps toward principled integration*, V. Honavar and L. Uhr (ed), Academic Press, 1994.
- [91] E.B. Hunt, J. Marin, and P.J. Stone. *Experiments in Induction*. Academic Press, New York, 1966.
- [92] F.F. Ingrand, M.P. Georgeff, and A.S. Rao. An architecture for real-time reasoning and system control. *IEEE Expert*, 7(6):34–44, 1992.
- [93] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.
- [94] C.G. Jansson. *Taxonomic Representation*. PhD thesis, The Royal Institute of Technology, Stockholm, Sweden, 1987.
- [95] L. Kanal and T. Tsao. Artificial intelligence and natural perception. In L.O. Hertzberger and F.C.A. Groen, editors, *Intelligent Autonomous Systems*, pages 60–70. North-Holland, 1987.
- [96] M. Kearns, M. Li, L. Pitt, and L.G. Valiant. Recent results on boolean concept learning. In *Fourth International Workshop on Machine Learning*, pages 337–352. Morgan Kaufmann, 1987.
- [97] D. Kibler and D. Aha. Learning representative exemplars of concepts. In *Fourth International Workshop on Machine Learning*, pages 24–30, Irvine, CA, 1987.
- [98] D. Kirsh. Second-generation AI theories of learning. *Behavioral and Brain Sciences*, 9:658–659, 1986.
- [99] D. Kirsh. Foundations of AI: The big issues. *Artificial Intelligence*, 47(1):3–30, 1991.
- [100] D. Kirsh. Today the earwig, tomorrow man? *Artificial Intelligence*, 47(1):161–184, 1991.
- [101] T. Kohonen. The “neural” phonetic typewriter. *IEEE Computer*, 27(3):11–22, 1988.
- [102] J.B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage, 1978.

- [103] J.E. Laird, A. Newell, and P. Rosenbloom. SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33(1):1–64, 1987.
- [104] G. Lakoff. *Women, Fire, and Dangerous Things: What categories reveal about the mind*. The University of Chicago Press, 1987.
- [105] J.M. Lammens, H.H. Hexmoor, and S.C. Shapiro. Of elephants and men. In *NATO-ASI on the Biology and Technology of Intelligent Autonomous Agents*, Trento, Italy, 1993.
- [106] P. Langley. Machine learning and grammar induction. *Machine Learning*, 2(1):5–8, 1987.
- [107] P. Langley, J.M. Zytkow, G.L. Bradshaw, and H.A. Simon. Three facets of scientific discovery. In *IJCAI-83*, pages 465–468, Karlsruhe, 1983.
- [108] M. Lebowitz. Concept learning in a rich input domain: Generalization-based memory. In R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, editors, *Machine Learning: An AI Approach, Volume II*, pages 193–214. Morgan Kaufmann, 1986.
- [109] M. Lebowitz. The utility of similarity-based learning in a world needing explanation. In R.S. Michalski and Y. Kodratoff, editors, *Machine Learning: An AI Approach, Volume III*, pages 399–422. Morgan Kaufmann, 1990.
- [110] D.B. Lenat. *AM: An Artificial Intelligence Approach to Discovery in Mathematics as Heuristic Search*. PhD thesis, Stanford University, 1976.
- [111] D.B. Lenat. Automated theory formation in mathematics. In *IJCAI-77*, pages 833–842, Cambridge, MA, 1977.
- [112] D.B. Lenat and J.S. Brown. Why AM and Eurisko appear to work. *Artificial Intelligence*, 23(3):269–294, 1984.
- [113] L.M. Lewis. A time-ordered architecture for integrating reflexive and deliberative behaviour. *SIGART*, 2(4):110–114, 1991.
- [114] R.C. Lou and M.G. Kay. Multisensor integration and fusion in intelligent systems. *IEEE Trans. on Systems, Man, and Cybernetics*, 19(5):901–931, 1989.
- [115] D.M. Lyons and A.J. Hendriks. Reactive planning. In S.C. Shapiro, editor, *Encyclopedia of Artificial Intelligence, 2nd ed.*, pages 1171–1181. John Wiley and Sons, 1992.
- [116] D.M. Lyons, A.J. Hendriks, and S. Mehta. Achieving robustness by casting planning as adaption of a reactive system. In *IEEE Int. Conf. on Robotics and Automation*, 1991.
- [117] R.L. Madarasz, L.C. Heiny, R.F. Crompt, and N.M. Mazur. The design of an autonomous vehicle for the disabled. *IEEE Journal of Robotics and Automation*, 2(3):117–126, 1986.
- [118] D. Marr. *Vision*. Freeman, 1982.
- [119] C. J. Matheus, L. R. Rendell, D. L. Medin, and R. L. Goldstone. Purpose and conceptual functions: A framework for concept representation and learning in humans and machines. In A. G. Cohn, editor, *Proc. of the Seventh Conference of the Society for the Study of Artificial Intelligence and Simulation of Behaviour*, pages 13–20. Pitman and Kaufmann, 1989.
- [120] C.J. Matheus. Conceptual purpose: Implications for representation and learning in machines and humans. Technical Report UIUCDCS-R-87-1370, Department of Computer Science, University of Illinois at Urbana-Champaign, 1987.
- [121] M.W. Matlin. Concept learning. In R.J. Corsini, editor, *Encyclopedia of Psychology*. John Wiley & Sons, 1984.
- [122] G. Matthews and J. Hearne. Clustering without a metric. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(2):175–184, 1991.

- [123] M. McCloskey and S. Glucksberg. Natural categories: Well defined or fuzzy-sets? *Memory and Cognition*, 6:462–472, 1978.
- [124] D.L. Medin and M.M. Schaffer. Context theory of classification learning. *Psychological Review*, 85:207–238, 1978.
- [125] D.L. Medin and E.E. Smith. Concepts and concept formation. *Annual Review of Psychology*, 35:113–138, 1984.
- [126] R.S. Michalski. Pattern recognition as rule-guided inductive inference. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2(4):349–361, 1980.
- [127] R.S. Michalski. Concept learning. In S.C. Shapiro, editor, *Encyclopedia of Artificial Intelligence*, pages 185–194. John Wiley and Sons, 1987.
- [128] R.S. Michalski. 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, 1987.
- [129] R.S. Michalski and R.L. Chilausky. 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), 1980.
- [130] R.S. Michalski and Y. Kodratoff. Research in machine learning: Recent progress, classification of methods, and future directions. In Y. Kodratoff and R.S. Michalski, editors, *Machine Learning: An AI Approach, Volume III*, pages 3–30. Morgan Kaufmann, 1990.
- [131] R.S. Michalski and J.B. Larson. 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, 1978.
- [132] R.S. Michalski and R.E. Stepp. Learning from observation: Conceptual clustering. In R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, editors, *Machine Learning: An AI Approach*, pages 331–363. Springer-Verlag, 1983.
- [133] L. Miclet. *Structural Methods in Pattern Recognition*. North Oxford Academic, 1986.
- [134] M. Minsky. A framework for representing knowledge. In P.H. Winston, editor, *The Psychology of Computer Vision*, pages 211–277. McGraw-Hill, 1975.
- [135] M. Minsky and S. Papert. *Perceptrons*. MIT Press, 1969.
- [136] S. Minton. On modularity in integrated architectures. *SIGART*, 2(4):134–135, 1991.
- [137] T.M. Mitchell. Version spaces: A candidate elimination approach to rule learning. In *IJCAI-77*, pages 305–310, Cambridge, MA, 1977.
- [138] T.M. Mitchell. Becoming increasingly reactive. In *AAAI-90*, pages 1051–1058, 1990.
- [139] T.M. Mitchell, R.M. Keller, and S.T. Kedar-Cabelli. Explanation-based generalization: A unifying view. *Machine Learning*, 1(1):47–80, 1986.
- [140] R. Mooney, J. Shavlik, G. Towell, and A. Gove. An experimental comparison of symbolic and connectionist learning algorithms. In *IJCAI-89*, pages 775–780, 1989.
- [141] S. Moscatelli and Y. Kodratoff. Advanced machine learning techniques for computer vision. In *Advanced Topics in Artificial Intelligence, Lecture Notes in Artificial Intelligence 617*, pages 161–197. Springer Verlag, 1992.
- [142] H. Murase and S.K. Nayar. Learning object models from appearance. In *AAAI-93*, pages 836–843, 1993.

- [143] G.L. Murphy and D.L. Medin. The role of theories in conceptual coherence. *Psychological Review*, 92:289–316, 1985.
- [144] N. Muscettola, S.F. Smith, A. Cesta, and D. D'Aloisi. Coordinating space telescope operations in an integrated planning and scheduling architecture. *IEEE Control Systems Magazine*, 12(1):28–37, 1992.
- [145] P.B. Musgrove and R.I. Phelps. An automatic system for acquisition of natural concepts. In *ECAI-90*, pages 455–460, Stockholm, Sweden, 1990.
- [146] D.J. Nagel. Learning concepts with a prototype-based model for concept representation. In T.M. Mitchell, J.G. Carbonell, and R.S. Michalski, editors, *Machine Learning: A Guide to Current Research*, pages 233–236. Kluwer Academic Press, 1986.
- [147] D.J. Nagel. *Learning Concepts with a Prototype-based Model for Concept Representation*. PhD thesis, Department of Computer Science, Rutgers, The State University of New Jersey, 1987.
- [148] U. Neisser, editor. *Concepts and Conceptual Development: Ecological and intellectual factors in categorization*. Cambridge University Press, 1987.
- [149] K. Nelson. Concept, word, and sentence: Interrelations in acquisition and development. *Psychological Review*, 81(4):267–285, 1974.
- [150] A. Newell. *Unified Theories of Cognition*. Harvard University Press, 1990.
- [151] N.J. Nilsson. A mobile automaton: An application of artificial intelligence techniques. In *IJCAI-69*, pages 509–520, 1969.
- [152] D. Osherson, M. Stob, and S. Weinstein. *Systems that Learn*. A Bradford Book, MIT Press, 1986.
- [153] P.W. Pachowicz. Integration of machine learning and vision into an active agent paradigm. In *AAAI Fall Symposium on Machine Learning and Computer Vision: What, Why, and How?*, (FS-93-04). AAAI Press, 1993.
- [154] K. Pahlavan. *Active Robot Vision and Primary Ocular Processes*. PhD thesis, Royal Institute of Technology, Stockholm, Sweden, 1993.
- [155] L. Pitt. Introduction: Special issue on computational learning theory. *Machine Learning*, 5(2):117–120, 1990.
- [156] L. Pitt and R.E. Reinke. Criteria for polynomial-time (conceptual) clustering. *Machine Learning*, 2(4):371–396, 1988.
- [157] T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266, 1990.
- [158] A.R. Pope and D.G. Lowe. Learning 3-d object recognition models from 2-d images. In *AAAI Fall Symposium on Machine Learning and Computer Vision: What, Why, and How?*, (FS-93-04). AAAI Press, 1993.
- [159] Z.W. Pylyshyn. *Computation and Cognition: Toward a Foundation for Cognitive Science*. MIT Press, 1984.
- [160] M.R. Quillian. Semantic memory. In M. Minsky, editor, *Semantic Information Processing*, pages 227–270. MIT Press, 1968.
- [161] W.V.O. Quine. *Ontological Relativity and other Essays*. Columbia University Press, 1969.
- [162] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
- [163] L. Rendell. A new basis for state-space learning systems and a successful implementation. *Artificial Intelligence*, 20(4):369–392, 1983.

- [164] L. Rendell. A general framework for induction and a study of selective induction. *Machine Learning*, 1(2):177–226, 1986.
- [165] L. Rendell. Learning hard concepts. In *Third European Working Session on Learning*, pages 177–200, 1988.
- [166] L. Rendell. Comparing systems and analyzing functions to improve constructive induction. In *Sixth International Workshop on Machine Learning*, pages 461–464, 1989.
- [167] G. Rey. Concepts and stereotypes. *Cognition*, 15:237–262, 1983.
- [168] G. Rey. Concepts and conceptions: A reply to Smith, Medin and Rips. *Cognition*, 19:297–303, 1985.
- [169] E. Rosch. Principles of categorization. In E. Rosch and B.B. Lloyd, editors, *Cognition and Categorization*, pages 28–49. Erlbaum, 1978.
- [170] E. Rosch and C.B. Mervis. Family resemblances. studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605, 1975.
- [171] E. Rosch, C.B. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.
- [172] R. Rosen. *Anticipatory Systems – Philosophical, Mathematical and Methodological Foundations*. Pergamon Press, 1985.
- [173] S. Rosenschein and L. Kaelbling. The synthesis of digital machines with provable epistemic properties. In J.F. Halpern, editor, *The 1986 Conference on Theoretical Aspects of Reasoning and Knowledge*, 1987.
- [174] Y. Roth-Tabak and R. Jain. Building an environment model using depth information. *IEEE Computer*, 22(6):85–90, 1989.
- [175] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning internal representations by error propagation. In D.E. Rumelhart and J.L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol.1: Foundations*. MIT Press, 1986.
- [176] B. Russell. *The Problems of Philosophy*. Oxford University Press, 1912.
- [177] R.C. Schank and R.P. Abelson. *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum, 1977.
- [178] R.C. Schank, G.C. Collins, and L.E. Hunter. Transcending inductive category formation in learning. *Behavioral and Brain Sciences*, 9:639–686, 1986.
- [179] J.C. Schlimmer and P. Langley. Machine learning. In S.C. Shapiro, editor, *Encyclopedia of Artificial Intelligence, 2nd ed.*, pages 785–805. John Wiley and Sons, 1992.
- [180] P.G. Schyns. A modular neural network model of concept acquisition. *Cognitive Science*, 15:461–508, 1991.
- [181] K. Sengupta and K.L. Boyer. Information theoretic clustering of large structural model-bases. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 174–179, 1993.
- [182] R.N. Shepard. Representation of structure in similarity data: Problems and prospects. *Psychometrika*, 39:373–421, 1974.
- [183] M. Shneier. A compact relational structure representation. In *IJCAI-79*, pages 818–826, 1979.
- [184] S. Sjölander. Some cognitive break-throughs in the evolution of cognition and consciousness, and their impact on the biology of language. *Evolution and Cognition*, 1994. In print.



- [185] A. Sloman. Why we need many knowledge representation formalisms. In M. Bramer, editor, *Research and Development in Expert Systems: Proc. of the BCS Expert Systems Conf. 1984*. Cambridge University Press, 1985.
- [186] E.E. Smith. Category differences/automaticity. *Behavioral and Brain Sciences*, 9:667, 1986.
- [187] E.E. Smith. Concepts and thought. In R.J. Sternberg and E.E. Smith, editors, *The Psychology of Human Thought*. Cambridge University Press, 1988.
- [188] E.E. Smith. Concepts and induction. In M.L. Posner, editor, *Foundations of Cognitive Science*, pages 501–526. MIT Press, 1989.
- [189] E.E. Smith and D.L. Medin. *Categories and Concepts*. Harvard University Press, 1981.
- [190] E.E. Smith, D.L. Medin, and L.J. Rips. A psychological approach to concepts: Comments on Rey's "Concepts and stereotypes". *Cognition*, 17:265–274, 1984.
- [191] P. Smolensky. On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11:1–74, 1988.
- [192] P. Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1-2):159–216, 1990.
- [193] P. Smyth and J. Mellstrom. Detecting novel classes with applications to fault diagnosis. In *Ninth International Workshop on Machine Learning*, pages 416–425, Aberdeen, Scotland, 1992.
- [194] R.J. Solomonoff. A formal theory of inductive inference. *Information and Control*, 7(1,2):1–22, 224–254, 1964. Parts I and II.
- [195] R.L. Solso. *Cognitive Psychology, third edition*. Allyn and Bacon, 1991.
- [196] R.E. Stepp. Concepts in conceptual clustering. In *IJCAI-87*, pages 211–213, Milan, Italy, 1987.
- [197] R.E. Stepp and R.S. Michalski. Conceptual clustering: Inventing goal-oriented classifications of structured objects. In R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, editors, *Machine Learning: An AI Approach, Volume II*, pages 471–498. Morgan Kaufmann, 1986.
- [198] R.E. Stepp and R.S. Michalski. Conceptual clustering of structured objects: A goal-oriented approach. *Artificial Intelligence*, 28(1):43–69, 1986.
- [199] D. Subramanian. Representational issues in machine learning. In *Sixth International Workshop on Machine Learning*, pages 426–429, 1989.
- [200] S. Tachi and K. Komoriya. Guide dog robot. In S.S. Iyengar and A. Elfes, editors, *Autonomous Mobile Robots: Control, Planning and Architecture*, pages 360–367. IEEE Computer Society Press, 1991.
- [201] K. Thompson and P. Langley. Incremental concept formation with composite objects. In *Sixth International Workshop on Machine Learning*, pages 371–374, 1989.
- [202] K. Thompson and P. Langley. Concept formation in structured domains. In D.H. Fischer, M.J. Pazzani, and P. Langley, editors, *Concept Formation: Knowledge and Experience in Unsupervised Learning*, pages 127–161. Morgan Kaufmann, 1991.
- [203] G.G. Towell, J.W. Shavlik, and M.O. Noordewier. Refinement of approximate domain theories by knowledge-based neural networks. In *AAAI-90*, pages 861–866, 1990.
- [204] M.M. Trivedi, C. Chen, and S.B. Marapane. A vision system for robotic inspection and manipulation. *IEEE Computer*, 22(6):91–97, 1989.
- [205] A. Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977.

- [206] S. Ullman. Visual routines. *Cognition*, 18:97–106, 1984.
- [207] P.E. Utgoff. Shift of bias for inductive concept learning. In R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, editors, *Machine Learning: An AI Approach, Volume II*. Morgan Kaufmann, 1986.
- [208] L.M. Vaina and M-C. Jaulent. Object structure and action requirements: A compatibility model for functional recognition. *International Journal of Intelligent Systems*, 6:313–336, 1991.
- [209] L.G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.
- [210] S. Watanabe. *Knowing and Guessing: A Quantitative Study of Inference and Information*. John Wiley and Sons, 1969.
- [211] S.M. Weiss and I. Kapouleas. An empirical comparison of pattern recognition, neural nets and machine learning classification methods. In *IJCAI-89*, pages 781–787, 1989.
- [212] S.M. Weiss and C.A. Kulikowski. *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann, 1991.
- [213] P.H. Winston. Learning structural descriptions from examples. In P.H. Winston, editor, *The Psychology of Computer Vision*, pages 157–209. McGraw-Hill, 1975.
- [214] P.H. Winston. Learning new principles from precedents and exercises. *Artificial Intelligence*, 19(3):321–350, 1982.
- [215] L. Wittgenstein. *Philosophical Investigations*. Basil Blackwell, 1953.
- [216] A.D. Woozley. Universals. In P. Edwards, editor, *Encyclopedia of Philosophy*. Macmillan, 1967.
- [217] S. Wrobel. Towards a model of grounded concept formation. In *IJCAI-91*, pages 712–717, Sidney, Australia, 1991.



## **Part II**

# **A Framework for Autonomous Agents Based on the Concept of Anticipatory Systems**





















## **Part III**

# **Toward a General Solution to the Symbol Grounding Problem: Combining Machine Learning and Computer Vision**

















## **Part IV**

# **A Framework for Organization and Representation of Concept Knowledge in Autonomous Agents**

























