

PREDICTION AS THE BASIS OF LOW LEVEL
COGNITIVE ORGANIZATION

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER
ENGINEERING AND INFORMATION SCIENCE
AND THE INSTITUTE OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Armağan Yavuz
January, 1998

Q
335
.Y28
1998

PREDICTION AS THE BASIS OF LOW LEVEL COGNITIVE ORGANIZATION

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER
ENGINEERING AND INFORMATION SCIENCE
AND THE INSTITUTE OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Armağan Yavuz
January, 1998

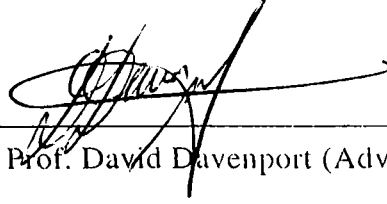
Armağan Yavuz

Armağan Yavuz

Q
336
-Y38
1998

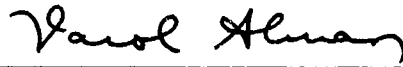
B 040200

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



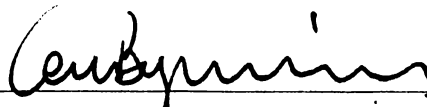
Asst. Prof. David Davenport (Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



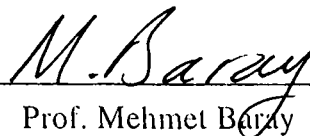
Prof. Varol Akman

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



Asst. Prof. Cem Bozşahin

Approved for the Institute of Engineering and Science:



Prof. Mehmet Baray
Director of Institute of Engineering and Science

ABSTRACT

PREDICTION AS THE BASIS OF LOW LEVEL COGNITIVE ORGANIZATION

Armağan Yavuz

M.S. in Computer Engineering and Information Science

Advisor: Asst. Prof. David Davenport

January, 1998

I suggest that the brain's low-level sensory-motor systems are organized on the basis of prediction. This suggestion differs radically from existing theories of sensory-motor systems, and can be summarized as follows. Certain simple mechanisms in the brain predict the current or future states of other brain mechanisms. These mechanisms can be established and disposed dynamically. Successful prediction acts as a kind of selection criteria and new structures are formed and others are disposed according to their predictive powers. Simple mechanisms become connected to each other on the basis of their predictive power, possibly establishing hierarchical structures, and forming large complexes. The complexes so formed, can implement a number of functionalities including detecting interesting events, creating high-level representations, and helping with goal-directed activity. Faculties such as attention and memory contribute to such processes of internal predictions and they can be studied and understood within this setting. All of this does not rule out the existence of other mechanisms, but an organization driven by prediction serves as the backbone of low-level cognitive activity.

I develop a computational model of a sensory-motor system that works on this basis. I also show how this model explains certain interesting aspects of human perception and how it can be related to general cognitive capabilities.

Keywords: prediction mechanism, emergent representations, constructivism, perception, cognition, cognitive science.

ÖZET

BİLİŞSEL DİZGENİN ALT DÜZEY ÖĞELERİNİN TAHMİN ETME TEMELİNDE ORGANİZASYONU

Armağan Yavuz

Bilgisayar ve Enformatik Mühendisliği, Yüksek Lisans

Danışman: Yrd. Doç. Dr. David Davenport

Ocak 1998

Bu tezde beyinin alt düzey duyusal-motor işlevlerinin tahmin etme temelinde organize olduğu öne sürülmektedir. Bu öneri duyusal-motor sistemler hakkındaki varolan kuramlardan oldukça farklıdır ve şu şekilde özetlenebilir: Beyindeki birtakım basit mekanizmalar diğerlerinin o anki ya da gelecekteki durumlarını tahmin ederler. Bu mekanizmalar dinamik olarak oluşabilir ya da yok olabilirler. Tahminlerdeki doğruluk derecesi bu süreçte seçim kriteri olarak rol oynar. Basit mekanizmalar bu şekilde birbirlerine bağlanır ve hiyerarşik kompleksler oluştururlar. Bu kompleksler, ilginç olayları tanıma, yüksek düzey gösterimler oluşturma, ve bir hedefe yönelik etkinliklere yardımcı olma gibi bir dizi işlevi yerine getirirler. Dikkat ve bellek gibi diğer dizgeler bu işlemlere yardımcı olur ve bu sistemden yararlanırlar. Tahmin temelinde gerçekleşen böyle bir organizasyon alt-düzeysel bilişsel etkinliklerin temelini oluşturur.

Bu tezde, tahmin temelinde çalışan bir duyusal-motor sistem modeli verilmekte ve bu modelin algıya ilişkin bazı ilginç problemleri nasıl çözümlendiği ve diğer bilişsel etkinliklerle nasıl ilişkili olabileceği tartışılmaktadır.

Anahtar Sözcükler: tahmin mekanizması, gösterimlerin oluşumları, konstruktivizm, algı, biliş, bilişsel bilim.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to Dr. David Davenport, who grew in me the interest towards cognitive science, gave me the original insight and ideas of this thesis, and supervised and encouraged me through all the stages of my study with wholehearted kindness and sincerity. Without his guidance, this thesis would not be possible.

I would like to thank my dear family and my friends for their assistance and continuous support. I would especially like to thank Azer Keskin who has always been a great help, Burak Acar and Toygar Birinci, who have contributed to this thesis with excellent insights, comments, and ideas, and my office mates Huseyin Kutluca, Murat Temizsoy, and Bora Uçar.

Contents

1 Introduction	1
1.1 The Motivation	1
1.2 The Principle Statement.....	2
1.3 Why Prediction?	3
1.4 Prediction in the Small.....	5
1.5 Organization of the Thesis	6
2 On Modeling Cognition	7
2.1 Problems of Modeling Cognition	7
2.1.1 Epistemic Access	8
2.1.2 The Frame Problem	9
2.1.3 The Whole and Its Parts.....	10
2.1.4 Interaction Among Cognitive Mechanisms	11
2.1.5 Learning	11
2.1.6 Adaptation.....	12
2.2 Models of Cognitive Activity	12
2.2.1 Connectionist Models	13
2.2.2 Symbolic Models	16
2.2.3 Some Other Directions in Cognitive Science	17
2.2.4 Constructivism: The Search for “Order from Noise”	19
3 A Prediction Mechanism	22
3.1 An Intuitive Opening	22
3.2 A Hierarchy of Predictions	23
3.3 Functions of the Prediction Mechanism	24
3.3.1 Carrying Out Goal-Directed Activity.....	25

3.3.2	Generating High Level Percepts	25
3.3.3	Revealing “Interesting” Information.....	26
3.3.4	Remarks	27
3.4	Some Guiding Principles	28
3.5	Setting up the system	29
3.5.1	The Composition of the Cognitive System.....	29
3.5.2	Time Scales.....	29
3.5.3	The Interaction Between the Agent and the Environment	30
3.5.4	Events	31
3.6	The Basic Predictor.....	32
3.6.1	Establishing Basic Predictors.....	32
3.6.2	Goal-Directed Activity with Basic Predictors	34
3.7	Meta-Predictors.....	36
3.7.1	Establishing Meta-Predictors	38
3.7.2	Prediction with Meta-Predictors	40
3.7.3	Goal-Directed Behavior with Meta-Predictors	41
3.7.4	Timing Issues.....	46
3.7.5	Examples with Meta-Predictors.....	46
3.8	Generalizers	48
3.8.1	Emergence of a Representation of Obstacles.....	50
3.8.2	Low-level Organization of a Visual System	51
3.9	State-Holders	54
3.9.1	Establishing State-holders.....	56
3.9.2	Sequences of Actions.....	57
3.9.3	Re-trying an Interaction	57
3.10	Remarks	58
4	Discussion	59
4.1	Questions Regarding the Prediction Mechanism.....	59
4.2	Functions of the Prediction Mechanism	62
4.2.1	Supporting Goal-Directed Activity.....	62

4.2.2 Finding Interesting Events	62
4.2.3 Generation of High Level Notions.....	63
4.3 Capabilities of the Prediction mechanism	64
4.3.1 Pattern Recognition.....	64
4.3.2 Composite Actions.....	64
4.4 Towards High Level Cognitive Skills.....	66
4.4.1 An Attention System.....	66
4.4.2 An Episodic Memory System	67
5 Conclusions	68

List of Figures

2.1. Components of an artificial neural network.....	13
2.2. Traditional vs. behavior-based decomposition of a mobile robot control system.	19
3.1. Graphical representations of sensors.....	30
3.2. Graphical representation of actuators.....	31
3.3. An example predictor.....	33
3.4. Rules for basic predictors to propagate goalness and avoidedness.....	35
3.5. A robot hand connected to sensors and actuators	36
3.6. Predictor p is established between <i>touch</i> and <i>close-hand</i>	38
3.7. Rules for meta-predictors predicting the success or the failure of their targets	41
3.8. An example of goal-propagation with meta-predictors.....	41
3.9. Inhibition of goal-propagation with negative meta-predictors.....	42
3.10. Rule 3.	42
3.11. Rules 4.a and 4.b.....	43
3.12. Rules 5.a and 6.a.	44
3.13. Rule 7.	44
3.14. Rules 8.a and 9.a	45
3.15. Predictors for recognizing synchronous activation.	47
3.16. An example for the fine control of an action	48
3.17. Establishment of a generalizer.	49
3.18. A pattern of activation on the grid of sensors	51
3.19. Predictors established between proximate sensors.	52

3.20. Predictors collected through a generalizer.....	52
3.21. The original activation pattern and the primal image constructed by generalizers.....	53
3.22. A state holder.....	55
3.23. The usage of a state-holder together with a generalizer.....	56
3.24. Predictors that recognize a temporal sequence.....	57
3.25. A scheme for the robot hand to remember its state.....	58
4.1. Sensory motor control of the robot hand.....	65

Chapter 1

Introduction

In a seminar, my thesis supervisor Dr. Davenport had claimed that there are three types of fundamental questions: questions regarding the origin and the nature of the universe; questions regarding the origin and the nature of life; and questions regarding the origin and the nature of the human mind. Like him, I am deeply interested in the last one of these. In this thesis, I try to explore a new perspective for investigating the workings of the mind. Briefly, I claim that the brain is a system that (among doing other things) continually predicts its future states and re-adjusts itself for improving its predictions. This basic idea can be exploited to study how different levels of the cognitive mechanism can be organized. To begin with, I apply it to the low-level perception system and develop a computational model. I also discuss how this idea can help in modeling the workings of the mind in a broader setting.

1.1 The Motivation

One of the primary theories that claim to account for human intelligence is constructivism, the idea that concepts, categories, skills, and the like are not in-born, but are actively “constructed” by the agent in an attempt to put an order to the seemingly chaotic nature of its interaction with the environment. As yet, there is no viable computational model that accounts for a constructivist development. As a matter of fact, what such a model would look like is generally unclear. However, it seems that the requirements constructivism places on a cognitive model are so harsh that a constructivist model must necessarily exploit certain fundamental principles and methods, like those typically employed in scientific

research. Of course, what exactly these principles and methods are and how they could be applied to cognitive functioning is not evident.

Carrying out predictions seems to be a good candidate to serve as an underlying principle of a construction process. So, I try to explore this possibility and see how prediction can contribute to a constructivist development process. However, the framework I present in this thesis should not be regarded as a general development theory, but as a precursor to such a theory. The methodology I adopt in this study is perhaps more relevant to constructivism than the framework itself.

1.2 The Principal Statement

The question “How does the mind work?” has attracted the interest of philosophers and scientists for many centuries. Modern cognitive science tries to answer this basic question, but it departs from past endeavors on two important points: Firstly, it stresses the importance of interdisciplinary collaboration. Secondly, it uses the computer metaphor for analyzing, understanding and modeling the mind. However, apart from these two points, there seems to be no single idea within cognitive science that is the subject of a general consensus. Researchers and philosophers disagree with each other on every possible aspect of studying cognition, on matters such as the appropriate level of analysis, the relevance of learning to intelligence, the nature of representations, and so on. However, all these disagreements are, really, natural if we consider the difficulty of the problem at hand; and in fact, one sometimes feels relieved that the cognitive science community is not betting all its money on a single horse.

What I want to do in this thesis is to present a new perspective to studying cognition, introduce a new horse to the game, if you like. The basic idea that I base my study on, that prediction is somehow relevant to intelligence, is really an old one, but I try to polish it up and re-introduce it by suggesting a new way for making use of it. So, here is, what I can call, the principal statement of my thesis:

Certain simple mechanisms in the brain predict the current or future states of other brain mechanisms. These mechanisms can be established and disposed dynamically. Successful prediction acts as a kind of selection criteria and new structures are formed and others are disposed according to their predictive powers. Simple mechanisms become connected to each other on the basis of their predictive power, possibly establishing hierarchical structures, and forming large complexes. The complexes so formed, can implement a number of functionalities

including detecting interesting events, creating high-level representations, and helping with goal-directed activity. Faculties such as attention and memory contribute to such processes of internal predictions and they can be studied and understood within this setting. All of this does not rule out the existence of other mechanisms, but an organization driven by prediction serves as the backbone of low-level cognitive activity.

What I want to establish in this thesis is that this statement is plausible, and a solution to some of the age old problems of cognition may be found along this way.

1.3 Why Prediction?

Prediction is, to put it simply, making guesses about the future. An individual can be said to be predicting, if he is not just making up the guesses out of the blue, but his guesses are based on some ground. It seems quite intuitive that the power to predict has something to do with intelligence. However, this possibility has not been fully explored so far within cognitive science research. The reason, I believe, can be found in the following excerpt by McCarthy and Hayes [16:40] that shows the typical criticisms towards the idea:

A number of investigators ... have taken the view that intelligence may be regarded as the ability to predict the future of a sequence from observation of its past. Presumably, the idea is that the experience of a person can be regarded as a sequence of discrete events and that intelligent people can predict the future. Artificial intelligence is then studied by writing programs to predict sequences formed according to some simple class of laws (sometimes probabilistic laws). ... [However,] what we know about the world is divided into knowledge about many aspects of it, taken separately and with rather weak interaction. A machine that worked with the undifferentiated encoding of experience into a sequence would first have to solve the encoding, a task more difficult than any sequence extrapolators are prepared to undertake. Moreover, our knowledge is not usable to predict exact sequences of experience. Imagine a person who is correctly predicting the course of a football game he is watching; he is not predicting each visual sensation (the play of light and shadow, the exact movements of the players and the crowd). Instead his prediction is on the level of: team A is getting tired; they should start to fumble or have their passes intercepted.

This criticism is fair enough, and if we take the idea of prediction in this sense, it is of little interest to the study of cognition. However, there exists another sense of prediction, one I can call prediction-in-the-small, that may be helpful in ex-

plaining intelligence. Before going on with that though, I would like to look at two “merits” of prediction that make it worthwhile of study.

Self-contained Error Criterion

A major difficulty in dealing with knowledge is identifying what is non-knowledge. The usual practice in the field of machine learning is using separate training and test sets, the first set, to teach the machine something, and the second to test if it has learned it. Artificial Neural Networks using back-propagation learning rule require error signals for their functioning. Similarly, negative examples are used in many other learning paradigms.

The good thing about predictions is that they contain their own error signals. The predicted event either occurs, in which case there is no error, or does not occur, in which case there is an error. There is no need for external error signals, negative examples, or other kinds of error criteria. The act of prediction is unique in its simplicity of detecting errors.

Prediction and Knowledge

Knowledge, at least in its simpler sense, can be reduced to the power to predict. Assume I am dealing with some sophisticated system X . If I can correctly predict how X is going to behave in all possible circumstances (note that I am not specifying how I come to be able to predict those. For my present purposes, the way I do this may be as sophisticated as ever), then this fact may stand in for the fact that I know the nature of X . These two things are in fact distinct, because X may have many inner details that are not relevant to the behavior I observe. However, if I am a practical kind of person, I can just forget about the distinction. After all, if I can predict X 's behavior correctly, I can use it for achieving my goals, and manipulate it as I see fit. As long as I am interested in the practical aspects of X , those aspects that are relevant to my everyday pursuits, the difference does not matter.

What about real knowledge of X ? I suppose, in order to have a real understanding of X , I have to, somehow, look into X , and this time observe and be able to predict the behavior of its components together with its external behavior. If I can predict the transitions between X 's internal states, I believe, this fact shows that I have some real understanding of X . Obviously, I can repeat the same exercise for the components of X , components of components of X , and so on, recursively, until I hit the quantum level. And even the quantum level is defined in terms of prediction, by the impossibility of making accurate predictions about particles.

1.4 Prediction in the Small

In everyday language, we usually use the word prediction to refer to some conscious, deliberate act of guessing the future by using our past memories, experiences, or pieces of knowledge we have obtained in other ways. We predict that it's going to rain when we see black clouds in the sky. We predict that the boss will be frustrated when he hears about the delay or that the next world cup will go to a south American team. Such predictions are no doubt interesting acts, but there seems to be no reason to attribute them any special importance in explaining the mechanisms of cognition. However, just as there are things in the world that can add two numbers without attending primary school, there may be things that can predict something without having conscious experience or without having access to human kind's accumulated wisdom. It seems that the only constraints on calling a guess a prediction are that: (1) the guess is based on a somewhat sound basis (possibly statistical data, but of course, not limited to that) and (2) the outcome of the prediction (the success or the failure) is regarded as what it is: the outcome of a prediction. If we adopt this weaker sense of the concept of prediction, then, it will be obvious that what carries out a prediction can be extremely simple. A simple-minded machine can predict something about another simple-minded machine on the basis of the statistical data it keeps. For example, the first machine may continually predict that the second one will display behavior *B* if the second machine displays behavior *B* more than 50% of the time. The failure of the prediction in this case will indicate that the second machine is displaying some unusual behavior. I will use the term prediction in this sense from now on.

Going back to McCarthy and Hayes's criticism, it is obvious that the criticism loses its power once we take prediction as prediction-in-the-small. Firstly, we do not need fantastic sequence extrapolators in order to carry out simple predictions. Extremely simple elements that capture extremely simple statistical relations will suffice. Then, it will be problematic to explain how intelligent behavior comes out of such simple predictions, but that is a problem regarding the theory, not the idea of prediction. Secondly, that the information from the world is encoded is everybody's problem and not just prediction's. Thirdly, simple prediction machines can not predict the outcome of a football match but they can indeed predict the play of light and dark and the movements of the players and the crowd. (The blob of light over there will continue to exist. That round shape moving right will continue to move right.) The predictions carried out will be of the most simple kind and will fail frequently, but they will still be predictions.

I have, therefore, removed the problems that make the idea of prediction implausible as a basis of cognitive activity. Notice, however, that prediction-in-the-small still holds the two merits I described in the previous section, just as prediction-in-the-large does.

1.5 Organization of the Thesis

In the next chapter, I will discuss some of the findings of psychology and certain models and ideas that have been proposed within cognitive science. In Chapter 3, I will try to show that low-level cognitive activities may be implemented by a prediction mechanism, a system organized on the basis of predicting sensory states. In Chapter 4 I will discuss how prediction can be related to cognition in a broader setting. Finally Chapter 5 summarizes my conclusions.

Chapter 2

On Modeling Cognition

In this chapter I wish to present the reasons and the background work that have motivated me to develop a model of sensory-motor organization. A thorough presentation of all the data related to cognition and all the models of cognitive activity can not fit into a book, let alone a single chapter. Therefore I will limit this discussion to subjects that I find only directly relevant. Surveys and discussions of the field with better coverage may be found elsewhere [1,12].

In the first part of this Chapter I will enumerate certain problematic issues and findings in cognitive science that await solutions, or just act as constraints on possible cognitive models in other ways. In the second part I will discuss some existing models and ideas that seem particularly relevant to the content of this thesis.

2.1 Problems of Modeling Cognition

Understanding and explaining cognitive activity is a difficult task. However, it may not appear to be so at first sight. After all, how we seem to be thinking looks introspectively explainable to us. Such an intuition, surely, will fail to prove correct. Our introspective picture tells us really very little about how we carry out cognitive tasks. (Dennett's "Consciousness Explained" provides a rather rich and deep discussion of this topic [8].) If one looks at the wealth of problems encountered in cognitive science research, it will become apparent that no simplistic method can explain the full range of human cognitive activity.

In this section, I will present and discuss a number of problematic issues regarding the attempt to model cognition. These are in no way all the problems one

will encounter. They are probably not even a representative set of such problems. They are only the ones I am most interested in. This presentation, I hope, will hint at certain difficulties, and will give an idea of the kind of problems a cognitive model has to face.

It is, of course, possible to dismiss any or all of these problems as side-issues that do not directly relate to the central problems of cognitive science. However, I take all of them seriously, and believe that, each one poses serious constraints on plausible models of cognition.

Models of cognitive activity can be roughly divided into two classes: those that model cognitive activity “on the surface” (or, what is usually called, “the phenomenological level”), and those that model it “in the deep”. The first type of models concentrate on mental events that are available to consciousness. They describe human behavior in terms of concepts such as beliefs, desires, and plans, and employ, what is usually called, a folk-psychological vocabulary. The second type of models describe cognition in terms of processes and mechanisms that do not map directly to (conscious) mental states, but that can perhaps explain how such states can come into existence, from the interaction of simpler processes. The approach I take is closer to this second type. However, if one opts for the deeper level, this means that he will have to do without all that introspective picture of the mind that acts as some kind of constraint on possible models. Therefore, one needs something else to explain, something else that can constrain the model. The problems and findings discussed in this section serve as my “something else”.

The points described below, I hope, will also suggest why the first line of approach, the one that works with surface-level concepts, will not suffice, and why things are a lot more complex than they seem “up there”.

2.1.1 Epistemic Access

Models of intelligent activities developed by Artificial Intelligence researchers are usually criticized on the ground that they really know nothing about the world. A computer program can process the predicates *mother(jane,joe)* and *mother(mary,jane)* and infer the predicate *grandmother(mary,joe)*. However, unlike a human being who would do the same inference, a program does not understand what mothers, grandmothers and children are. For it, all these predicate symbols are meaningless strings. We could install many thousands of rules and facts into the program, but what could possibly make those rules and facts pieces of knowledge about the world, rather than just more strings to process?

By the term epistemic access I mean devoidness of any epistemological problems (like the problems I mentioned above). For example, a model that supplies a detailed explanation of human behavior in terms of the workings of neurons, etc., would, I believe, be considered to have epistemic access by most contemporary philosophers. This would be so, because a computer implementing such a model would be doing exactly what the brain is doing and since nobody ever suspects that brains can have real knowledge about the world, the same has to be true for the computer. (However such a model may not provide solutions to problems within epistemology, since it would possibly not explain how human beliefs, knowledge, etc., are implemented by the brain's physiology.)

Not all researchers agree on the relevance of epistemic access to cognitive science. Fodor, for example, dismisses the problem by claiming that it is too difficult to solve as long as there remain problems in other natural sciences that await solutions [10].

2.1.2 The Frame Problem

Another problem suffered by traditional artificial intelligent models is the frame problem. The problem arises from the fact that the so-called frame axioms of a formal reasoning system seem to be unlimited if the system is meant to capture human-style common sense reasoning. It turns out that, most activities that we carry out with ease in our daily life require a huge amount of background knowledge. For example, as Dennett points out, one has to somehow take into account the fact that the friction between a tray and a plate is non-zero, if he wishes to carry the plate on the tray [7]. There is simply no end to the amount of knowledge needed to carry out even the simplest real-world tasks and if we try to list all the trivial things we know about the world, we will probably never finish the list. Moreover, a huge knowledge base will introduce problems of efficiently retrieving and using the knowledge. If these problems are attacked by default reasoning methods, this time, it will be problematic to cover the non-default cases (for example carrying a plate on a tray made of melting ice). In order to be unaffected by the frame problem, I believe, a system must successfully address at least the following issues:

- It must be able to learn new information from the environment and be able to learn new ways of learning.
- It must be able to distinguish between relevant and irrelevant aspects of a problem effortlessly.

- When its default behavior fails, it must be able to try new strategies for solving the problem.

2.1.3 The Whole and Its Parts

In this category, I take all problems regarding the relation between the whole and its parts. The matter has mostly been given importance by Gestalt psychologists and while most of their original claims and ideas have since been rejected, problems regarding the whole-part relation remain as serious as ever. If I try to state it simply: (1) a whole is not merely a collection of parts and (2) a percept can not be analyzed by itself if it is also part of a whole. There are a wealth of experiments in psychology showing that a human's perception of 'whole's and 'part's is not a simple matter, and the way high-level percepts are composed out of low-level percepts ought to be a complex matter. Therefore simplistic formulations such as:

triangle = closed + rectilinear + figure + three-sided

do not even approach to capturing the richness of whole-part relations in human cognition. The first part (1) of my statement of the problem stresses that parts of a whole can be in complex relations with each other in order to make up the whole. Fodor and Pylyshyn have used this point to show the inadequacy of connectionist models by suggesting that "Mary loves John" can not be represented as distinct from "John loves Mary" in a connectionist network [11]. A note of importance here: such complex relations between constituents are by no means unique to language and perhaps appear in their most sophisticated form in visual perception.

The second part of the problem states that perceptions are mediated once they are recognized as part of a whole (a fact realized by Gestalt Psychology). There exists a large body of psychological evidence in favor of this principle: Subjects briefly presented scattered **I** and **S** symbols report having seen **\$** symbols. Similarly, when shown two light sources that successively turn on and off, they report seeing a moving light. Higher-level mental representations also seem to follow the same scheme: subjects that are told stories with nonsense elements recall the stories with nonsense elements transformed to sensible ones. (This problem is also related to the problems I discuss in the next sub-section, Interaction Among Cognitive Mechanisms.)

2.1.4 Interaction Among Cognitive Mechanisms

Older models of the brain viewed cognition as a feed-forward process: Information entered the brain from the senses, was processed by layers of perceptual mechanisms, and arrived at high-level cognitive areas. These areas chose appropriate actions and sent them to motor mechanisms, which were responsible for carrying out those actions. Signals only traveled in one direction, up from the senses to the higher levels and then down from the higher levels to the motors. However, more recent neurological studies of the brain show that the situation is quite different. The presumed pathway of signals contains many loops and evidence suggests that high-level cognitive exercises (like imagery) involve a significant amount of processing in low-level areas [19].

This picture of the brain challenges traditional, strictly hierarchical views of cognitive activity where low-level processes and high-level thought are neatly separated. Such evidence suggests going for cognitive models that study the interaction between different levels of cognitive activity rather than models that are based on a notion of “information flow”.

2.1.5 Learning

Any theory of human cognition has to answer an elemental question regarding the nature of intelligence. The question can be posed in multiple ways: Is intelligence a collection of many different tools or is it something simple, regular, and principled? Has the evolutionary path that has resulted in the human brain created numerous specialized structures, adding, removing and fixing them along the way, or has it discovered a group of simple organizing principles that are powerful enough to solve any problem? Is the rich mental world we experience an outcome of the innate complexity of the brain, or is it a reflection of the complexity of our physical and social environment? In short: What is the importance of learning within cognitive activity?

It is difficult to give an exact answer to these questions, and it is probable that both sides of the questions have an element of truth in them. However it would not be wrong to say that learning has not been given the credit it deserves within existing computational models of cognition, if we compare the learning capabilities of such models with the learning capacity of human beings.

A number of experiments indicate that the role of learning in cognition is more significant than it might appear like. For example, congenitally blind people (people who are blind from birth, but may gain vision by a surgical opera-

tion), instead of instantly starting to see upon regaining their sight, develop their ability to see, rather slowly and painstakingly, possibly after several years.

But the most important line of evidence comes from the experiments of psychologist Jean Piaget. Piaget found out that very young children do not have as inborn knowledge, such elemental facts as the permanent existence of objects, but rather come to learn this by experience. For example, if a toy is shown to an infant and then is hidden under a piece of cloth while the infant is watching, the infant can reveal the toy by pulling the cloth away. This is repeated a number of times after which the toy is hidden under a second piece of cloth, again while the infant is watching. The infant, surprisingly, looks for the toy under the first piece of cloth.

This experiment and similar ones suggest two important things about the nature of learning. Firstly, children learn even such elemental things as the existence of objects. Secondly our knowledge about the basic regularities in the world is not in the form of declarative rules, but is in the form of procedures.

2.1.6 Adaptation

Another interesting aspect of human cognition is the incredible adaptive skills displayed by humans. For example, subjects wearing special glasses that make them see the world upside-down, can adapt to this situation in a few weeks and begin to use their sight without any special effort. When the glasses are removed, they again have to laboriously adapt to their original state of seeing.

This and similar examples of adaptive skills pose interesting problems for cognitive models. The extent of such adaptive skills is not known. However, they seem to be important enough to act as constraints on cognitive modeling.

2.2 Models of Cognitive Activity

There has been a vast amount work done on understanding and modeling cognition. These efforts and studies, however, have not converged on a single, generally accepted model. We can instead talk about many different, usually conflicting models, theories and ideas. In this section I will discuss certain models and ideas that are somehow relevant to the rest of this thesis.

2.2.1 Connectionist Models

The term “connectionist model” brings to mind the image of a model that involves a network of some sort. This is true. However the term is generally used in the literature to refer to certain types of networks called parallel distributed processes (PDP) or sometimes artificial neural networks (ANN). These models are said to involve distributed representations. There is also another class of connectionist models that use local representations instead of distributed ones. However these have neither received the wide popularity enjoyed by models using distributed representations, nor do they have a general framework. Therefore, I will only discuss models with distributed representations in this sub-section.

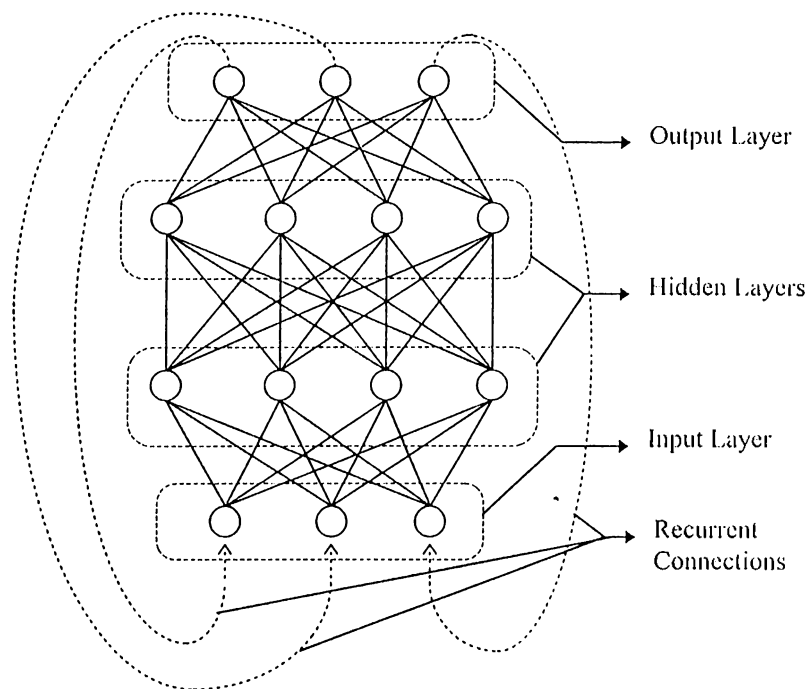


Figure 2.1. Components of an artificial neural network.

Figure 2.1 shows a typical ANN. The ANN is composed of an input layer, several hidden layers and an output layer. The recurrent connections drawn with dashed lines appear only in some of the models. The connections between nodes have associated weights. The nodes usually perform a weighted summation of their inputs and pass the result through a threshold function to generate their outputs. The network is started with a pattern of activation on the input layer. The output is then collected at the output layer, either after a single pass, or after multiple iterations. Such networks can be “trained” by using certain learning rules to approximate certain functions. Typically, they are trained to recognize certain classes of patterns.

It seems important here to discuss why these networks are said to involve distributed representations. If we call something in a model a representation, it is possible that this representation is naturally defined as a combination of other representations. For example, a representation of orangeness, could be somehow a combination of representations of redness and yellowness. This much obviously does not make a representation a “distributed” one. In PDP models, what generates distributed representations is the existence of hidden layers. The weights on the connections of hidden layer nodes affect all the nodes in the subsequent layers, and thus, they contribute to the patterns generated at the output layer. However the contribution of any single hidden layer node does not correspond to an observable quality. Therefore, it does not represent anything by itself. But taken together, the hidden layer nodes are responsible for accomplishing a certain function, and they can be said to be representing something (the class of patterns they are trained to recognize, for example). Since this representation is a function of many hidden layer nodes, it is called a distributed representation.

A number of problematic issues strike me about Artificial Neural Networks. In Chapter 3, I will try to sketch a model of perceptive and motor systems, where ANNs are usually considered as the right tools for modeling. Therefore, I would like to discuss ANNs in some depth here and try to show that they in fact have important shortcomings that make such models implausible. ANN models proposed in the literature differ in many aspects. Therefore I will limit my discussion limited to two important aspects that are shared by most of these models.

Hidden Layers

The predecessor of current Artificial Neural Networks was the Perceptron model suggested by McCulloch and Pitts. The Perceptron model consisted of only two layers of nodes: an input layer and an output layer. However, the model was severely criticized by Minsky and Papert, who showed that the Perceptron model was unable to compute the XOR function. Therefore, connectionist networks became dismissed for modeling cognitive activities and it was not until the introduction of hidden layer nodes, which incidentally made the computation of the XOR function possible, that Artificial Neural Networks re-gained popularity.

It should be clear that the explanatory power of ANNs derives largely from the existence of hidden layer nodes. This aspect has been sometimes criticized, since it makes the computation carried out by an ANN unintelligible. This is really a minor problem since the internal workings of a cognitive mechanism does not have to be intelligible to us, the observers. However, it becomes a severe problem once we realize that those workings are also unintelligible to other

cognitive mechanisms. There is no way the information captured by a hidden layer node can be re-used by a system other than the one the node is part of.

Let us consider a network that recognizes alphabetical letters. In order to successfully do this, the hidden layer nodes must somehow be processing the way line segments that make up the characters are joined with each other, their orientation, etc. (If they are not processing these kinds of relations, they can not reliably recognize letters.) But if this is the case, then it would be nice to use this wonderful piece of machinery in other tasks, say face recognition. However, this is not possible since that information is “hidden”.

The way to work around this problem is simple: represent (and recognize) line segments, junctions and orientations **explicitly**. Make the letter recognizer work on such representations, as well as the face recognizer. Moreover, make the face recognizer use the letter recognizer for recognizing v-shapes o-shapes or T-shapes. In short, organize the system in a systematic and hierarchical way to maximize the utility. But if we do all of this, then there is very little left of the spirit of PDP models that are based on generating the ultimate result from the raw input at one big step. If one accepts this picture, then the really interesting point becomes the hierarchical organization of mechanisms and the interactions between them, rather than the wondrous distributed representations.

Supervised Learning

Most ANN simulations are based on a certain kind of training called supervised learning. In this method, the result computed by the network is compared against the correct result. If the output of the network is wrong, then the weights in the network are updated according to some learning rule (for example, error back-propagation).

What I want to argue is that this is not a correct account of human learning. Let us compare the way two agents—a human and a robot controlled by an ANN— learn to drive a car. Both of them will make frequent errors to begin with. The way the robot learns will be error-driven. For example, if it wrongly steers the car left in situation A, it will have to be told something like: “Whoops, you should really have steered right in situation A”. The robot will then update its weights to take that information into account. After being trained about how it should steer in many different situations, and after many weight updates to produce the correct steering behavior, it will presumably become a skillful driver. What about the human? She will probably make similar errors and, for example, will also steer the car left in situation A. Like the robot, she will hear a warning

from the teacher and will try to take it into account. But one more thing: The human will also record what happens when she steers left in situation A. As a matter of fact, what the human learns will not be in the form of: “What should I do in this situation” but rather in the form of: “What happens when I do that in this situation”. Unlike Artificial Neural Networks, human brains record the outcome of actions and therefore require, not the correct output, but the outcome of their action in order to learn new skills.

2.2.2 Symbolic Models

In their article entitled “Connectionism and Cognitive Architecture”, Fodor and Pylyshyn criticized connectionist models of cognition, on the grounds that such models were unable to account for representations with combinatorial structure [11]. Fodor and Pylyshyn argued that representations we employ can not merely be **collections** of active objects; they should somehow be **structured** entities themselves. For example, the representation of the idea “John loves Mary.” should be different from the representation of the idea “Mary loves John.”, although both ideas involve the same active objects (“Mary”, “John” and “loves” in this case). A symbol system could represent the two ideas in distinct ways (for example with the representations *loves(john, mary)* vs. *loves(mary, john)*). However, a connectionist system could not make this distinction in a systematic and productive way, since its representations were essentially vectors of such active objects with no additional structure. Connectionist models, therefore, lacked the kind of representational power we seem to possess.

Fodor and Pylyshyn’s claims have been answered by authors in the connectionist school, who referred to a number of connectionist systems that processed combinatorial representations [21]. I do not want to look further into the details of this discussion. However, I wish to point out that, in this short history of a recent discussion one can find the reason why the computer metaphor is so critical to the study of cognition.

Computers are not only simulation tools for cognitive science. They share with human beings, the interesting capacity of employing combinatorial representations. So, they are in some way different from other, older metaphors, like clocks or electric fields that were proposed to serve the same purpose. Moreover, this relation between computers and human brains is not one of simple analogy. That is, it is not that computers and brains have a common property (the property of being sensitive to combinatorial structure) that establishes the relation between these. Rather, if something is sensitive to combinatorial structure as such,

then its operation with regard to this combinatorial structure can be characterized as **computation**.

This insight has been the basis of a great amount of research in cognitive science and artificial intelligence. Symbolic models have been proposed covering a wide variety of human skills such as reasoning, problem solving, vision, and language, giving rise to many practical systems, frameworks, and fields of study.

However, the symbolic models, I believe, have a number of serious shortcomings. There has been too much emphasis on knowledge installation (consequently, too little emphasis on learning), a general reluctance to distinguish between what is just a practical system, and what is meant to faithfully model human behavior, and too much confusion on problems of epistemic access. In general, symbolic models fail to explain almost all the problematic issues I presented in the previous section, as well as many others.

These shortcomings should not be blamed on the fundamental assumption that symbol manipulation is relevant to cognition. They are mostly due to the difficulty of the problem at hand and the unavailability of a sound methodological framework. It is not at this moment evident which aspects of symbolic models will prove to be relevant to cognition in the future and which will not. In any case, it is not my concern here to make an in-depth analysis of these issues.

2.2.3 Some Other Directions in Cognitive Science

Certain other approaches to cognitive modeling seem to be worthwhile to look at here. There exists another class of connectionist models apart from the one I discussed in Section 2.2.1, namely, those that involve local representations. The first exemplar of such models is Hebb's cell assemblies [14]. Hebb suggested that the simple organizing rule for the brain was the establishment of connections between neurons that got activated simultaneously. This resulted in groups of densely connected neurons that Hebb called cell assemblies. Cell assemblies, he argued, would act as recognizers for frequently occurring patterns and would trigger each other to start chains of thought (a phase sequence). Hebb's theory had the exciting character of trying to explain complex mental phenomena with an extremely simple rule. However his ideas were rather implausible and lacked sufficient explanatory power.

Since the time of Hebb, many connectionist models that employ local representations have been proposed and studied. Recently, there has been a good deal of interest in networks where synchronization between the activation patterns of

nodes is used for establishing bindings between them [20]. Such synchronization has also been observed to take place between actual neurons in the brain [13, 19].

Another interesting line of research is carried out in the field of situated robotics. The practitioners of the field try to build robots that act and display skills in “real” environments (like crowded offices). An interesting architecture for situated robots has been proposed by Rodney Brooks, who has given it the name subsumption architecture [2,3,4,5]. Subsumption architecture models an agent with a set of hierarchical layers that sit on top of each other (see Figure 2.2). Each layer is implemented with a number of simple mechanisms (typically finite automata), whose interaction results in a certain type of behavior that corresponds to a level of competence for the robot. For example, the bottom-most layer may be responsible for generating some walking behavior, in which case, particular finite automata will control the harmony between the legs of the robot. Higher layers generate behaviors by modulating the workings of lower layers. For example, a collision-avoidance layer can monitor sensors of the robot and adjust the workings of the walking layer such that the robot will not bump into objects. Each layer is debugged and made robust in itself so that it can carry out the activity it is responsible without the interference of higher layers. As a result, the architecture of higher layers can be kept relatively simple since they do not have to attend to the low-level behaviors.

Brook’s subsumption architecture is successfully used in many robots that can operate in the real world. However, the important question regarding this line of work is whether this approach will scale up to more complicated skills or not. Currently, Brooks and his colleagues are working on an android that is planned to have a physical structure similar to a human being and a level of competence in some low-level tasks comparable to humans.

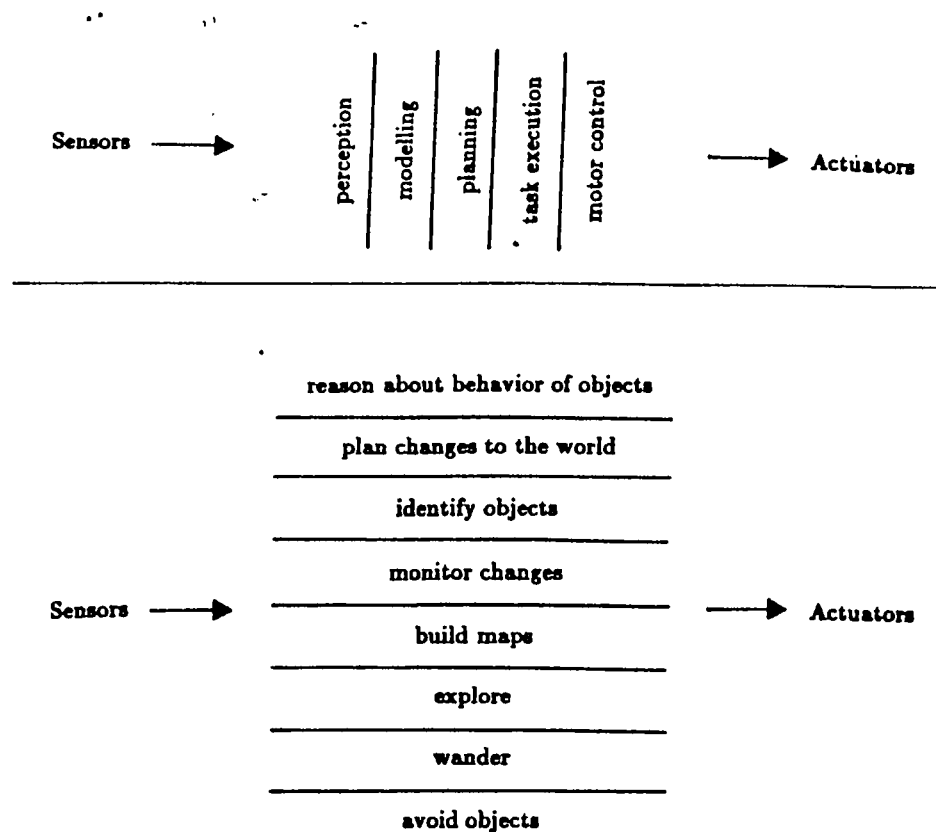


Figure 2.2. Traditional vs. behavior-based decomposition of a mobile robot control system.

2.2.4 Constructivism: The Search for “Order from Noise”

One of the most important theories on the nature of human intelligence is constructivism, which has been developed by Jean Piaget through extensive studies of child development. According to constructivism, our knowledge and skills are not innate but are actively constructed through our interaction with the world. Piaget describes the goals of the constructivist line of research as follows [17]:

Fifty years of experience have taught us that knowledge does not result from a mere recording of observations without a structuring activity on the part of the subject. Nor do any a priori or innate cognitive structures exist in man; the functioning of intelligence alone is hereditary and creates structures only through an organization of successive actions performed on objects. Consequently, an epistemology conforming to the data of psychogenesis could be neither empiricist nor preformationist, but could consist only of a constructivism, with a continual elaboration of new operations and structures. The central problem, then, is to under-

stand how such operations come about, and why, even though they result from nonpredetermined constructions, they eventually become logically necessary.

The two problems of constructivism that await explanation, (1) how we are able to put an order to the massive amount of information we have to deal with, and (2) how is it that everybody agrees on the same “order”, can be reduced to one and the same problem by assuming that order resides in the environment and the agent constructs his own version of reality by “transfer of structure” (though not all constructivists would agree on this reduction). However by combining the two problems, we do not arrive at a fundamentally easier problem. The task that awaits constructivism, explaining how order can be constructed out of what seems as noise, is still one of the hardest that can be envisaged.

A translation of constructivist ideas into the domain of computer science was attempted by Drescher [9]. Drescher represented the knowledge of the agent as schemas: rules that consisted of a context, an action, and a result. The context and result were conjunctions of “items”, which were the outputs of low-level recognizers. A schema encoded the knowledge that when the items in the context were active, carrying out the action would result in the activation of the items in the result with a better-than-chance probability. Drescher showed that such schemas could be found by simple statistical analysis. Once reliable schemas were found, they could be used for generating goal directed behavior, by looking for chains of schemas that took the agent from the current context to a certain goal.

Although Drescher’s work contains many interesting ideas, it lacks sufficient explanatory power to account for human development. First of all, his schema mechanism does not contain hierarchical structures and is therefore limited in its extent. Moreover, every instance of a rule is treated alone and these are not aggregated into more general schemas. Lastly his system functions in a too simplified setting and it is not obvious whether his solutions will scale up or not.

Although I believe that constructivism is correct in its premises, and that the elaboration of a constructivist framework should be the primary research goal before cognitive science, I will not attempt to sketch such a framework in this thesis. My work is related to constructivism in a different manner. In the following chapter, I will describe a system that can account for human low-level sensory-motor system that acts as some kind of pre-requisite for higher-level cognitive systems that I believe should have a more “constructivist” flavor. I do not regard my system as a constructivist one because there will be relatively little stress on “active construction” as opposed to “passive impression”. However my

work hopefully suggests how order can emerge out of noise and how the development of a constructivist model of cognition can be guided by general principles and ideas.

Chapter 3

A Prediction Mechanism

This chapter describes a hypothetical system that is meant to be a model of human low-level sensory-motor mechanisms. The first three sections of this chapter are introductory: they lay out a number of intuitions, observations and ideas that will hopefully present the rationale behind positing such a system. In Section 3.4, I present a number of guiding principles that have aided me in the development of the hypothetical prediction mechanism. Section 3.5 describes an abstract setting for the environment and the sensors/actuators of an agent. The remaining sections incrementally develop the prediction mechanism, by introducing new types of components.

3.1 An Intuitive Opening

Imagine yourself climbing a staircase. When you are about to climb the last few steps, the lights suddenly go off. Without bothering yourself too much, you go on to climb the stairs in the dark. But just when you climb that last step something bizarre happens: your foot does not touch the ground when it should have. In fact, you have already finished climbing all the steps. Next, imagine that you see a very clean window. Unable to stop yourself, you attempt to touch it, knowing that this will leave a nice, dirty finger-print. But you feel really strange as your finger touches nothing. You realize that the window frame is empty.

Why am I giving these rather trivial examples? Because I believe that there is something interesting going on here. In both of these examples what we are sensing is not the existence of a stimulus, but the absence of one. So what must be going on these cases must be something like the following: A mechanism within our cognitive system should be predicting the time and the type of the

stimulus we should get if everything went well, and signaling an error when the particular stimulus does not arrive at the right time. (That is, I take it that the strange feeling we experience in these cases is somehow related to the error signal.)

The examples I gave above are not really special in any respect. (In fact I can recall many other cases where I had similar experiences.) So, I infer from these that predictions are carried out routinely during all our activities and something within our mind is constantly keeping track of what we are going to sense in the near future. Of course, only one in a million of such predictions ever fail and we are seldom alarmed. The bulk of the predictions are carried out silently without us noticing them, but they are needed to ensure that we become aware if anything goes wrong.

So, we can say that there exist at least three mechanisms regarding our low-level sensory-motor system: one that initiates actions, one that processes sensations, and one predicting future sensations. (These might, however, be one and the same mechanism, but we do not know that yet.) I will call this latter one **the prediction mechanism**. We know that the prediction mechanism, in order to fulfill its function must adapt to new sensory-motor tasks and must be able to predict sensory patterns in complicated activities with good precision. This means that we are talking about a delicate piece of machinery here. And if there exists such a machinery then its functional role must be far beyond providing error signals in one in a million trials. If I were to design the architecture of a cognitive agent, and if I had such a prediction mechanism at my disposal, I would find many important uses for it. Nature is a far better designer than me, and it seldom wastes resources. So I conclude that the prediction mechanism has a crucial function within human cognition.

3.2 A Hierarchy of Predictions

Predicting the typical feedbacks of our actions is no doubt an important task. However, we should realize that we really benefit from such a prediction, not when everything goes as predicted, but when the prediction fails. The failure of a prediction creates an important piece of information that would not be available if the prediction had not been carried out: that there is something atypical, strange, and unusual in what happened.

Consider the following example: An organism can sense two pieces of information, *A* and *B*, and typically *B* is sensed right after *A* is. Now, assume that *B*

starts to be predicted whenever *A* is sensed. Let us call the failure of this prediction *C*. Of course, *C* is just another piece of information, and as a matter of fact, it is a “higher-level” piece of information than both *A* and *B*. Therefore, instead of saying that the prediction has failed, we may simply say that “*C* is sensed”. This means that, the outcomes of predictions are not different from original sensations, and they can themselves be the subjects of yet higher level predictions, giving rise to even more high level pieces of information. There is, at least in principle, no problem in talking about a **hierarchy of predictions**, that is composed of many layers. Information that belongs to the top of the hierarchy, then, is the information that is supposed to be the most high-level.

I suggest that, our low-level sensory-motor system implements just such a hierarchy. At the bottom of the hierarchy lie simple sensory and motor activations, where any single piece of information does not carry much importance and can be discarded without any serious problems. On the other hand, the top of the hierarchy contains pieces of information that are much closer to the percepts we get from the environment.

Notice, how such a picture of perception differs radically from traditional views that deal with issues such as recognition, feature detection, and the like. Unlike traditional ones, there is no question of what to recognize, and what to ignore, which features to detect and which to discard. It, as a primitive picture, has the potential to account for interesting properties of human perception such as flexibility, plasticity, and interactivity.

3.3 Functions of the Prediction Mechanism

I have claimed that a mechanism that can predict future sensory states would have a number of possible usages within a cognitive system. In order to discuss this, the idea of ‘predicting future sensory states’ must be somewhat clarified. Basically sensory states can be predicted in three ways:

- 1) As direct feed-backs of motor commands: Almost all muscles in the human body send acknowledgment signals to the brain when they are activated. Therefore the arrival of the acknowledgment signal can be predicted from the initial motor activation.
- 2) As indirect feed-backs of motor commands: Motor commands affect the sense organs in a systematic way. For example when we are making drawings or writing with a pencil, we move our hand by initiating motor commands and see how our hand (and the pencil) moves in return. The move-

ments seen by the eyes can be predicted again from the initial motor activations.

- 3) As continuations of previous sensory states: Sensory states can be predicted from previous sensory states. For example, if one is listening to a rhythmic beat, he may predict when the next beat is going to be heard.

Predictions of type (2) and type (3) can not be the result of totally innate (genetically fixed) mechanisms. Humans continually learn new behaviors and skills which result in new feed-back patterns. Therefore, the prediction mechanism that is responsible for these predictions must be non-innate.

A prediction mechanism that successfully predicts future sensory states could serve a number of purposes within the cognitive system apart from providing error signals. We may list some of these possible uses as: carrying out goal directed activity, generating high level percepts, and revealing interesting information. Below, I will discuss each of these.

3.3.1 Carrying Out Goal-Directed Activity

A mechanism that predicts future sensory states can perhaps be "executed" in reverse direction for reaching "desired" sensory states. For example, if the mechanism somehow encodes that sensory state B comes after motor action A, it may start action A whenever state B is desired. Similarly, such a mechanism can be used for avoiding "undesired" states. If state B in the previous example is this time undesired, the mechanism can inhibit the activation of action A.

Goal directed activity can, of course, be carried out also by higher level cognitive mechanisms. The scheme that is suggested here for this is that the automatic, repetitive acts that do not require special attention and planning are carried out by the prediction mechanism whereas activities that require complex planning and attention are carried out by higher level mechanisms. However higher level mechanisms can use the prediction mechanism for carrying out parts of such activities.

3.3.2 Generating High Level Percepts

Considering the complexity of human senses and activities, successfully predicting future sensory states is an extremely difficult task. Therefore the mechanism that is accomplishing this with a good degree of success must also be organizing the information on the sensory-motor level, aggregating initially "meaningless" sensory activations into more "meaningful" percepts that can then be used by

higher level cognitive mechanisms. For example, consider how the changing of the image on one's visual field can be predicted while moving forward. In this case the projections of objects that are closer to the subject would move rather quickly whereas projections of objects far away would remain relatively stable. A mechanism that successfully predicts these changes, then, must be able to distinguish between near and far objects. Therefore useful concepts such as nearness and farness may be emerging at this level of cognitive activity. If powerful organizing rules can be found that result in a successful prediction mechanism, one of the side results of these rules may be the emergence of interesting, high-level percepts.

One of the most difficult aspects of studying cognition is dividing it into functional parts where each part may be studied more or less independently. Studying low level prediction lets us to cut a slice in cognitive activity where we replace the question of "How can we build an intelligent machine?" with the question of "How can we make a prediction mechanism that successfully predicts sensory states?" Still a very difficult question but at least a more rigorous one.

3.3.3 Revealing “Interesting” Information

Sensory information that a cognitive agent acquires from the world can be roughly divided into two categories: information that is "predictable" from past sensory states and actions, and information that is not predictable as such. Information of the latter type, in general, is the type that is worthy of more interest. Consider the example of eye movements. When somebody moves his eyes a few degrees to the left, the image he sees will slide a little bit to the right. This new image can be roughly predicted from the previous image and the amount of the eye movement. However if part of the image changes in an unpredicted way, i.e. if predictions about part of the image fail, it may be concluded that something in the external world has moved. It would be extremely difficult to recognize this movement (since the whole image has changed anyway) if no prediction had been done about the new image.

A prediction mechanism, thus, can act like a filter that reveals interesting events in the environment. It automatically distinguishes between what is the normal outcome of events and what is non-standard, original and interesting.

3.3.4 Remarks

I have listed three very important functions of a hypothetical prediction mechanism. It may be evident by now that I ascribe a very important and special role to this prediction mechanism. If the prediction mechanism does support these functions, then it could account for most of what we call as low-level processing in the brain. This view radically differs from conventional theories of low-level systems, so it may be worth elaborating this point a little bit.

To some readers, the claim that low-level processes in the brain are carried out by a prediction mechanism may seem absurd. After all, the claim is not based on any neuro-scientific evidence. In fact, it is not based on any real evidence other than my introspections (which would hardly count as solid evidence). I suppose the real evidence in support of my claim lies in its explanatory power. The existence of a prediction mechanism that continually adjusts itself to make better predictions and which, on the way, carries out the functions I have described above explains a number of peculiarities regarding human cognition. It explains, firstly, how we are able to perceive the world, carry out motor actions, and sense interesting things. It explains how we get better at doing things through experience. It explains why our perceptions change through time and how they get more refined as we keep on perceiving a certain class of things. All of these explanations, and still others, I believe, make my claim a viable one.

Now, how can I give such explanations? I must first show that the prediction mechanism I am suggesting is not a miracle, and if one sits and tries to design such a mechanism, he will at the end come up with a number of simple rules, whose systematic application will result in the system we are looking for. If we can design a prediction mechanism, then a similar one could exist within the brain, and if the prediction mechanism is so good at providing explanations, then it seems plausible to consider all of this as serious evidence for that existence.

So, we come to the question of what such a prediction mechanism would look like. Unfortunately this question is seriously difficult and is not fully answered by the work presented here. However, this chapter will present a way to attack the problem. We are going to start by presenting a very simple prediction mechanism that can only make successful predictions in the simplest cases. Then we will incrementally develop it to cover more sophisticated cases. This development, hopefully, will convince the reader that a powerful, and yet simple prediction mechanism, is possible.

3.4 Some Guiding Principles

It is a very difficult task to design a system that does not have a clear specification. The existence of a prediction mechanism is something we have taken as a hypothesis. However we are unable to clearly state its function and behavior. For this reason, our designing process will have the nature of an exploration whose outcome or destination is not really known. In order to keep this exploration focused, we will state some guiding principles which we will use throughout the development of the prediction mechanism.

In Section 3.3, I listed three functions of a prediction mechanism: enabling goal-directed activity (of a simple kind), generation of internal states that correspond to high level notions, and lastly, distinguishing between interesting and non-interesting events. While we are working out the details of a prediction mechanism, we will make use of three guiding principles founded on these functions.

Guiding Principle 1

For carrying out goal-directed activity, elements that make predictions must also work in the reverse direction in order to find out how to achieve a goal state. When the elements work together for making better predictions, they must also enhance the goal-directed behavior of the overall system. Therefore, our first guiding principle will be enhancing goal-directed activity.

Guiding Principle 2

For successfully distinguishing between interesting and non-interesting information, the prediction mechanism must exploit all opportunities for making successful predictions. The more powerful the prediction mechanism is, the more interesting will be the things that the system failed to predict. Therefore, another guiding principle we will use is maximizing predictory power.

Guiding Principle 3

If the prediction mechanism can predict with a good degree of success by using a limited amount of resources, this means that it is making use of notions that are also useful for higher level cognitive skills. Therefore, minimizing the amount of resources used should be one of our guiding principles in designing a prediction mechanism.

3.5 Setting up the system

The human cognitive system and its interaction with the body and the environment are complex issues that contain numerous details. Studying the organization of a prediction mechanism in this complex setting is rather difficult. Therefore I will define a much simpler setting where the development of the ideas in this chapter will be easier.

3.5.1 The Composition of the Cognitive System

The human brain is a complex structure that is made up of densely connected cells, called neurons. Some of the neurons in the brain are stimulated by signals coming from various receptor cells in the body or the sense organs. These are called sensor neurons. Still others, when stimulated from within the brain, send signals that control muscles. These are called motor neurons. The connection patterns between the neurons is known to be dynamic, i.e. some of the existing connections can be broken and new ones can be established. Here I will present a scheme for a cognitive system that covers these aspects.

The “brain” of the cognitive agent is a complex network of simple elements and can be roughly divided into three layers. The bottom layer, called the interface layer, consists of a fixed set of sensors and actuators. Certain fixed structures, that directly link sensors to actuators to create simple reflexes, also belong to this layer. The middle layer, called the prediction mechanism, can grow or shrink dynamically by the addition or removal of elements. The basic organizational principle of this layer is predicting the activity of the interface layer. There is also a higher layer which, I assume, contains various higher level mechanisms that are not critical to our discussion.

3.5.2 Time Scales

For simulating the passing of time in a simple way, we will assume that the environment operates in discrete time steps. The sensors, actuators and other elements of the cognitive system will also operate in discrete time steps. It is convenient to assume that these internal elements operate faster than the things in the environment, so that a piece of information about an external entity may be processed before the information becomes obsolete. Therefore, we will distinguish between these two time scales. The environment will operate in external time steps, whereas the elements will operate in internal time steps. We will as-

sume that a single external time step corresponds to multiple internal time steps. Sometimes, I will refer to external time steps as simply, time steps.

3.5.3 The Interaction Between the Agent and the Environment

The agent interacts with the environment through its sensor and actuator elements. Below, I will describe these two types of elements.

Sensor Elements

Every sensor element is connected to a physical sensor. This physical sensor is assumed to check a binary condition in the world (such as if a finger touches something or not). When this condition is satisfied, it activates the sensor element. In this case, the sensor element stays active throughout one external time step. Sometimes I will call sensor elements simply sensors. I will graphically represent sensors with a circle that has an upward arrow inside. The name of the sensor will be written below this circle. If I want to show that the sensor is active I will place the symbol **A** to the left of the circle.



Figure 3.1. Graphical representations of sensors

Figure 3.1 shows two sensors, $s1$ and $s2$. Sensor $s2$ is explicitly shown to be active.

Actuator Elements

Actuator elements are connected to motors (or muscles) in the body of the cognitive agent. These motors make parts of the body move in a certain way. For simplicity, we will assume that motors work in a binary way and they are either active, carrying out a certain action in a certain way, or they remain inactive. Whenever the actuator element is activated, it will start the motor it is connected to. Sometimes I will call actuator elements simply actuators. I will graphically represent actuators with a circle that has a downward arrow inside. The name of the actuator will be written below this circle. If I explicitly want to show that the actuator is active I will place an **A** to the left of the circle.



Figure 3.2. Graphical representation of actuators

Figure 3.2 shows two actuators, $a1$ and $a2$. Actuator $a2$ is explicitly shown to be active.

3.5.4 Events

If we have an interface element (either a sensor or an actuator) named k , the event that k is active within a certain external time step is denoted by $k.A$ (We read this as the activeness of k). Similarly, the event that k has **become** active (when it was not active at the previous time step) is denoted by $k.A\uparrow$ (read activation of k). Conversely, the event that k has become inactive at a time step is denoted by $k.A\downarrow$ (read deactivation of k).

All these events have certain frequencies. The prediction mechanism will record and use these frequencies to statistically observe the relations between such events. However, in order to simplify the presentation of the prediction mechanism, I am going to talk about probabilities instead of frequencies. It must be remembered that the prediction mechanism approximates all such probabilities with frequencies. For example, $P(k.A\uparrow)$ will denote the probability of element k becoming activate at a certain (external) time step. A prediction mechanism has no way of knowing this probability. Therefore we will assume that it will record and use the corresponding frequency instead.

The sensors and actuators work within a causal and mechanistic world. Therefore the activation patterns of these elements are related with each other.

I will use conditional probabilities to express statistical relations between events. However I will slightly change the notation of conditional probability in order to capture the time differences between events. If we have three events A , B and C :

- The expression $P(C|_n A)$, read n -delay conditional probability of C under condition A , will denote the probability of event C happening in a certain time step, if event A had happened n time steps ago.

- Similarly, the expression $P(C \uparrow_n | A, B)$, read n -delay conditional probability of C under condition A and B , will denote the probability of event C happening in a certain time step, if event A and B had happened n time steps ago.

Again, the prediction mechanism will approximate such conditional probabilities with the corresponding frequencies if they are ever needed.

3.6 The Basic Predictor

In this section, a component of the prediction mechanism, the basic predictor, is introduced. In the first sub-section I discuss how basic predictors can be established. In the second sub-section, I try to show how they can help with goal-directed activity.

3.6.1 Establishing Basic Predictors

Let us remember our setting: We have an agent situated in a complex environment. The agent has a set of sensors and actuators. Initially, the actuators are activated by certain reflexes or by other mechanisms that are irrelevant here. While the agent is interacting with the environment, the sensors and actuators will become activated and deactivated. Some of these activations and deactivations will be causally related with each other. This enables us to predict such internal events by looking at other internal events. For example, a sensor a may be sensing the acknowledgment signal of a muscle that is activated by an actuator z . In this case, the sensor will become active exactly one time step after the actuator becomes active. Therefore,

$$P(k.A \uparrow | z.A \uparrow) = 1$$

Of course, not all elements are so directly related with each other. The important point is that, related elements will almost always be statistically related with each other. We may establish predictors between statistically related elements. Basically, we can establish a predictor q from interface element k to interface element l if :

$$P(l.A \uparrow | k.A \uparrow) > P(l.A \uparrow)$$

The above inequality means that l gets active more often than usual if k had got active one time step ago. This is the basic rationale behind the idea that something can **predict** the activation of l when it sees the activation of k . We

will call predictors established between two such interface elements basic predictors. I will call the conditional probability $P(l.A\uparrow \mid k.A\uparrow)$, the **reliability** of basic predictor q . I will call the elements k and l , the **source** and the **target** of q respectively. Similarly, I will call the events $k.A\uparrow$ and $l.A\uparrow$, the **source event** and the **target event** of q respectively. Lastly I will call the time difference between the source and target events the **delay** of q .

I will graphically represent predictors with an arrow from the source element to the target element. The event at the source or target will be shown to the right of the arrow's connections with the elements, by either an upward arrow sign (if the event is an activation), by a downward arrow sign (if the event is a deactivation), or by the absence of a sign (if the event is activeness). The delay of the predictor will be written over the arrow, while the name of the predictor will be written under the arrow. For example, a predictor named $p1$ from an actuator named x to a sensor named a , whose source and target events are the activation of x and the activation of a respectively would be shown graphically as follows:

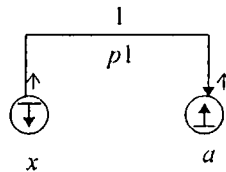


Figure 3.3. An example predictor

So far, we have looked at how to establish a predictor that predicts an element's activation from another element's activation. Of course, it is in principle possible to use events other than activation for making predictions. Establishing all possible predictors is in accordance with Guiding Principle 2 that I have discussed in Section 3.1. However, such an approach would not necessarily comply with guiding principles 1 and 3. This leads to an important point that will recur later in this thesis: Deciding which types of predictors to establish is an engineering decision and must be studied in detail if we want to implement a working prediction mechanism. However, such a detailed analysis is beyond the scope of the work presented here.

The same line of argument applies to the delay between source and target events. We can try to establish predictors between events that happen in the same time step, as well as events that are two, three or more time steps apart. In this thesis, I will only use predictors with either 0 or 1 delay. However, predictors with longer delays may be found to be useful in a more detailed study.

In the rest of this chapter, I will assume that only the types of predictors we discuss are established between elements, though other types are also possible. However I will keep the notations and graphical representations as general as possible.

3.6.2 Goal-Directed Activity with Basic Predictors

By Guiding Principle 3 the predictors that are established must also contribute to goal-directed activity. This is carried out in a rather simple way:

All elements (sensors, actuators or predictors) have an associated **goalness**. The goalness of an element will be assumed to have a value in the continuous interval between 0 and 1. I will say that an element is a **goal** if its goalness is over a predefined threshold (we may assume that this threshold is close to 1). Graphically, I will show that an element is a goal by placing a **G** to the left of its graphical representation. Now, let us assume we have a predictor that predicts a target element's activation from its source element's activation. Let the predictor's reliability be equal to 1, which means that the target element's activation **always** follows the source element's activation at the predicted time. Such a predictor will not only carry out predictions, but it will also carry the goalness of its target to its source. So, if the target element is a goal, the predictor will make the source element a goal as well. If the source element is a sensor, the goalness can be further propagated to other elements. I will also assume that, if actuators become goals, they are automatically activated. Therefore, if the source of the predictor is an actuator, the predictor will activate that actuator upon carrying the goalness. Since the predictor is completely reliable (that is, its predictions always prove correct), the target will also become active after the activation of the source.

The same idea applies to predictors with reliabilities that are less than 1. However this time, the predictor will not carry the goalness of the target to the source as it is, but will decrease this goalness on the way. The less reliable a predictor is, the more reduction will be made to the goalness. This can be easily achieved if the predictor multiplies the goalness of the target with its reliability, before carrying it to its source.

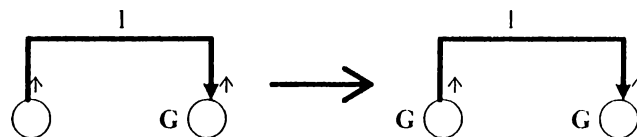
The same mechanism also applies to propagation of what I will call **avoidedness**. Avoidedness is in some sense the dual of goalness and these two attributes cancel each other out. I will say that an element is an **avoided** if its avoidedness is over a predefined threshold. Graphically, I will show that an element is an avoided by placing a **V** to the left of its graphical representation. Avoidedness is

propagated by predictors from their target to their sources in the same way goalness is. It is also decreased according to the reliability of the predictor.

In order to better present these ideas graphically, we will somehow simplify the workings of predictors by distinguishing between those predictors whose reliability is close to 1 and those predictors whose reliability is relatively far from 1. I will call the first type **reliable** predictors and the second type **unreliable** predictors. I will graphically represent reliable predictors with thick arrows and unreliable predictors with thin ones. Also, I will represent interface elements whose type (sensor or actuator) need not be specified with an empty circle.

A predictor's carrying goalness and avoidedness is shown graphically in Figure 3.4. I will call such graphical representations "rules", since they capture important aspects of workings of the prediction mechanism.

Rule 1.a



Rule 1.b

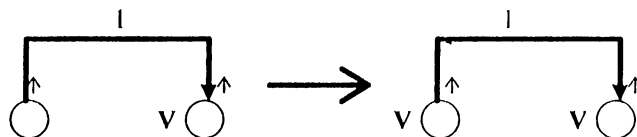


Figure 3.4. Rules for basic predictors to propagate goalness and avoidedness

The rules above represent how goalness and avoidedness are propagated in a simplified way. The right side of the rules show the state of the elements one internal time step after the state shown by the left side. Note that the predictors used in the rules are reliable predictors (they are drawn in thick lines). In reality, predictors should work on graded attributes. However, this kind of representation is easier to understand and such rules will be used throughout the presentation of the prediction mechanism.

An interesting situation arises when two different predictors carry goalness to the same element. Should these two goalness values be added, or should we take the maximum one? It turns out that we should add them if the target events of the

two predictors are independent. If they are not independent, we should take their maximum. It is in principle possible to learn this by a statistical analysis of the relation between the two events. However, such a statistical analysis cannot be performed by simple predictors, and if the analysis is done by an external system, then the prediction mechanism I am proposing will lose its simplicity. Therefore, we will always take these events to be dependent ones and take the maximum of the two values. This is because, it is possible to have a great number of dependent events and adding the goalness values associated with all of these would result in artificially increased goalness values. We also take the maximum of avoidedness values, if multiple predictors carry avoidedness to the same element.

3.7 Meta-Predictors

In the previous sub-section I have described how to establish predictors between two statistically dependent interface elements. Such predictors can capture only the simplest relations between elements. In this section, I will describe how predictors can be built in a hierarchical way to form a more complex system that can make better predictions.

Let us use a simple example to better present this idea: The example concerns sensory-motor control of a simple robot hand for grabbing objects.

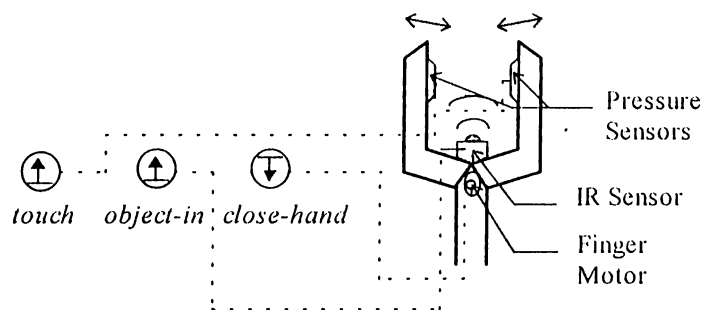


Figure 3.5. A robot hand connected to sensors and actuators

The robot hand is a simple device that has two fingers, a finger closing motor that can temporarily close the fingers, a pair of pressure sensors that sense if the fingers touch anything and lastly, an infrared sensor that senses if there is anything between the two fingers. The pair of pressure sensors, the infrared sensor

and the motor are connected to two sensor elements and one actuator element that are named *touch*, *object-in*, and *close-hand* respectively.

We assume that the agent is acting in the world and trying motor actions (through reflexes or other means) at different times. Meanwhile, we record the frequencies of different events to approximate conditional or unconditional probabilities. One of the motor actions carried out by the agent is activating *close-hand*, which makes the motor hand close its fingers. In this case, if there is an object between the fingers, the fingers will touch the object and pressure sensors will sense this and activate the sensor element *touch*. Let us call the probability of there being an object between the two fingers P_{object} . This probability will be equal to (assuming that the infrared sensor that recognizes this works perfectly) $P(\text{object-in})$. Therefore:

$$\begin{aligned} P(\text{touch} \cdot \mathbf{A} \uparrow \mid \text{close-hand} \cdot \mathbf{A} \uparrow) = \\ P(\text{touch} \cdot \mathbf{A} \uparrow) + P(\text{close-hand} \cdot \mathbf{A} \uparrow) \cdot P_{\text{object}} \\ - P(\text{touch} \cdot \mathbf{A} \uparrow) \cdot P(\text{close-hand} \cdot \mathbf{A} \uparrow) \cdot P_{\text{object}} \end{aligned}$$

which is strictly greater than $P(\text{touch} \cdot \mathbf{A} \uparrow)$. So, *touch* is activated more often than usual, if *close-hand* was activated one time step ago. This can be intuitively explained as follows: Some of the trials of closing the robot hand will take place when there is no object between the fingers. If we had done our statistical analysis only in these instances, activation of *touch* and activation of *close-hand* would turn out to be independent events. However in those trials where there is an object between the fingers, closing the hand **causes** the pressure on the fingers and thus activation of *touch* and activation of *close-hand* will be statistically dependent. Those trials will affect the overall probabilities, and in the general setting, these two events will be statistically related. As described in the previous sub-section, a predictor from *close-hand* to *touch* will therefore be established. Let us call this new predictor p .

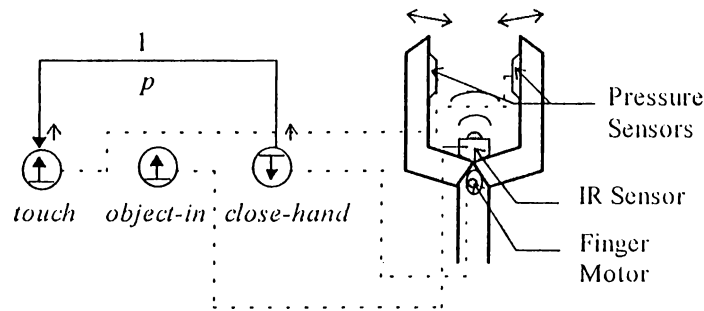


Figure 3.6. Predictor p is established between *touch* and *close-hand*

The predictions carried out by predictor p will sometimes be successful (the predicted event will happen.) and sometimes fail. We may consider the success and failure of this predictor as two new events. Let $p.S$ denote the event that p succeeds at a certain time step. Similarly, $p.F$ will denote the event that p fails at a certain time step. What we must do to enhance the prediction capability of our system is regarding these events as yet new events that need to be predicted.

Evidently p will be a rather unreliable predictor, because its predictions will be successful only if there is an object between the fingers (which is only rarely possible). However in the setting I have described above, there is a simple way to reliably predict the activation of *touch*. What we must do for that is to use sensor *object-in* (which is connected to the infrared sensor) together with the activation of *close-hand*. This fact can be exploited with a simple addition to the system: We must simply look for something that serves as some kind of context for the success of predictor p . Basically when that context is present, p 's chance of succeeding must be higher than usual. Activeness of *object-in* is just such a context. Therefore, what we will do is to establish a “meta-predictor” from *object-in* to p . When *object-in* is active, the meta-predictor will temporarily increase the reliability of p . Thus in those instances that *object-in* is active, p will function just like a reliable predictor. In other time instants, p will work like any other unreliable predictor.

3.7.1 Establishing Meta-Predictors

Now let us look at the details of this process. Firstly, we should elaborate on how to “find” meta-predictors. Let x and b be two interface elements and p be a predictor that predicts b 's activation from x 's activation. The reliability of p will be given by $P(b.A \uparrow | x.A \uparrow)$. Now, let a be another interface element. In order to

decide whether to establish a meta-predictor from a to p , we record the frequency corresponding to $P(b.A\uparrow | x.A\uparrow, a.A)$, which gives how reliably b 's activation can be predicted from x 's activation and a 's activeness. If this value is strictly greater than $P(b.A\uparrow | x.A\uparrow)$, which is p 's reliability, we establish a meta-predictor q and $P(b.A\uparrow | x.A\uparrow, a.A)$ becomes q 's reliability. This value shows how reliably b 's activation can be predicted from x 's activation and a 's activeness. And the check we perform makes sure that this is better than how reliably b 's activation can be predicted from x 's activation alone.

As stated before, we use frequencies to approximate probabilities. The way this approximation is done is not important for our theory, however, we must make sure that such frequencies can be obtained in a simple way. The expression we use for computing reliabilities of meta-predictors causes a problem here, since it involves events ($b.A\uparrow$ and $x.A\uparrow$) that are not directly recognizable by q (since q is not connected to either b or x). However we can express q 's reliability in another way:

$$\begin{aligned}
 \text{reliability}(q) &= P(b.A\uparrow | x.A\uparrow, a.A) \\
 &= \frac{P(x.A\uparrow | a.A) \cdot P(b.A\uparrow | x.A\uparrow, a.A)}{P(x.A\uparrow | a.A)} \quad (1) \\
 &= \frac{P(p.S | a.A)}{P(x.A\uparrow | a.A)} \quad (2) \\
 &= \frac{P(p.S | a.A)}{P(p.S | a.A) + P(p.F | a.A)} \quad (3)
 \end{aligned}$$

At step (1), we multiply the nominator and the denominator with the same expression. At step (2), we exploit the fact that $P(p.S)$ can be rewritten as the product of $P(x.A\uparrow)$ and $P(b.A\uparrow | x.A\uparrow)$. Finally, at step (3), we make use of the idea that whenever x is activated, p either succeeds or fails.

All the events involved in formula (3) are accessible by the meta-predictor. Hence the meta-predictor can adjust its reliability in its local setting.

As well as meta-predictors that predict their goal's success, we can have meta-predictors that predict their goal's failure. In order to discuss these we may make use of the notion of unreliability. Unreliability is the dual of reliability and is defined as $(1 - \text{reliability})$. A meta-predictor from an element a that predicts the failure of a predictor p is established if:

$$\frac{P(p.F | a.A)}{P(p.S | a.A) + P(p.F | a.A)}$$

is greater than the unreliability of p . In this case, this value becomes the meta-predictor's reliability.

I will call meta-predictors that predict their target's success, positive meta-predictors. Similarly, meta-predictors that predict their target's failure are called negative meta-predictors. When graphically representing positive meta-predictors, a plus sign (+) will be placed to the right of the endpoint of the arrow. Negative meta-predictors will be shown with a minus sign (-) at the same place.

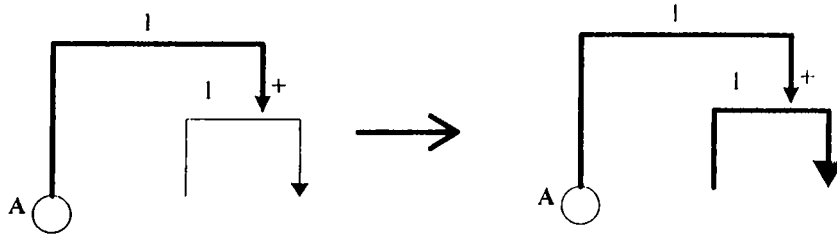
3.7.2 Prediction with Meta-Predictors

Whenever its source is active, a positive meta-predictor will temporarily set the reliability of its target to its own reliability. Conversely, a negative predictor in this case will set the reliability of its target to its own unreliability. That is, the reliability of the target predictor will become, for that internal time step, $(1 - \text{the reliability of the meta-predictor})$. An important observation here is that, a positive meta-predictor must be more reliable than its target, since this is the constraint for its establishment. Conversely, the reliability of a negative meta-predictor must be greater than the unreliability of its target. This makes sure that, if the meta-predictor predicts the failure or success of its target, it is guaranteed that the reliability of the target will change (unless another, more reliable predictor has already changed it).

It is also possible to establish meta-predictors whose targets are other meta-predictors. Such meta-predictors work and are established just like the first level meta-predictors.

I will give a new graphical rule for demonstrating how meta-predictors affect their targets in a simplified way.

Rule 2.a



Rule 2.b

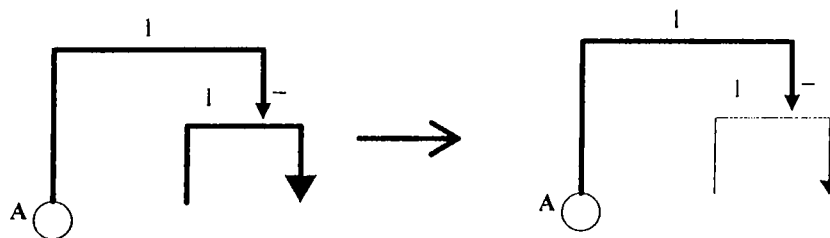


Figure 3.7. Rules for meta-predictors predicting the success or the failure of their targets.

3.7.3 Goal-Directed Behavior with Meta-Predictors

In line with our Guiding Principle 1, meta-predictors must also contribute to the goal-directed activity.

Reliable positive meta-predictors contribute to goal-directed activity by temporarily making their targets reliable when predicting their success. The example in Figure 3.8 demonstrates this idea:

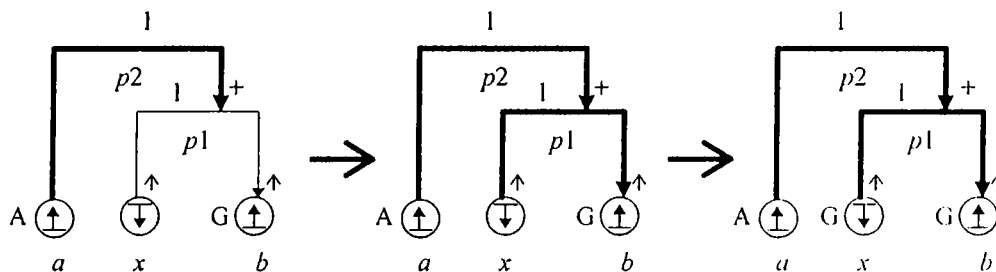


Figure 3.8. An example of goal-propagation with meta-predictors

When sensor *a* is active, meta-predictor *p2* makes predictor *p1* reliable. Since sensor *b* is a goal, predictor *p1*, which is now reliable, makes actuator *x* a goal.

Actuators, when they become goal, automatically become activated. Therefore actuator x will become active as a result. If $p2$ had not predicted the success of $p1$, x would not get activated.

Similarly, reliable negative meta-predictors help with goal-directed activity by temporarily making their targets unreliable. Figure 3.7 gives an example for this idea:

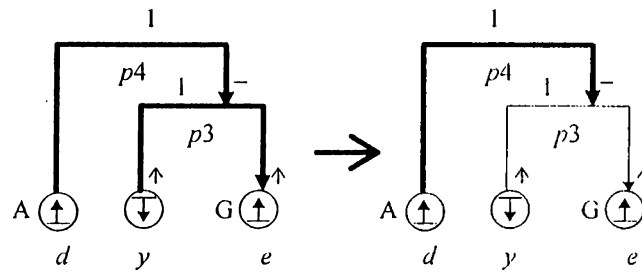


Figure 3.9. Inhibition of goal-propagation with negative meta-predictors

When sensor d is active, meta-predictor $p4$ makes predictor $p3$ unreliable. Sensor e is a goal, however, since $p3$ becomes temporarily unreliable, it can not carry goalness to actuator y . If $p4$ had not predicted the failure of $p3$, y would get activated.

Meta-predictors can contribute to goal directed activity in even more powerful ways. However, to do that, the functioning of basic predictors must be extended. This extension to the functioning of predictors will be different according to whether the predictor's source is a sensor or an actuator. I will give two new rules for the two different cases and present the rationale behind them.

Rule 3

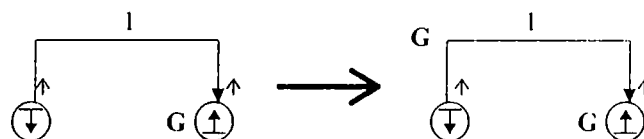
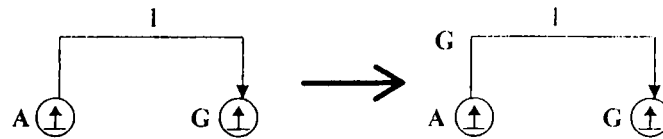


Figure 3.10. Rule 3.

The predictor in Figure 3.8 encodes the knowledge that the target will be activated when the source actuator becomes active. Rule 3 states that if this predictor is unreliable, the predictor itself becomes a goal. A predictor's becoming a goal means that other elements in the system will be activated to be able to predict its success. Once the predictor is predicted and becomes, as a result, reliable, it will be able to activate the actuator.

Rule 4.a



Rule 4.b

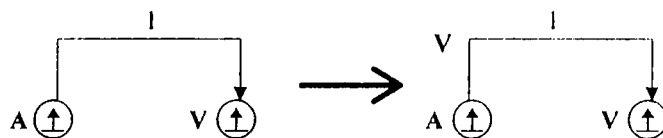
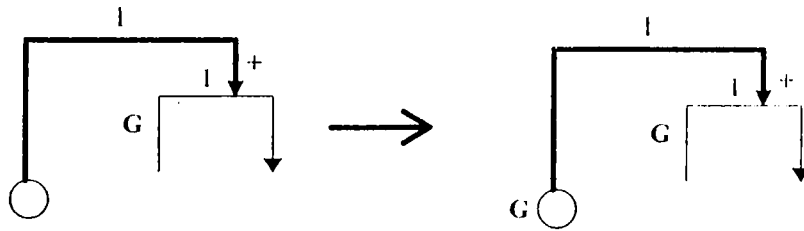


Figure 3.11. Rules 4.a and 4.b

Rules 4.a and 4.b in Figure 3.9 define how unreliable predictors can propagate goalness and avoidedness if their sources are sensors. Notice that the source and target events of the predictors are activeness events and not activations. Such predictors become goal or avoided only if their sources are active. When their sources are active, such predictors encode the information that their targets will become active if they can become reliable. Therefore, if their target is a goal, they make themselves a goal so that other predictors will predict their success. If their target is avoided, they make themselves avoided so that other predictors will not accidentally predict their success.

Once we have defined these rules, we may give goalness and avoidedness propagation rules for the meta-predictors.

Rule 5.a



Rule 6.a

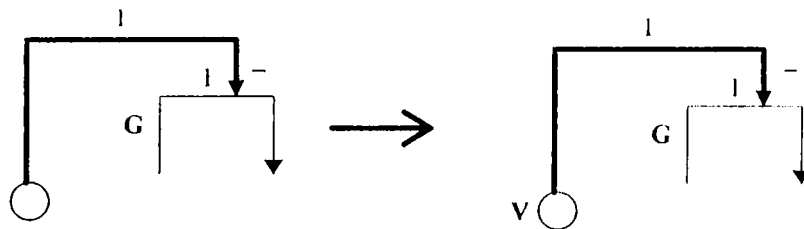


Figure 3.12. Rules 5.a and 6.a.

The rules in Figure 3.10 describe how reliable meta-predictors propagate goalness and avoidedness. Rule 5.a looks very much like Rule 1.a. The meta-predictor makes its source a goal so as to make its target predictor reliable. If its source is an actuator, it will automatically become activated. If it is a sensor, goalness will be further propagated. Rule 6.a uses a similar idea. However, it makes its source avoided if its target is a goal. Rules 5.b and 6.b can be obtained by replacing goalness and avoidedness with each other.

Rule 7

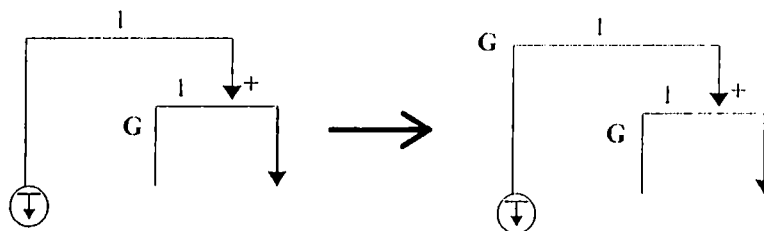
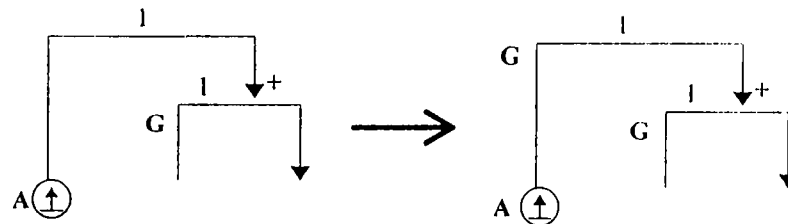


Figure 3.13. Rule 7.

Rule 7 describes how unreliable predictors whose sources are actuators propagate goalness. This rule is based on the same idea as Rule 3. The meta-

predictor makes itself a goal so as to become reliable and activate the actuator at its source.

Rule 8.a



Rule 9.a

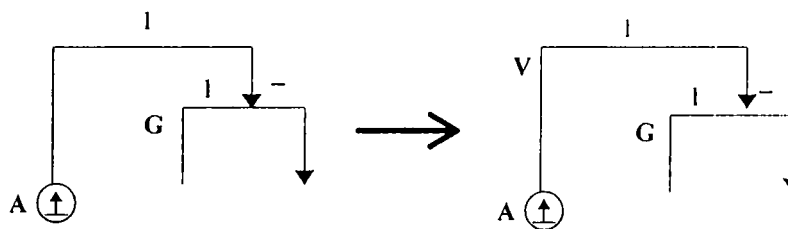


Figure 3.14. Rules 8.a and 9.a

Rules 8.a and 9.a apply the idea in Rules 4.a and Rule 4.b to meta-predictors. These rules are valid for unreliable predictors that have sensors as their sources. The meta-predictor in rule 8.a encodes the information that its targets will be reliable if the meta-predictor itself becomes reliable. Therefore, it makes itself a goal. Similarly, the meta-predictor in Rule 9.a encodes that, its target will become unreliable, if the meta-predictor itself becomes reliable. Therefore, it makes itself avoided. Rules 8.b and 9.b are based on similar ideas and can be derived by replacing goalness and avoidedness with each other in Rules 8.a and 9.a.

These rules are generated according to intuitive ideas. Like most other rules and ideas in this thesis, they should not be taken as if they are precisely specifying an architecture. The main objective in stating these rules is communicating how a prediction mechanism can function. Some of these rules can be discarded or modified, and many other rules can be found useful while engineering a working prediction mechanism.

3.7.4 Timing Issues

While describing the setting for the system, I stated that internal and external time steps are distinguished from each other and an external time step corresponds to multiple internal time steps. A process carried out by a predictor (a prediction or a goal propagation) takes one internal time step. Therefore, multiple internal time steps are necessary for the hierarchical system of predictors to finish their processing. However, this is not the only reason for having multiple internal time steps within a single external time step. Another reason is that, it is more fruitful to view a system of predictors as a dynamic system that stabilizes after a certain time, rather than seeing it as a system implementing a feed-forward, monotonic process. We can take this dynamic character of the prediction mechanism into account in developing certain details. One of the results of such an approach is that it would be wrong, in this case, to think that the state of any element in the unstable phase of the system has any significance.

An actuator may become a goal during the unstable phase of the prediction mechanism. The characteristics of the actuator must be such that, it must not immediately become activated and initiate a motor action the moment it becomes a goal, but must wait until the current external time step. Only when the external time step ends, must it become active and start the action. We could have created the same effect if we had not distinguished between external and internal time, by specifying the behavior of the actuator so that it would become activated only if it remains as a goal for a long enough time interval.

Having defined the behavior of the actuators in this way, we can add a new principle: An actuator which is a goal will behave inside the prediction mechanism, as if it was active. I will call this rule, the “**simulation principle**”, since this makes the prediction mechanism carry out prediction as if the motor action was really initiated.

3.7.5 Examples with Meta-Predictors

In this sub-section we will look at a number of problematic example cases to demonstrate how predictors can carry out goal-directed activity.

The Need For Synchronous Activation

For accomplishing certain tasks, it may be necessary to activate multiple actuators synchronously. Figure 3.13 shows a system of predictors that can carry out such a task.

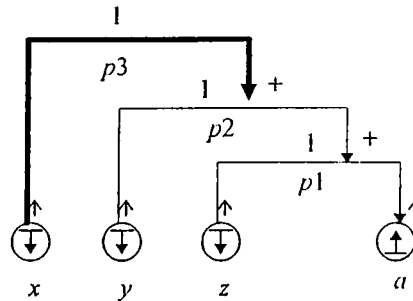


Figure 3.15. Predictors for recognizing synchronous activation.

The predictors in Figure 3.15 basically encode that sensor a will be activated one time step after the activation of z , if x and y are also active at that previous time step. When node a becomes a goal, predictor $p1$ will become a goal (by Rule 3) after which, $p2$ will also become a goal (by Rule 7). Since $p3$ is a reliable predictor, it will carry $p2$'s goalness to x (by Rule 5.a). Since x is an actuator and is a goal, the simulation principle will be applied and x will act as if it had been activated in this turn. Therefore $p3$ will predict $p2$ and make it reliable (by Rule 2.a). Since $p2$ is now reliable, it will propagate the goalness of $p1$ to y (by Rule 5.a). Once again by the simulation principle, y will act as if it was active. $p2$ will thus make $p1$ reliable (by Rule 2.a). Predictor $p1$ is now reliable and will make z a goal. At last x , y , and z have become goals and they will synchronously carry out their actions in the next external time step.

Fine Control Of Actions

Some tasks require precise sensory-motor control which can not be achieved by only activating certain activators. In order to achieve a fine level of sensory-motor control, certain other actuators must be inhibited. Below, we will look at such a problem.

Assume we have a sensor a and two actuators x and y . Let us also assume that activation of either x and y leads to the activation of a . However, when x and y are both activated at the same time, a does not become active. Such a scheme, will lead to the following system of predictors:

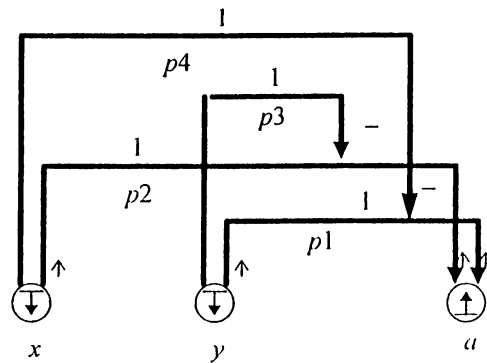


Figure 3.16. An example for the fine control of an action

Predictors $p1$ and $p2$ predict the activation of a from the activation of x and y respectively. Predictor $p4$ captures the knowledge that when x is active $p1$ will fail ($p1$ can not succeed if x was active in the previous time step, because this means that x and y were both active at the same time, which means that a would not become activated). Similarly, $p3$ captures the knowledge that $p2$ will fail if y was active one time step ago.

All these predictors will be established according to the statistical methods we have described. Moreover, the system will successfully carry out goal directed activity. When a is a goal, $p2$ and $p1$ will make x and y goals. By the simulation principle, $p3$ and $p4$ will then make $p1$ and $p2$ unreliable. Therefore x and y will not be goals in the next internal time step. However, this means that $p3$ and $p4$ will not predict $p2$ and $p1$'s failure anymore. Therefore, the same sequence will repeat again. Thus, x and y will oscillate between being goals and not goals, whereas $p2$ and $p1$ will oscillate between being reliable and unreliable. However, reliability and goalness are, in fact, continuous attributes, and if, for example, $p1$ and $p3$ are slightly more reliable than $p2$ and $p4$, y will get the upper hand in this competition. Even if all predictors have the same reliability, the system will be in a delicate balance that will be disturbed by even a small effect from outside. Thus, eventually, the balance will be broken and the system will stabilize, activating one of the actuators.

3.8 Generalizers

So far we have focused our attention on Guiding Principles 1 and 2, that is, on successfully predicting sensory states and guiding goal-directed activity. However, the mechanisms we have described are also in line with Guiding Principle 3. The important point here is to observe that a predictor's states almost al-

ways capture a higher level piece of information than the states of its source and target. In this section we will focus on Guiding Principle 3.

One important property of intelligent behavior is treating similar things in similar ways. It would be absurd to claim that high level cognitive skills, such as analogy making, categorizing, etc., that are responsible for this kind of behavior can be explained at the level of the prediction mechanism. However, it would be plausible to think that the prediction mechanism, in its own simple way of doing things, exploits this idea. It tries to find similar things, similar schemes etc., to ease the task of prediction. In other words, it “generalizes”.

How can a prediction mechanism generalize, how can it find out that two things are in some way, “the same”? There can be multiple answers to these questions, but a simple one that comes to mind is the following: “Same” things usually have the same consequences. We will now extend our theory to exploit this idea.

A simple way to implement this idea is collecting predictors that have the same target, in one common “pool”. In order to do this, it is sufficient to insert a new node before an element, make all the predictors that predict the element in the same way predict that new node, and make that new node, in turn, predict the element. This idea is demonstrated in Figure 3.17:

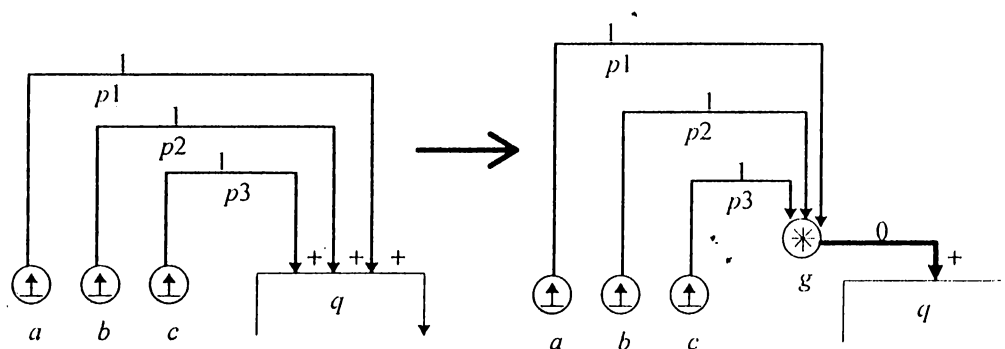


Figure 3.17. Establishment of a generalizer.

I will call such temporary nodes generalizers and graphically represent them with circles that have an asterisk sign inside. I suggest inserting generalizers only before predictors, since it is difficult to see how inserting generalizers before sensors would be beneficial. A generalizer is connected to the predictor after it, (Let us call this predictor the indirect target of the generalizer) with a special 0-delay predictor, called the primary arc of the generalizer. The generalizer becomes activated when it is predicted by a reliable predictor. Once it is active, it in

turn (indirectly) predicts either the success, or the failure of its indirect target. If the primary arc of the generalizer now succeeds, the generalizer signals a success so that the predictor(s) that activated it also succeed. Similarly, if the primary arc fails, the generalizer signals a failure, and the predictors also fail. Generalizers also propagate the goalness or avoidedness of their indirect targets.

3.8.1 Emergence of a Representation of Obstacles

Let us, once again, continue our discussion with an example. Our agent this time will be a mobile robot that moves on wheels through a complex environment. The agent is equipped with different kinds of sensors like sonars, infra-red sensors, touch sensors etc., as well as sensors that measure its speed, acceleration, deceleration, etc. As the agent navigates in the environment, a predictor is established from its actuator that makes it move forward, to a sensor that senses the forward movement. Therefore, this predictor simply captures the knowledge that, when the agent tries to move forward (by activating the appropriate actuator) it will do so. This predictor will usually succeed, but sometimes it will fail, probably because an obstacle is blocking the way. In time, meta-predictors will be established that predict when the predictor fails. There will, of course be a number of such meta-predictors; each one using different sensory inputs. Some meta-predictors will have, as their sources, sonar outputs, others infra-red sensor outputs, still others the outputs of touch sensors on artificial tentacles. An important point is that many different things can be obstacles: a wall, a column, a glass door, a stair, the legs of a chair, to name a few. All of these obstacles will not always be recognized by a single sensor. However, we may assume that an obstacle will be sensed by at least one of these sensors. If we do not establish generalizers, every meta-predictor will work on its own to predict the failure of the predictor. However, if we do establish them, all these meta-predictors will first predict a generalizer. The activeness of the generalizer will signal the existence of a (generic) obstacle in front of the agent: a piece of information that could not be captured by any single meta-predictor. So, generalizers can produce signals that are much closer to our high-level notions than the signals generated by individual predictors.

There is also another very important function of generalizers. They lead to an economy in the number of predictors used in the system. A generalizer, once established, can be used in the prediction of multiple elements. For instance, in our example, the generalizer that recognizes the existence of an obstacle in front of

the agent can be used to predict that if the agent is moving speedily, the touch sensors on its front will sense a violent crash.

3.8.2 Low-level Organization of a Visual System

Let us now try to apply the organization rules covered so far to the visual system of an agent. For simplicity, I will assume that the visual input of the agent consists of a rectangular grid of light sensors. Every sensor becomes active if the amount of light it senses is over a certain threshold and is inactive otherwise. So if the agent is looking at a black shape over a white background, the grid of sensors would have the following pattern of activations.

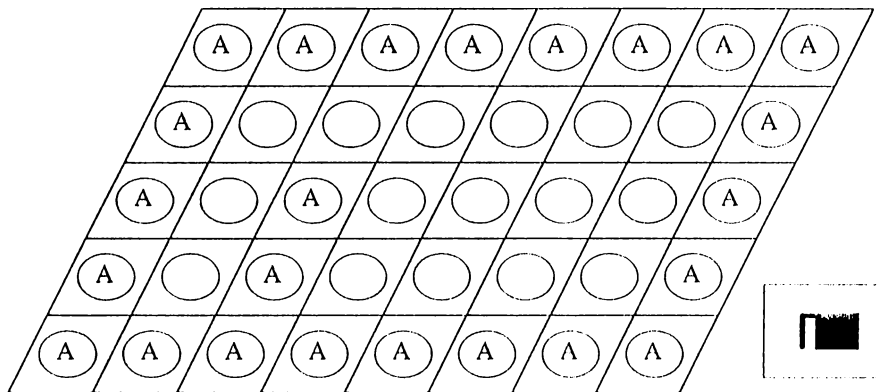


Figure 3.18. A pattern of activation on the grid of sensors

If the agent is situated in a typical environment, then the sensors in the grid will be statistically related to neighboring sensors. (This is so because objects in the world usually have uniform colors or uniformly colored sub-regions and such objects usually do not stimulate a single visual sensor, but a group of them.) The activations of two such sensors will be statistically dependent within the same time step. Therefore we have to establish 0-delay predictors in order to capture such relations. So, 0-delay predictors will be established between sensors that are close to each other. I will assume that predictors between two neighboring sensors will be reliable, and predictors between sensors that are close but not neighboring will be unreliable. (Remember that reliable predictors are those whose reliability is greater than a certain threshold. For the sake of the argument, I will assume that the threshold is appropriately fixed for these predictors.) Therefore, a network of predictors will be established as in Figure 3.19.

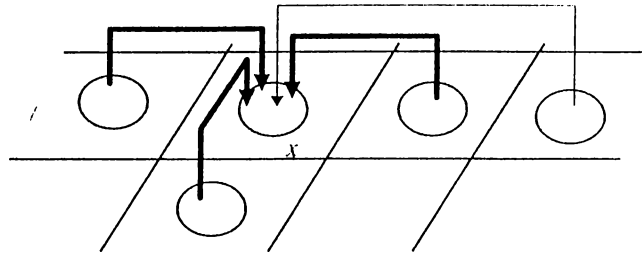


Figure 3.19. Predictors established between proximate sensors.

These predictors will fail whenever their source is active but their signal is not. Therefore, failures of these predictors in some sense recognize points of contrast. Once again, we obtain some higher level information by using predictors.

As I have discussed in this section, we can collect the predictors shown in Figure 3.19 (and other predictors that predict x 's activeness) together by introducing generalizers. In this scheme, there will be one generalizer for every sensor and predictors will be connected to the generalizer as shown in Figure 3.20.

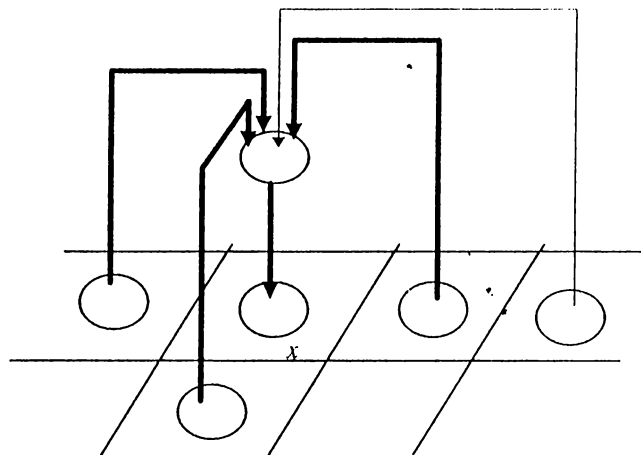


Figure 3.20. Predictors collected through a generalizer.

Since there will be one generalizer for every node, we may think of generalizers as constituting a second layer over the layer of sensors. Now, what information does this new layer encode? According to our discussion, the generalizers will become active when they are predicted by a reliable predictor, and they will signal a failure (let us simply say, they fail) when their primary arc fails. There-

fore, a generalizer will fail whenever one of the neighboring sensors is active but its indirect target is not. The failure of a generalizer indicates that the sensor it generalizes is a point of contrast. Therefore, the second layer of nodes is a map of contrasts, something like the primal sketch used by Marr as an intermediate representation of the visual input.

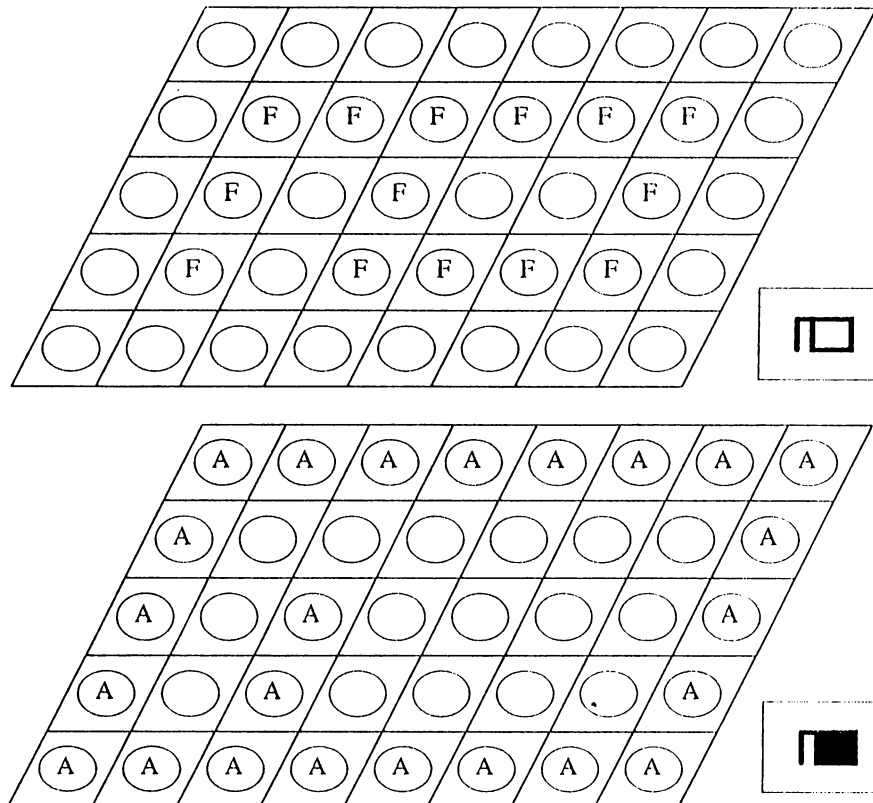


Figure 3.21. The original activation pattern and the primal image constructed by generalizers.

If we observe the statistical relations between the nodes in the second layer, there will be again interesting dependencies between the generalizers. The crucial observation at this step is that points of contrast would not in general come in random patterns but would be organized as continuous edges. This regularity would depict itself on the statistical relations between the generalizers, and the failure of neighboring generalizers would turn out to be statistically dependent. Thus 0-delay predictors would be established between neighboring generalizers whose success would indicate **short segments of contours**.

I have not described the entire story here. Hierarchical build-up would continue from this point. Short segments of contours should be also interdependent and this would be captured by the prediction mechanism. Furthermore, I have

only discussed the establishment process without paying attention to other aspects of predictors like inhibitions or enablings. I also did not discuss how higher cognitive systems could interact with this process. Considering these aspects, the visual system described here would develop in complicated ways that are not clear to me. My purpose was to demonstrate how visual primitives (like segments of contours) could emerge out of the organization rules of the prediction mechanism.

3.9 State-Holders

So far I have described a number of elements of a prediction mechanism. These elements share a common property: They only use the information that is readily available on the sensors and actuators for carrying out predictions. Many events in the environment can be predicted that way. But the prediction mechanism can be made a lot more powerful if we add a small memory component to it, enabling the system to use this memory component together with the information on the sensors and actuators.

Adding such a memory component would enable the prediction mechanism to make better predictions, because things in the world tend to keep their states. Even if there is no readily available sensory information about a certain thing, the memory component could make predictions about that thing by using the idea that, the thing is still preserving its state since the last time sensory information was last obtained about it.

I call such memory components “state holders”, because once they become activated, they will remain to be active for a short time, even if the activation signal dies out. The need for state holders will become apparent if we show how non-state holding elements fail to accomplish certain tasks.

Basically, an element may have two types of activation patterns: First, the activation pattern may consist of short spikes; once the element becomes active, it becomes inactive again in the next time step. Secondly, the activation pattern may consist of relatively long plateaus, once the element becomes activated, it remains to be active for some time

Let us call the first type, spike maker elements, and the second type plateau maker elements. Meta-predictors that have these two different types of sources have different characteristics. If the source is a plateau-maker element, the meta-predictor encodes the fact that the target predictor will succeed (or fail) as long as the source remains active. Thus, in a way, the plateau maker element acts as a

context. If the source of the meta-predictor is spike maker, the meta-predictor encodes the fact that the target predicted will succeed (or fail) at the next time step after the appearance of the source signal. This means that the source of the meta-predictor and the source of the target predictor must be **simultaneously** activated. Thus the spike-maker acts in a way as a trigger. This means that there is no way to use a spike-maker element as a context. This is rather problematic because spike-maker elements usually signal a change in the environment. If the changed “thing” remains to be sensed by the agent, there is no problem, however if it is not, then we lose an important piece of information.

Not only most activators and some sensors, but also predictors and meta-predictors are spike-making elements (since success and failure events can be regarded as spikes). The activation, success and failure events generated by these elements signal states in the environment, and such states can be used as a context. State holders will help us to capture this kind of knowledge.

Figure 3.22 shows a typical state-holder. The state-holder is connected to a predictor (we call this the indirect target of the state-holder) with a 1-delay predictor (called the primary arc of the state-holder). The primary arc does not have to be fully reliable unlike the primary arc of a generalizer. The state-holder can be predicted or “depredicted”. Once it is predicted by a positive predictor, it becomes active and remains to be active until either it is depredicted by a negative predictor or a certain time limit is reached. The state-holder can also propagate goalness and avoidedness.

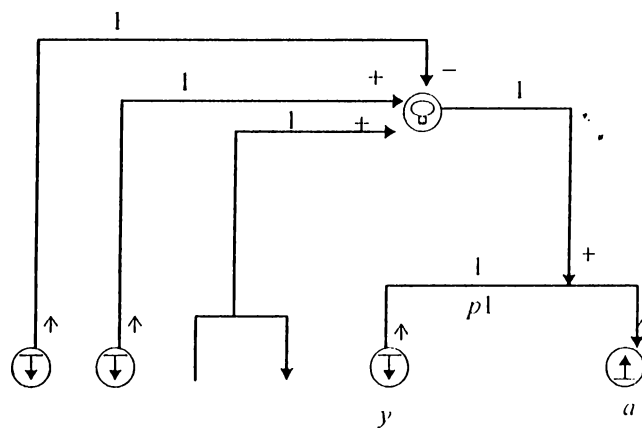


Figure 3.22. A state holder.

We graphically represent a state-holder with a circle that has a light bulb inside. Predictors that have a state-holder as their target may have predictors as

their sources. These are triggered by either the success or failure of their source predictors.

3.9.1 Establishing State-holders

A state-holder represents a general, possibly hidden context for the success or the failure of a predictor. Therefore, it is possible to establish state-holders for every predictor. However it could be better to use heuristics in selecting which predictors to establish state-holders for. State-holders can also be used in conjunction with generalizers as shown in Figure 3.23.

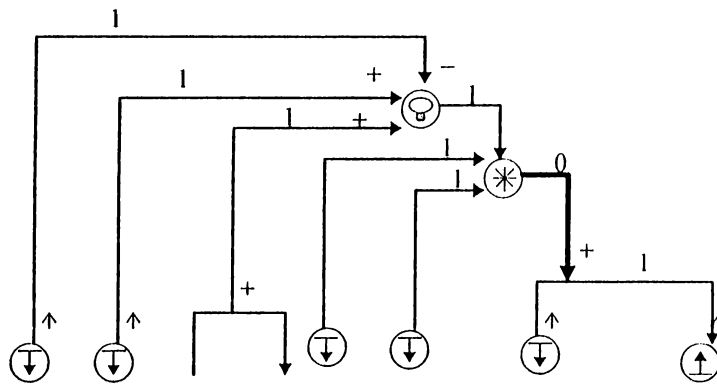


Figure 3.23. The usage of a state-holder together with a generalizer

The generalizer in the figure uses the state of the state-holder together with other sensors and actuators to carry out predictions. Thus, the state-holder can be generalized as well. This scheme is not problematic, however, I will ignore generalizers in this discussion for simplicity.

The real problem with using state-holders is finding those events that could predict or depredict it. For this, we may think that, instead of making a statistical observation of the elements at two time steps, we observe statistical relations within a time range. If, say, the success of a predictor x turns out to be statistically dependent on the activity of an element y , not only for a single time step, but for a number of time steps afterwards, we can establish a predictor from y to the state-holder attached to x .

Also activities of high-level cognitive systems may be guiding the state-holder establishment process. Such an idea makes sense. However, how such a guidance can be carried out is unclear.

3.9.2 Sequences of Actions

State-holders enable us to encode the knowledge about a sequence generating a result, within a prediction mechanism. Assume we have three actuators, x , y , and z , and a sensor a . When x , y , and z are activated strictly in this sequence (not necessarily in consecutive time steps), sensor a becomes activated. Such knowledge can't be faithfully encoded without state-holders, if intermediate states are not recognized by sensors. However, the system in Figure 3.24 can encode it.

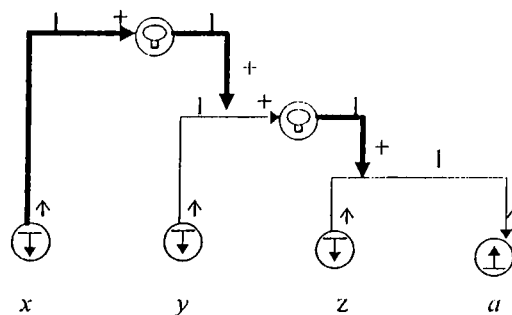


Figure 3.24. Predictors that recognize a temporal sequence

Predictions that deactivate the state-holders are not shown in the figure, but they are assumed to exist. The system will not only successfully predict the activation of a at the right time, but also it will initiate the right sequence of motor actions if a becomes a goal. The delays on the arms of the state-holders will provide that the actuators are not activated simultaneously but in a sequence.

3.9.3 Re-trying an Interaction

Consider the example about the robot hand discussed in Section 3.7. Let us assume that we have a robot hand exactly like that one, but without an infra-red sensor to sense objects between the fingers. In this case, the predictor from *touch* to *close-hand* can not be made reliable, since there is no sensor that signals the existence of an object. However, using state-holders may help with this difficulty. When $p1$ succeeds (it touches an object) it is probable that it will succeed again in the near future (the object will still be there). Conversely if it fails it will probably fail again. The following system captures this knowledge.

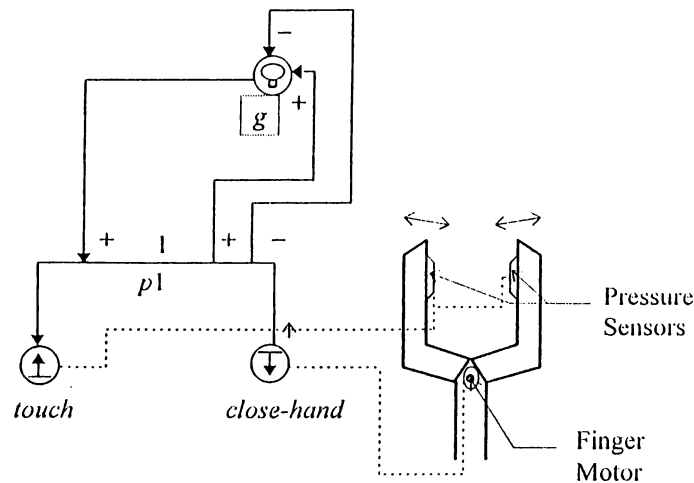


Figure 3.25. A scheme for the robot hand to remember its state

If $p1$ succeeds it will activate its state-holder g . Conversely, if it fails, it will deactivate g . If g has become activated, it will continue to predict the success of $p1$ for some time. Overall, the system encodes the knowledge that $p1$ is likely to succeed if it succeeded a short time ago. There may of course be other predictors that predict g by looking at other inputs.

3.10 Remarks

This finishes the presentation of the prediction mechanism. I should add a disclaimer here. This sketch of a model should not be taken as a finished proposal that will work once it is implemented. A real prediction mechanism can be developed only through detailed experiments and continuous refinements. However, I believe this chapter suffices to demonstrate how the problem can be attacked and serves to show the ideas and principles at work.

Chapter 4

Discussion

My claim in the introduction of this thesis was that a prediction mechanism serves as the low-level component of the cognitive system. I presented a number of arguments to support this claim. I argued that a prediction mechanism would help with goal-directed activity, distinguish between interesting and non-interesting events and generate and use high level notions. Then, in the third chapter, I sketched a model of a prediction mechanism. I suggested mechanisms and organizational rules in order to support these three functions. In this chapter, I will look at the idea of a prediction mechanism in relation to the general problems of cognition. In the first section, I will try to reply to some possible questions and objections regarding the prediction mechanism. Next, I will review the functions it fulfills within the cognitive system. In the third section, I will discuss some of its additional capabilities. Finally, the last section of this chapter is devoted to a discussion of higher cognitive systems and their relation to the prediction mechanism.

4.1 Questions Regarding the Prediction mechanism

In this section, I would like to point out some possible questions regarding the prediction mechanism and suggest some answers.

Q1: The operations of the prediction mechanism seem to be associationist. But, we know that simple associationism cannot account for the complexity and productivity of human behavior. How can the prediction mechanism overcome the limitations of associationism?

Although it may seem so, the theory presented here is not an associationist one. For example, the establishment of a predictor and the establishment of a Hebbian cell assembly look very similar in that, they are both based on some kind of simultaneous activation of elements [14]. However, unlike associationist mechanisms a predictor does not merely “associate” its source and target (for example, it does not activate its target when its source is active). Prediction is a much richer organization principle than association and mechanisms that carry out predictions, therefore, can serve a broader functionality than associationist mechanisms can. For example, the success and failure of a prediction creates a piece of information that is of a higher level than the original signals, whereas association does not create any information as such. It, at most, extends the coverage of already known rules to new objects.

Q2: It seems, the prediction mechanism has to catch delicate statistical relations between activation patterns of sensors and actuators in order to work properly. However, most “interface elements” will be only loosely related with each other. For example, the activation of a receptor in the retina and the activation of a motor that moves the arm are not independent events since there are certain delicate conditions in which one of these will trigger the other. However such relations are statistically so insignificant that it is hard to imagine how they can be caught accurately and systematically by a simple mechanism.

I would like to answer this in a series of steps; (1) The prediction mechanism must be a practical system. Its duty is not capturing all statistical relations between elements, but aiding the adaptation and survival of the agent by distinguishing interesting events, helping with goal-directed activity and generating high-level information. Therefore missing a certain relation will not matter as long as this does not hinder these functions significantly. Even if it does so, we would have the same problem without the prediction mechanism. So the question is: are there any dependencies between elements that are easy to catch, and would these be sufficient to carry out the three functions I have mentioned. (2) There are indeed dependencies between certain elements that are statistically significant. Consider this time two neighboring receptors in the retina. These two will have similar levels of activation if they are sensing the same surface and different levels of activation if one is sensing an edge or a spot. Obviously, at any given time, most of our receptors are sensing surfaces, while relatively few are sensing edges or spots. Therefore, such receptors will have **significant** statistical relations. Similarly, there will be other significant statistical relations regarding other sensory systems. (3) The mechanics of the prediction mechanism dictates that the success or failure of a predictor is a less frequent event than its source signal.

Therefore, as we go up in the hierarchy of predictors, the signals become less frequent, carry higher-level information and become more specific. Consequently, the relation between signals at higher levels can become statistically significant, although relations between their sources and targets are not.

Q3: It seems that we have lots of statistically related signals going around. If we have, say, 100 elements that are all related to each other, then, establishing predictors between each (ordered) pair will result in 9900 predictors. If the successes and failures of these predictors are also all related, we will get 9900×9899 predictors in the second level. Since, we do not know in advance what kind of statistical relations the environment will generate on the agent's senses, it seems possible that the number of predictors will explode combinatorially.

This, I must admit, is a serious problem. It is one of the reasons that makes the development of a “practical” prediction mechanism a tough engineering problem. However, I do not see any reason for not being able to control the explosion with new types of elements and heuristic measures.

Q4: How can the prediction mechanism ever be implemented in the hardware of the brain? The system requires that the elements carry out complicated tasks such as recording frequencies. Moreover, there is the problem of establishing new predictors between existing ones. How can the neurons, being relatively simple structures ever carry out such complex tasks?

First of all, the prediction mechanism does not have to map directly to organic mechanisms. The model I sketched is just one possibility for carrying out the functionalities I mentioned. There may be many other ways the same thing can be done, provided that these also work on a hierarchy of predictions. Second, we have not yet fully understood the workings of the brain and we should not make any a priori assumptions about its complexity. Third, frequency recording can be approximated by simple means; when a positive instance comes, we can just increase the frequency a little bit if it is already high, and increase it a lot if it is low. Similarly, when a negative instance comes, we may decrease the frequency a little bit if it already low, and decrease it a lot, if it is high. This is very easy to carry out. It will give us only a crude measure of frequency, but this is not problematic since the frequency itself is only an approximation of probability. Lastly, new connections can be established in many ways, even randomly, provided that they can be broken with ease. Finding statistically related elements can be, after all, just a search problem. And neurons do get connected and disconnected with each other.

4.2 Functions of the Prediction Mechanism

The prediction mechanism plays an important role in cognitive activity. I have listed three basic functions that are carried out by the prediction mechanism. To repeat, these are supporting goal-directed activity, finding interesting events by distinguishing between predicted and unpredicted states, and finally, generation of high level notions. Below, we will look at each of these functions.

4.2.1 Supporting Goal-Directed Activity

The prediction mechanism carries out goal directed activity by propagation of goalness through predictors. This type of parallel, simple, and quick goal-directed activity is quite different from the serial, sophisticated and slow goal directed activity that humans carry out consciously. The role of the prediction mechanism in generating actions must be, therefore, limited to carrying out automatic behaviors. When a human first learns a sensory-motor skill (such as walking, swimming, juggling, etc.) he consciously plans and initiates every little action at the beginning. However, with experience, he masters the skill and starts to use it effortlessly. I believe that it is the prediction mechanism that is responsible for this automation. While the skill is being experienced, the prediction mechanism connects the relevant elements to predict what happens during the application of the skill. While doing this, it starts to become capable of carrying out the skill. So, the responsibility of carrying out (at least the monotonous parts of the skill) are slowly transferred to the prediction mechanism. After a time the prediction mechanism starts to display the skill on its own.

While the theory of the prediction mechanism is developed for carrying out low level, sensory-motor skills only, it seems to be possible that the prediction mechanism also automates certain high-level skills. After all, the prediction mechanism can regard any internal element as a sensor or an actuator. It does not really matter if that element is “below” the prediction mechanism (connected to the world) or “above” it (connected to higher level systems). If a process that is carried out by high level systems has repetitive parts, these parts may, in principle, be transferred to the prediction mechanism as well.

4.2.2 Finding Interesting Events

We have claimed that one of the functions of a prediction mechanism is finding out interesting events by distinguishing between predicted and unpredicted ones.

If the cognitive agent is sitting in a room with its eyes fixed on a spot, the prediction mechanism will predict that the image seen will not change. If the prediction fails, that is, if part of the image changes, this means something or somebody in the room has moved: a piece of information that is in general more interesting than the image of a still room. If the prediction mechanism is powerful enough, it may, this time, predict subsequent movements of the moving thing. If these predictions fail, this means that the thing that is moving has changed the course of its movements, in other words, it has “acted”. Again, something more interesting is revealed by carrying out predictions.

In the prediction mechanism we have described, a meta-predictor always has to be more reliable than its target. Therefore it fails less often and makes stronger predictions. This further means that the failure of a higher level predictor is more “unpredicted” and hence, more “interesting” than the failure of a lower level one. Thus the hierarchical build-up of the prediction mechanism automatically contributes to the objective of finding more interesting knowledge.

4.2.3 Generation of High Level Notions

A powerful prediction mechanism must employ notions such as nearness, farness, movement, etc., so that it can make successful predictions. Therefore, some of the concepts that are meaningful to us may be grounded in elements at this level. The argument that this is a benefit of the prediction mechanism may seem to be circular at first look, because we first posit the existence of such notions without stating how they can be created, and then call that an advantage. The important point here is that the prediction mechanism alone can be responsible for the creation of such high level notions, by trying to predict better using limited resources. If we somehow engineer the system so that it finds the most economic way of making successful predictions, “high-levelness” of internal elements comes as a bonus.

Elements like state-holders and generalizers allow the system to make better predictions, at the same time decreasing the number of predictors used. It is in general more efficient to use these elements as intermediate states in prediction instead of using predictors directly.

While presenting the model of the prediction mechanism, I presented methods of establishing predictors but did not discuss methods of discarding them. Establishing every possible predictor may not be the best way to organize a prediction mechanism. An alternative perspective is seeing the prediction mechanism as an arena where different elements compete with each other to make more successful

predictions, those failing in the competition being eliminated. Such a system will be difficult to design, however, this effort must be undertaken if we wish to implement a practical model.

4.3 Capabilities of the Prediction Mechanism

In this section, I wish to discuss certain capabilities of the prediction mechanism described in the previous chapter.

4.3.1 Pattern Recognition

Perhaps the most important difference between the approach presented in this thesis and more classical approaches, is the treatment of time. One of the most important ideas underlying this study is that cognition is something that takes place over time. This idea is used in the design of even the simplest components of our model. However, the idea can also be used in a broader setting. In this subsection we will look at how this idea can be applied to pattern recognition.

In general, pattern recognition is studied, not in the context of an interaction between the agent and the environment, but as an isolated task on its own. However, pattern recognition, if taken as an interactive process, presents a quite different picture than the models generated in such studies.

For example, in a discussion, Dr. David Davenport had suggested a rather interesting scheme for visual recognition that explained recognition in terms of a series of simple predictions. Every prediction was something like: "If I look at that point now, I will see a vertical line (or a dot, a curve, etc.)." In the beginning of the recognition session, the predictions usually failed. However, in time, they became more and more successful. Finally, there came a time where the predictions always succeeded. This meant that the object at hand was recognized.

Applying the idea of prediction to recognition, therefore, seriously changes the nature of the problem. The agent can look at different points, move its head, touch an object, etc., in order to obtain more information. Thus, recognition can be studied as some type of interaction, rather than as an isolated problem.

4.3.2 Composite Actions

One problem with the model described in the previous chapter is that goals are represented by increasing the goalness of certain elements. How can such a

scheme represent a goal like “grab the bottle” that involves multiple constituents (such as **grabbing** and **bottle**)? It can be suggested that there are certain elements in the system that represent such composite actions. However, that would lead to a combinatorial explosion in the number of elements needed, since we would have to establish elements that represented arbitrary combinations of such constituents. A better solution can be found by letting actions take their arguments from the context.

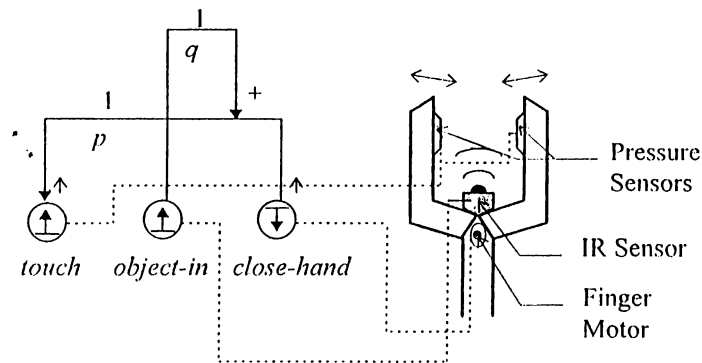


Figure 4.1. Sensory motor control of the robot hand

In order to present this idea, let us remember the robot hand example that we used in the previous chapter. In this example, when the sensor *grab* is made a goal while there is no object in-between the fingers, the predictor *p* will become a goal itself (since its source is an actuator). In turn, the meta-predictor *q* will make the sensor *object-in* a goal. Thus, we will now have the goal of having something in-between the fingers. We may assume that, the goalness of *object-in* will be propagated to other parts of the system, probably resulting in the behavior of moving the hand towards the nearest object so that it now lies between the fingers. Once that is achieved, *q* will make *p* reliable, and *p* in turn will activate *close-hand* so that the object is grabbed. In short, when the sensor *grab* is made a goal, the agent will grab the nearest object. This means that, an activity like “grabbing the bottle” can be implemented by bringing the hand close to the bottle and then carrying out the generic grabbing action. Of course this is not a complete solution to the problem because it leads to the problem of “bringing the hand close to the bottle” which is just another composite action. But that action too can take its argument from the context. For example, the hand movement action may move the hand to the point where the eyes are fixed. Such reductions of composite actions may eventually lead to simple actions that take their objects from the context.

4.4 Towards High Level Cognitive Skills

A final point that needs to be addressed in this discussion is whether the approach I have taken in this thesis will scale up or not. This is a rather difficult question and is only possible to answer within a constructivist picture of cognition.

Constructivism proposes that all human skills and behaviors are learned afresh by every individual. However, it differs from older “tabula rasa” theories of learning in that, it views the agent as an active constructor of knowledge rather than as a passive, empty space where experience is transcribed upon.

There is, as yet, no satisfactory computational model of the constructivist theory. This is probably due to the difficulty of developing such a model. The requirements of a constructivist model are extremely difficult to meet. As a matter of fact, (and this is the basic point that I am trying to get at) the requirements the constructivist theory places on cognitive models are so harsh that such models can be deduced by a top-down analysis. This is so because, such models can only be based on well-founded and general ideas (like the one I use, i.e. a hierarchy of predictions), and there should be very few such well founded and general ideas. I do not have at this moment any picture of what such a model looks like, but I believe it can be discovered, if Constructivism is after all, correct.

Below I will look at two cognitive systems that are believed to exist in the brain and discuss how they may be related to the prediction mechanism.

4.4.1 An Attention System

I have listed one of the functions of the prediction mechanism as distinguishing between interesting and non-interesting information. However, interesting information that the prediction mechanism reveals must be of use somewhere else other than the prediction mechanism itself. This “interestingness” property of the information generated by the prediction mechanism, must serve as a basis for the interaction between the prediction mechanism and other components of the cognitive system.

An attention system, whose basic function should be keeping the agent focused on a certain stream of information flow by suppressing other streams, seems to be an ideal candidate for interfacing to the prediction mechanism. The nature of the interface between these two systems can be expected to be in the following way: In general, information that was not predicted by the prediction

mechanism will be interesting for the agent. However “predictedness” or “unpredictedness” are not all-or-none attributes. We may suppose that the failure of predictors becomes more “unpredicted” as they get more reliable. Unpredictedness, and thus, interestingness, should therefore be graded attributes. Therefore threshold limits must operate to decide which pieces of information can be taken as interesting. An attention system can simply regulate these thresholds for controlling the information flow from the prediction mechanism to other systems.

4.4.2 An Episodic Memory System

An important observation is that the prediction mechanism may capture only frequently occurring and relatively simple regularities in the environment. It may, for example, recognize simple lines or predict when my foot will touch the ground when I am walking. However, it can not possibly generate a recognizer for an armadillo or predict the solution of a differential equation, because such things are either too infrequent or too complex to be captured statistically.

The episodic memory system is another component of the cognitive system that seems to be related with the prediction mechanism. The prediction mechanism, according to our discussion, works according to statistical relations. However capturing such statistical relations is not the only way to carry out predictions. For example, if we take a tour through a house we have not been at before, we will be able to make predictions like: “If I walk down the corridor and turn right I will arrive at the bathroom.” Such predictions can not be done by the prediction mechanism, since there is not enough data to derive statistical relations from.

An episodic memory system can help with doing better predictions by recording which predictions should be made in which context. In the above example, the episodic memory could store the path taken to the bathroom and regulate the prediction mechanism to do the right predictions at the right time.

Chapter 5

Conclusions

I started this thesis by stating that the capacity to carry out predictions should be built into an intelligent system. I tried to support this claim by showing that prediction had certain epistemological virtues (it was a practical measure of knowledge and it did not need any external error criterion) and therefore it could have uses within the cognitive system. In the second chapter, I tried to show that traditional models of cognition lacked sufficient explanatory power by referring to certain pieces of evidence and some well known problems. I discussed specific merits and shortcomings of certain approaches to modeling cognition. In the third chapter, I presented a hierarchical prediction mechanism that I argued could account for human low-level sensory-motor organization. I tried to demonstrate how the system could help in generating actions, reveal interesting events and result in the emergence of high-level representations. Finally, in the fourth chapter I tried to answer some possible questions regarding the prediction mechanism, discussed its functions and capabilities and how it could be related to the general cognitive system.

A number of themes emerge from this discussion:

The prediction mechanism described in Chapter 3 may actually exist in the brain in some or other way, and its rules and workings can help us in developing a better understanding of at least lower levels of the brain. I see this as a serious possibility, one more plausible than the existence of traditional models of perception given their limited scope and explanatory power.

The constructivist pursuit to build “order from noise” seems to be feasible, at least for the simplest cases. The framework presented in this thesis may be

somehow linked to a general development theory. The general methodology employed in the thesis may also be useful in studying higher levels of organization.

It is possible that there are other cognitive mechanisms whose functionings are based on simple and sound principles (like prediction) and thus can be inferred by a top-down analysis.

The constructivist theory of cognitive development sufficiently constrains possible computational models. However, existing frameworks in AI seem to be inadequate for studying constructivist models. There exists the need for a different kind of framework that offers different approaches to problems like learning and representation.

I find it rather difficult to discuss what kind of “future work” may follow this study. Obviously the hypothesis of the prediction mechanism has to be verified again and again. Computational models of the prediction mechanism must be built without hesitating to fill-in what I dismissed as “implementation details” in this thesis. These models must be tested within simulated or real environments and engineered towards better functionality. Computational models must be matched against neuro-scientific evidence with the hope of finding correlations or perhaps hints. All of this, of course, is no easy job, and can not be justified without evidence in favor of the feasibility of the ultimate enterprise.

This thesis, I believe should be regarded as a question rather than an answer. My concern has been, in a sense, to show that it actually is a good question, one that is worth investigating. I have tried to demonstrate that there may be something interesting along this side of the “search space” for a model of cognition. I hope that the material I have presented has fulfilled this purpose.

Bibliography

- [1] M. H. Bickhard and L. Terveen. *Foundational Issues in Artificial Intelligence and Cognitive Science*. North-Holland, 1995.
- [2] R. A. Brooks. *A Robust Layered Control System for a Mobile Robot*. MIT AI Memo No. 864, 1985.
- [3] R. A. Brooks. Intelligence without Representation. *Artificial Intelligence*:47:139–159, 1991
- [4] R. A. Brooks. *Intelligence without Reason*. MIT AI Memo No. 1293, 1991.
- [5] R. A. Brooks and L. A. Stein. *Building Brains for Bodies*. MIT AI Memo No. 1439, 1993.
- [6] D. Davenport. Inscriptors: Knowledge Representation for Cognition. In *Proceedings of ISCIS-8*, L. Gün, R. Önvural & E. Gelenbe (editors), Istanbul Nov. 1-3, 1993.
- [7] D.C. Dennett. Cognitive Wheels: The Frame Problem of AI. In C. Hookway (editor). *Minds, Machines, and Evolution: Philosophical Studies*, pages 129–151. Cambridge University Press, 1984.
- [8] D. C. Dennett. *Consciousness Explained*. Little, Brown and Co., 1991.
- [9] G. L. Drescher. *Genetic AI —Translating Piaget into Lisp—*. MIT AI Memo No. 890, 1985.
- [10] J. Fodor. Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology. In J. Haugeland (editor), *Mind Design*, pages 307–338. MIT Press, 1981.
- [11] J. Fodor and Z. W. Pylyshyn. Connectionism and Cognitive Architecture: A Critical Analysis. In S. Pinker and J. Mehler (editors), *Connections and Symbols*, pages 3–71. Bradford Books, MIT Press, 1988.

- [12] H. Gardner. *The Mind's New Science*. Basic Books, 1985.
- [13] C. M. Gray, P. König, A. K. Engel and W. Singer. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* 338:334–337, 1989.
- [14] D. O. Hebb. *The Organization of Behavior*. John Wiley, New York, 1949.
- [15] B. Kalb, I. Whishaw. *Fundamentals of Human Neuropsychology*.
- [16] J. McCarthy and P. J. Hayes. Some Philosophical Problems from the Standpoint of Artificial Intelligence. In B. Meltzer and D. Michie (editors), *Machine Intelligence 4*, pages 463–502. Edinburg University Press, 1969.
- [17] J. Piaget. The Psychogenesis of Knowledge an Its Epistemological Significance. In M. Piatelli-Palmarini (editor), *Language and Learning*, pages 23–34. Routledge and Kegan Paul plc., London, 1980.
- [18] M. I. Posner and M. E. Raichle. *Images of Mind*. Scientific American Library, New York, 1994.
- [19] *Science*, 274:339–340, 1996.
- [20] L. Shastri and V. Ajjanagadde. From Simple Associations to Systematic Reasoning: A Connectionist Representation of Rules, Variables and Dynamic Bindings Using Temporal Synchrony. *Behavioral and Brain Sciences*, 16:3: 417–451, 1993.
- [21] T. Van Gelder. Compositionality: A Connectionist Variation in a Classical Theme. *Cognitive Science*, 14:355–384, 1990.
- [22] A. Yavuz and D. Davenport. PAL: A Constructivist Model of Cognitive Activity. In A. Riegler and M. Peschl (editors), *Proceedings of the International Conference New Trends in Cognitive Science (NTCS 97)*, pages 207–213 Vienna, Austria, 1997.