Die approbierte Originalversion dieser Dissertation ist an der Hauptbibliothek der Technischen Universität Wien aufgestellt (http://www.ub.tuwien.ac.at).

The approved original version of this thesis is available at the main library of the Vienna University of Technology (http://www.ub.tuwien.ac.at/englweb/).

DISSERTATION

From Neuro-Psychoanalysis to Cognitive and Affective Automation Systems

Submitted at the Faculty of Electrical Engineering and Information Technology, Vienna University of Technology in partial fulfillment of the requirements for the degree of Doctor of Technical Sciences

under supervision of

Prof. Dr. Ing. Dietmar Dietrich
Institute of Computer Technology
Vienna University of Technology
Austria

and

Prof. Etienne Barnard Ph.D.
Human Language Technologies Research Group
Meraka Institute
Republic of South Africa

by

Dipl.-Ing. Brigitte Palensky
Matr.Nr. 8925308
Rembrandtstraße 22, 1020 Vienna

Vienna, February 28, 2008

Kurzfassung

Der Bedarf nach besserer Prozessinformation führt zu einer stetigen Erhöhung der Anzahl der Sensoren in Automatisierungssystemen. Die dadurch entstehende Komplexität wird durch dynamische, ungewisse und komplexe Einsatzumgebungen noch verstärkt. Traditionelle, regelbasierte Steuerungen stoßen dabei an ihre Grenzen, neue, adaptive und flexible Lösungen sind zukünftig gefragt. Diese Arbeit präsentiert einen neuen Ansatz technische Systeme mit den Erkenntnisfähigkeiten des menschlichen Geistes auszustatten. Die Basis dafür sind Ergebnisse einer jungen, aber produktiven Wissensdisziplin – der Neuro-Psychoanalyse. Diese ermöglicht die Entwicklung einer funktionellen, kognitiven Architektur – angelehnt an das Ich/Es/Über-Ich Modell von Sigmund Freud – mit der technische Systeme zur Entscheidungsfindung ausgestattet werden können. Ein zentraler Punkt ist die ganzheitliche Sicht von Wahrnehmung und Aktion. Automatisierungssysteme oder Roboter werden mit affektiven Mechanismen der Bewertung (Triebe, Emotionen, Wünsche, etc.) versehen, die es ihnen erlauben, wahrgenommene Sensordaten in bedeutungsbehaftete Informationen und kontext-spezifisches Wissen zu verwandeln, das wiederum die Basis für die Auswahl geeigneter Handlungen bildet. Der Kern der Architektur ist ein Gedächtnis, das individuelle Erfahrungen eines konkreten Systems auf emotional besetzte Weise abspeichert. Der stetige Fluss der Wahrnehmungen wird durch diese bewertet abgespeicherten Erinnerungen von Ereignissen und eigenen Handlungen, inklusive deren Folgen, gefiltert, um hochwertige Entscheidungen für die aktuelle Situation zu finden. Erfahrungen, die in der Vergangenheit als positiv empfunden wurden, sowie Entscheidungen, die indirekt zu einer positiven Empfindung geführt haben, werden wieder angestrebt. Bereits erfahrene und abgespeicherte Sequenzen dienen auch dazu, den Ausgang von aktuellen Vorgängen zu antizipieren und im eigenen Sinne zu beeinflussen. Es wird untersucht, wie der gewählte neuro-psychoanalytische Ansatz das Design der kognitiven Architektur bestimmt, sowohl bezüglich der konstituierenden Elemente (Konzepte, Funktionen, Datenstrukturen, etc.), als auch der strukturellen Organisation und der darauf ablaufenden Prozesse. Erste prototypische Tests der neuen Konzepte werden beschrieben um zu demonstrieren, wie die einzelnen Teile der viele Ebenen umfassenden Architektur interagieren und während des Betriebs aufeinander einwirken. Eine zukünftige, vollständige Implementierung der Architektur stellt in Aussicht, dass technische Systeme dieser Art, trotz aller Komplexität, nicht nur ein kontext-sensitives Verständnis ihrer Umgebung entwickeln können, sondern auch ein Bild ihrer *selbst* als handelnde Akteure.

Abstract

Automation systems are becoming increasingly complex, driven by a steadily increasing numbers of sensors for better process information. Additionally, such systems will be required to act in dynamic, uncertain, and complex environments. Traditional, rule-based models are too limited to create suitable adaptive systems; consequently, more flexible descriptions and solutions are necessary. This thesis presents a new approach of functionally translating useful human mental capabilities to technical systems via the construction of a unified cognitive architecture based on a combination of neurological and psychoanalytic findings and concepts – two sciences which have only recently embarked in promising cooperative directions. Psychoanalytically, out of the many possible frameworks, the architecture is inspired by the id-ego-superego model of S. Freud. A central aspect of the approach is an integrative view of perception and action. Automation systems or robots are equipped with evaluative psychic mechanisms (e.g. drives, emotions, and desires) that enable them to autonomously and adaptively turn perceived sensor information into meaningful pieces of knowledge, needed for the selection of appropriate actions. An important part of the architecture is the introduction of a system-specific memory storing individual experiences in an emotionally tagged way to constantly process the perceptual present through the filter of the past in order to reach decisions on 'what is currently the best thing to do'. Previously successful experiences are desired to be repeated. Known sequences of actions and events are projected ahead to anticipate what will happen next, and to evaluate alternative behavioral paths in an off-line fashion. The proposed cognitive architecture is informed by several aspects of the chosen approach concerning its constitutive elements (e.g. concepts, functions, data structures), its organization, and its processes. A simple prototypical implementation of the architecture is described to illustrate how the various functions on the many levels of the architecture work together. This serves to demonstrate the potential of the proposed architecture when 'in action' and supports the hope that – despite all complexities – one day, when effectively implemented, the architecture can lead to technical systems that construct a context-sensitive picture about their environment – and also about themselves as subjectively planning and feeling agents.

Preface

The number of sensor values automation systems have to deal with per time unit will increase dramatically in the not so distant future. Moreover, there is also the demand for systems that can act in highly dynamic, complex, and uncertain environments. Traditional, rule-based models mainly used in the field so far are not adaptive enough to meet these requirements, more flexible descriptions and solutions are necessary.

The fields of artificial intelligence (AI) is vast and has already seen several changes of the prevailing paradigms, from classical symbolic AI, over neuronal nets and other distributed and statistical approaches, to embodied cognitive science. Several cognitive mechanisms and architectures have been proposed. More recently, there has been an increased focus on the role of emotions in AI. Again several systems have been proposed. Most of them are either too low-level, ethology-inspired, or too rule-based, appraisal-oriented. *What is missing is a comprehensive model* unifying low and high-level capabilities. Some people have already suggested such models, however, almost no one (with very rare exceptions) has done this by consulting the insights of psychoanalysis. The so far suggested comprehensive models are either a) not coherent enough, or b) they stay too vague, just arguing the need for this or that mechanism without specifying how it could be realized in detail. Psychoanalysis can remedy both of these shortcomings.

The work intends to design a complex, autonomous control system by taking a prominent biological system as inspiration: the human mind, being able to filter vast amounts of information and to make good decisions in confusing and conflicting situations. The suggested cognitive architecture is based on neurological findings as well as on psychoanalytic principles, translated into a technical language. Drives and desires to motivate actions, basic and complex emotions to evaluate situations, different types of memories, planning ('acting-as-if') capabilities, and conflict resolution mechanisms are introduced as important functional elements. All these elements are arranged using the id-ego-superego model of Freud as template. The model helps to determine how to combine the processing of current external demands with the processing of current internal needs and currently available actions. A successful combination of these three elements (that is, one that serves the system's goals, the most fundamental of which being 'survival') makes up the core of intelligent behavior.

Of particular importance for situation recognition and categorization – two key capacities of intelligent behavior – is the use of predefined images as templates and the introduction of an emotionally afflicted episodic memory. Moreover, the system shall not only passively react, but actively build up expectations about what is supposed to happen next (*'focus of attention'*).

One intended target area of application is smart building automation, in particular care for handicapped and elderly people. Another potential application are mobile service robots. The suggested neuro-psychoanalytically inspired cognitive architecture is not only supposed to deliver more context-aware autonomous systems than the ones existing today. It shall also produce technical systems that possess some 'insight into the psychological functioning' of human beings. This is important for enabling technical systems to decide, for example, when a situation becomes potentially dangerous for a human being. Finally, the performed work will also be able to contribute to the field of psychological research.

The contents of the chapters is as follows:

Chapter 1 delivers the motivation for the work, followed by a description of the ARS (Artificial Recognition System) project of which the particular work is a part of. Finally, the goals and the methodology of the work are outlined.

Chapter 2 gives a brief overview of the history of artificial intelligence and cognitive science. The prevailing paradigms and their influence on research, modeling, and technical design are described.

Thereafter, a discussion of fundamental issues about the possibility of creating machines with mental capabilities ensues. How can the matter/mind problem, that is, the problem of symbol reference, be solved? How are the world and our images of it coordinated? What do we understand by information processing, and what is the essence of meaning? How can symbolic representations be constructed or 'emerge' out of distributed ones? The fundamental importance of *feedback loops* – a recurrent topic throughout the whole work – is for the first time stressed, investigating their role in the establishment of symbolic relationships and meaning. Finally, the hierarchic nature of symbolic relationships is outlined.

Chapter 3 describes, in its first part, important design principles of embodied cognitive science, among them most prominently system-environment coupling, embodiment, and value systems. These principles are stated because they also apply to the design of the proposed cognitive architecture. However, the new architecture goes beyond the framework of embodied cognitive science, by including many additional principles (coming from neuro-psychoanalysis). There is the claim that the new architecture can outperform solutions derived just under the paradigm of embodied cognitive science. The validity of this claim is partly investigated in the later chapters of this work, here, just a few hints are given. In the second part of Chapter 3, there is a broad discussion of drives and emotions. They are recognized as realization of the above required value systems. It is shown, how, during the course of evolution, they have emerged as a hierarchy of evaluation mechanisms, aimed to motivate, guide, and control behavior. It is also described how they can act as bridge between the body and cognitive capabilities.

Chapter 4 describes and discusses the technical state of the art, that is, cognitive architectures and computational systems using emotions. These systems can be compared to the newly proposed, psychoanalytically-inspired cognitive architecture because drives and emotions are in both cases key elements, although the new architecture also incorporates some further important functional principles. Related computational work directly referring to psychoanalysis (rather than neuro-psychoanalysis) is very rare. One exception is described.

Chapter 5 explains the neuro-psychoanalytic approach to the many level phenomenon human mind. It proceeds in both directions, combining bottom-up and top-down analysis. First, some selected facts about the human brain as described by the neurosciences are presented, then a description of the human mind as viewed by psychoanalysis follows, and, finally, the combined neuro-psychoanalytic picture is briefly sketched.

Chapter 6 is the main chapter of the work, presenting the new cognitive architecture. First, there is a discussion of the general design principles obeyed by the architecture. This serves to explain the general structure of the architecture. Thereafter, each of the modules of the architecture is described in detail. In particular, the implemented psychoanalytic principles are highlighted, showing how they guide the structural design as well as the functional design (the processes) of the various modules.

Chapter 7 deals with the implementation of the suggested architecture. First, general hardware and software design considerations and requirements are stated. So far, a simple version of the architecture has been implemented in the form of a software simulation. A virtual environment, called the 'Bubble Family Game', has been invented, producing cooperative as well as conflicting situations. A description of the Bubble Family Game serves to illustrate the architecture when *'in action'*.

Chapter 8 outlines the achievements of the approach along various dimensions. The potential of the architecture is compared with state-of-the-art systems, various strengths and weaknesses are outlined. Finally, potential applications are presented, and ideas for future work suggested.

Love and work.. work and love, that's all there is.
Sigmund Freud

Acknowledgements

I would like to thank Prof. Dr. Dietmar Dietrich for giving me the chance to write this work, for his continuous support, direction, and guidance. I would also like to thank Prof. Etienne Barnard Ph.D. for an inspiring discussion and many helpful comments.

This work would not have been possible without an intense collaboration with people from the field of psychoanalysis. My particular thank goes to Mag. phil. Mag. art. Brit Müller for critically reading the psychoanalytic parts of this work, and also to Mag. Patricia Pörscht, Dr. Georg Fodor, Dr. Elisabeth Brainin, Dipl.-Psych. Samy Teicher, and Mag. Dorothee Dietrich for sharing their profound psychoanalytic knowledge with me.

My fellow members of the ARS group at the Institute of Computer Technology deserve all my gratitude for being such a great team. Thanks for all the help, many fruitful discussions, as well as some really silly – but all the more funny – ones.

Love and thanks to all my family for continuous support, most particularly to my husband Peter, my son David, my brothers Helmut and Gerald, my mother and my father.

Contents

Abbreviations

AI	Artificial Intelligence
ARS	Artifical Recognition System
ARS-PA	Artifical Recognition System – Psychoanalysis
ARS-PC	Artifical Recognition System – Perception
A.T.	Action Tendency
BACS	Building Automation and Control Systems
B.E.	Basic Emotion
BFG	Bubble Family Game
Drv.	Drive
ERP	Enterprise Resource Planning
ICT	Institute of Computer Technology
MES	Manufacturing Execution System
NPSA	Neuro-Psychoanalysis
RAID	Redundant Array of Inexpensive Drives
R.I.	Recognized Image
T.I.	Template Image

Whatever course you decide upon, there is always someone to tell you that you are wrong. There are always difficulties arising which tempt you to believe that your critics are right. To map out a course of action and follow it to an end requires courage.

Ralph Waldo Emerson

1 Introduction

The objective of this work is to present a cognitive architecture for autonomous control systems inspired by two rather antagonistic sources of knowledge of the human mind: the *neurosciences* and *psychoanalysis* (see Chapter 5). It will be explored in which way a *combination of these two disciplines* can achieve the following:

- Motivate and inform the design of a multi-level but still unified, functional cognitive architecture
- Constrain specific mechanisms and processes within this architecture
- Inform the design of the data structures necessary to implement the architecture in a technical application

The resulting cognitive architecture for autonomous systems will include as a core element *affective* mechanisms such as drives and emotions. In recent years, there has been an increased interest in particular in the role of emotions within human cognition and decision making. This interest has spread out to the software and robotic agent community. It partly results from advances in cognitive neuroscience and emotion research, and partly from advances in agent technology and the desire for applications that require or benefit from the inclusion of different emotion-related aspects (e.g. autonomous agents, decision support systems, service robots, empathic human-computer interfaces, etc).

The surge of interest in emotions has led to a number of emotion-based architectures and applications. However, the work is often carried out in an 'ad hoc' manner. Due to the short history of the field and especially *the lack of an appropriate, coherent framework*, there is a still very limited understanding of how to design such architectures.

The aim of the work is to propose a new, *comprehensive and coherent* cognitive architecture that essentially includes affective elements and that still can be technically implemented, thereby getting closer to the idea of a machine that has a real understanding of what is going on around it. Sometimes in this work, such a machine is referred to as 'intelligent'. The term 'intelligence', however, cannot be given an exact definition that all people dealing with the topic are able to agree upon. See the beginning of Chapter 2 for a discussion of this issue. There, some aspects and manifestations of intelligence are reflected.

1.1 Motivation

Automation technology and automation networks [LDS01] – in particular building automation and control systems (BACS), industrial automation, and, to a lesser degree also mobile robots – were and still are one of the main research areas of the Institute of Computer Technology (ICT) at the Vienna University of Technology.

After contributing to the development and the international standardization of automation networks and systems in the 1990s (e.g. CLC/TC65, CEN/TC247[1]), for a period it seemed that, now, at the beginning of the new millennium, a range of established and mature technologies was available, and what was left to do in the coming years was just to improve, within the current conceptual state-of-the-art, existing algorithms, hardware components, embedded systems, protocols, or process management schemes. Technology-wise this might have even been (partly) true, but it should be remembered that technology is always just a vehicle for implementing services or functionalities people would like to have – and here, the list of demands is potentially unlimited.

BACS for example, being able to control lighting, air conditioning, and ventilation, were believed to be at the end of their evolutionary ladder. The reality is of course different. Currently, a modern, automated building management system is expected to include many more services than a traditional BACS can offer. For instance, the integration of the BACS data flow with that of other networks in a building, like office, security, or multimedia networks, is assumed to be beneficial but still a point where a typical BACS reaches its limits very quickly [Pal04], [DS00]. The interesting question now is: *If eventually all these networks and systems can seamlessly interoperate, what will result out of this*?

The answer is: There will be a plethora of heterogeneous data sources, flooding an application with information which partly may be of importance, and partly not. It is true that five cheap, maybe even diversified, sensors can often be more reliable and accurate than a single expensive one, but applying this strategy to future control systems will result in millions of 'data points' and the need of tremendous data fusion, comparison, level monitoring, rule checking, dependency processing, etc. Despite the fact that the necessary computing and data mining resources would need to be immense, it is expected that no one will be able to specify the operative rules of the system.

Take a complex, dangerous situation, like a fire in a very large building. Millions of sensors deliver large amounts of data. Some data might be of vital importance, some might be inaccurate or even wrong – but the lives of several hundreds of people can depend on the BACS' decision. The problem of how to decide gets more critical if the situation to be handled has not already been provided for by the system designer, like for instance via stored, simulated, and trained emergency and evacuation plans. What if the BACS should make decisions no one has thought of before? Human operators, if properly trained and working on the job for a long time, might be able to evacuate the floors in the best order, to close and open the right doors at the right moment, and to isolate the right parts of the building. At present, no one can imagine that a computer system could support or even replace humans in such tasks.

What is needed are technical systems that can mimic all the good capabilities of human decision making, e.g. the capacity to filter out what is relevant in a given situation, while avoiding human limitations, e.g. being tired or distracted).

[1] European committees for electrotechnical and general standardization, technical committees for fieldbus systems and building automation

1.2 Artificial Recognition System (ARS) Project

In 1999, D. Dietrich initialized at the ICT the idea to meet the future requirements of building automation by introducing bionic design concepts [DS00]. This was the first step towards the final goal of creating a highly fault-tolerant and performant control network for buildings (and also for other automation applications). Although the term *intelligent buildings* is already in use for simply automated or remotely controllable buildings, the ICT had and has something far more advanced as goal [DKM$^+$04], [DFKU07]. Real intelligent buildings shall be able to do the following:

- Recognize situations and scenarios
- Realize their impact
- Anticipate potential dangers
- Take appropriate countermeasures in time

As the size and costs to investigate an entire building are rather high for a research lab, a single room was chosen for a start: the kitchen. The Smart Kitchen Project [Rus03] was the first attempt to structure and to describe the problem of an 'intelligent' system in the sense as required above. The initial work dealt mainly with data acquisition, management, and categorization. It started out as a fairly classical bottom-up approach. The project, meanwhile renamed to ARS – *Artificial Recognition System* – was and is, however, not intended to develop only in this direction.

Out of the first steps, two main project sub-teams were defined:

- ARS-PA – ARS Psychoanalysis
- ARS-PC – ARS Perception

While the latter follows the initial Smart Kitchen Approach by clustering sensor information and by creating semantic symbols out of these [PLD05, Pra06, Bru07, Bur07], the ARS-PA team tackles the problem also from top to bottom [RLFV06, DL$^+$06, PLC07, Roe07, BLPV07, RLD$^+$07]. This work belongs to the ARS-PA part of the project, however, one special purpose of the work is *to present an integrative solution* that combines both approaches into one big picture.

The obvious strategy to construct a machine that is expected to keep up with humans when it comes to cognitive abilities is to take the human mind as a blueprint. Out of the many neurological, cognitive, psychological, sociological, and pedagogical models of 'how humans internally work', the findings of neuro-psychoanalysis (see Chapter 5) were chosen to be the most attractive for the ARS team.

The BACS of the future is still in the back of the ARS members' minds. The ARS system, however, is designed in a much broader sense. In particular, this work does not limit itself to building automation, but envisions intelligent systems in general, also including e.g. mobile service robots, or software agents autonomously mining the world wide web or other information resources, etc.

1.3 Problem Statement, Goal, and Methodology

> *'AI can have two purposes. One is to use the power of computers to augment human thinking, just as we use motors to augment human or horse power. Robotics and expert systems are major branches of that. The other is to use a computer's artificial intelligence to understand how humans think. In a humanoid way. If you test your programs not merely by what they can accomplish, but how they accomplish it, then you're really doing cognitive science; you're using AI to understand the human mind.'* [Sim69]

Although it is one of the objectives of the ARS project to *overcome* existing limitations of artificial intelligence by providing a completely new perspective – neuro-psychoanalysis – the two purposes mentioned in the quote above – the construction of problem-solving technical systems and the pursuit to understand human intelligence – still remain two principle directions of research that may be confluent but sometimes also contradictory.

As has been stated when describing the motivation for this work,

- *the problem* to be solved is to create a machine that can make sense of enormous amounts of data, and
- *the main goal* of this work is to present a comprehensive cognitive architecture based on neuro-psychoanalysis [nps07] that can potentially tackle this problem when fully elaborated.

Thus, the original motivation of the ARS project is to serve the first purpose mentioned in the quote. Further down below, it will be addressed that, in the long run, also the second purpose is aspired.

The *hypothesis* of the work is that the *combined neuro-psychoanalytic view* of the human mind can act as a basis for the design of an adaptive context-aware architecture. Autonomous systems equipped with the architecture shall get a basic understanding of what is going on around them, even in the challenging case that the environment they have to act in is dynamically changing. In such environments, traditional rule-based systems very quickly reach their limits. The new architecture shall provide a more flexible and context-sensitive solution. It is the belief within the ARS project, that, one day, out of the proposed neuro-psychoanalytic approach autonomous systems capable of dealing with unforeseen situations in a reliable way without human interference and control shall arise.

Although the usage of subjective insights about the functioning of the human mind is characteristic – and in fact defining – of psychoanalysis, the neuro-psychoanalytic view is the result of a truly scientific endeavor and as such, for example, fully compatible with an evolutionary perspective on intelligence and cognition. Out of this constellation, several sub-goals follow pursued in this work:

- One sub-goal is to point out what is novel and what can be specifically gained by using psychoanalysis (rather than some other psychological theory) as a basis for the design of the cognitive architecture.

- Another sub-goal, is to show how the chosen approach fits into a broad evolutionary perspective. Throughout the work, general bionic design principles derived from looking at intelligence as it appears in all its forms in nature are stressed and given a technical interpretation.
- A third sub-goal is to put this work in relation to existing approaches in the field of artificial intelligence and cognition.

When it comes to implementing the proposed general cognitive architecture, some further specifications have to be made, for example regarding the application in which the artificial autonomous systems shall prove themselves, the tasks they should fulfill, the desired behaviors they should produce, etc. All these aspects have to be referred to each other. In this respect, a fictitious demonstration environment for test purposes has been developed, called the *Bubble Family Game* (see Section 7.3). It is inspired by the idea of individual autonomous, embodied agents, coupled with their environment through sensors and actuators, potentially having to work together as a group, still making decisions on an individual basis, guided by the use of 'emotions', 'desires', and other functional concepts stemming from psychoanalysis, fulfilling individual as well as global tasks which require cooperative behavior.

As far as the control architecture is concerned, it shall also include 'psychological' elements. After all, the main idea of the presented approach is to use functional concepts for the creation of a cognitive intelligent artificial system that have been developed within a science studying the human psyche. Although autonomous systems or robots do not have the same bodily basis as living creatures, they obtain an abstract functionality of emotional assessment through appropriate implementation. Emotions act as a value system, improving goal-oriented behavior in conflicting and contradictory situations.

The suggested neuro-psychoanalytically inspired cognitive architecture is not only supposed to deliver more adaptive, context-sensitive autonomous systems than already existing. As stated in the introductory quote, apart from the goal of building a more intelligent machine, endeavors in the field of AI are also motivated by the old dream of mankind to technically create a human-like thinking and feeling being. Although this is not the original motivation of the ARS project, it applies to some degree. The proposed architecture shall result in technical systems that possess some *'insight into the psychological functioning'* of human beings. This is important for enabling a technical system to decide, for example, when a situation becomes potentially dangerous for a human being. Finally, because of this the work will also be able to contribute to the field of psychological research.

The *methodology* used in this work is given by the following principles:

Modularity – Component-oriented design is the key to simplified specification, development and testing. Complex systems cannot be built in any other way.

Bionic Approach – Nature shows a challenging level of functionality. We should consider the results of evolution to be very good solutions and aspire to copy them.

Strict Model – For describing the functionality of the human mind, the neuro-psychoanalytic model is considered to be the best choice. Only potential limitations will be covered with compatible alternative models.

Broad Theory – As this work is an early one in this branch of cognitive automation, an extensive theoretical survey and comparison with related work is necessary and will be given.

A beginning is the time for taking the most
delicate care that the balances are correct.
Princess Irulan
From *Dune* by Frank Herbert

2 Foundational Issues

Given that the thesis presents a new technical approach based on findings and models from a science studying the human psyche, the relation between physical and mental events and phenomena shall at least be briefly addressed. This is done in the second half of the present chapter with an exploration of the relationship between information processing and meaning, followed by a sketch of how meaning may arise out of matter. Before, in the first half of this chapter, an overview of the different research paradigms in artificial intelligence and cognition is given in order to shed a light on the difficulties of building a machine that can really understand.

The purpose of this chapter is three-fold:

- To put the newly proposed cognitive architecture into *context with existing approaches* and achievements
- To indicate how a *natural account of the relationship between matter and mind (psyche)* could look, and how it could diminish the gap between the objective sciences, like physics, biology, or neurology, on the one side, and the subjective sciences, like the sciences of the psyche, on the other side
- To *prepare the path* for showing how the psychoanalytic approach can fit into a technical picture and how it can contribute to the task of creating an understanding, context-aware machine

It shall be stated in the beginning that ‘intelligence’ is not a psychoanalytic term. Implicitly, intelligence is often viewed as property that can be ascribed to the conscious human mind when the latter is trying to solve problems. However, there is no generally agreed upon definition of intelligence [PS99, p. 6]. Below, two example definitions are stated. The first is rather passive, focusing solely on perception and inference processes, whereas the second explicitly refers to the fact that intelligence has to be viewed in the light of actions.

> **Intelligence**: *‘The essence of intelligence is the skill in extracting meaning from everyday experiences.’* [Unknown]

> **Intelligence**: *‘that which produces successful behavior’* [Alb96]

Often intelligence is just defined by a diverse list of topics that are required to be taken into account when trying to understand intelligence as it is characteristic for humans. In [PS99, pp. 7–12], the authors present such a list of *'what people in general think about intelligence'* which includes the following items:

- Graduated property
- Thinking and problem solving
- Learning and memory
- Language
- Intuition and creativity
- Consciousness
- Emotions
- Surviving in a complex world
- Perceptual and motor abilities

Note that in contrast to the general term 'intelligence', the list above does contain topics which are the object of psychoanalytic studies and models (see Chapter 5). Note furthermore that some of the above capabilities have been focused on by artificial intelligence (AI) and cognitive science ever since the beginning of these fields, whereas some of them have become a research agenda only more recently.

2.1 Paradigms of Artificial Intelligence

The endeavor to develop automata with intelligent behavior has a long and conflicted history [Kur90, RN04]. It gained considerable momentum with the appearance of the programmable computer in the middle of the twentieth century. The advent of computer science slightly preceded and strongly inspired artificial intelligence (AI) as well as cognitive science (see Section 2.2). In this work, artificial intelligence is understood as follows:

> **Artificial intelligence**: the science and engineering that seeks to build intelligent machines by copying, in one way or another, intelligence as it appears in nature.

Up to now, the history of artificial intelligence has seen many developments and also some turns in direction. It is not the task of this work to give a comprehensive overview. Only some aspects are highlighted which are important for understanding the premises on which the presented new cognitive architecture for autonomous systems is built. These aspects mainly deal with

- differences between top-down and bottom-up approaches, and
- questions concerning the various possible forms of representations, algorithms, and control mechanisms.

Along these lines, several 'phases' of artificial intelligence can be distinguished, see e.g. [LB07] for an overview. Natural intelligence has always been a role model and a source of inspiration, however, not in the same way for the different phases or approaches. In any case, no approach tries to imitate natural intelligence on a one-to-one structural level – which is no wonder considering the huge complexity of biological brains. Instead, each approach tries to build intelligent artifacts on a more abstract, functional basis, although some of the approaches look more 'brain-like' than others. In the following, a distinction will be made between 'symbolic AI', 'connectionist approaches to AI', and 'embodied approaches to AI', even though this distinction is often blurred in practice.

2.1.1 Symbolic Artificial Intelligence

The design process of symbolic artificial intelligence is *top-down.* Intelligence is viewed as *computations* which in turn are viewed as *rule-based manipulations on (mathematical) symbols* (see e.g. [Min91, Har90, Har02]). Objects of the external world, operations on them, but also internal goals – in short, everything is coded into symbols. Thus, out of their construction, the characteristic properties of classical AI systems are the following:

- An encoding of all the system-specific elements, relations, and operations
- Done in advance
- By the designer of the system (the programmer)
- In a completely top-down way
- Using an arbitrary code

For example, to model a game of chess, symbols for each of the black and white pieces, the squares on the board, and the moves each figure can make are introduced. To choose an appropriate move, various possible moves are calculated according to the encoded rules of the game. Then, the resulting configurations are evaluated, and the most favorable outcome – determined by a predefined 'fitness' function – is selected. All this is done on the 'symbolic level' whereby symbols are entities without internal structure and history, (pre-)defined in a purely syntactic way by how they relate to other symbols.

The chess example demonstrates two important components of the symbolic approach: *knowledge representation* and *search.*

From a mathematical point of view, knowledge representation is typically done by using concepts and tools from *formal systems theory* or from *logic.* However, it is simply too complex a task to represent real-world domains within the constraints of these formalisms [Min91]. Even when using tools such as semantic networks or frames for knowledge representation instead of some form of logical calculus, the general problems related to the classical view of cognition as computation still remain. These problems will be discussed below.

Problems and Advantages

The main problem with classical symbolic AI is that it can be only used when there is complete information about the part of the world to be modeled. For example, the set of *relevant* features whose changes have to be tracked has to be known in advance – in order neither to miss important changes nor to be forced to always evaluate every change occurring somewhere in the system. This is referred to as the *frame problem* (see e.g. [PS99, pp. 65–69]). Take the example of a person wanting to leave a room: It is irrelevant whether a bird has sat down on the window sill, but it is relevant whether someone has locked the room. Humans know a lot of such things, but machines don't. Having to process every change in the system not only dramatically increases processing time, it also requires a symbolic representation with sufficient representational power that it can capture all the possible changes of the domain to be modeled, i.e. all its components, relations, interactions, outcomes of interactions, outcomes of outcomes, etc. Beyond simple toy domains, it is not feasible to create such a representation *ab initio*.

Another big problem with classical systems is that they cannot be kept effectively in tune with changing environments, they lack *fault tolerance* and *generalization ability* [PS99, p. 63]. Their symbolic representations and the operations on them are too domain-specific, restricted, static, and time-consuming. This is mainly due to the fact that these representations are not grounded in the systems' interactions with their environment which is addressed by the *symbol grounding problem* referring to how symbols relate to the real world (see [Har90], also e.g. [PS99, pp. 69–71]). In symbolic AI, this relationship is never explicitly discussed which has been criticized most prominently by Searle [Sea80].

A further important point of criticism concerning classical symbolic AI is that its algorithms are exclusively *sequential* and (usually) *centrally controlled* [PS99, pp. 63–64].

The big advantage of classical symbolic systems is that their discrete knowledge representations are *explicit*, and *manipulable* in an open-ended manner [Pyl84].

2.1.2 Connectionist Systems

As an early rival of the symbolic model of mind, neural networks (Figure 2.1) were introduced [RM86]. Although they are still abstractions, compared to symbolic systems, neural networks are more inspired by the structure of real brains: What you have is a huge number of simple processing units (neurons) linked in parallel by an even bigger number of 'wires' and junctions (axons and synapses).[1]

With neural networks, the connectionist view of cognitive processes started taking shape [Smo88]. This goes hand in hand with the idea that knowledge can also be represented in a *distributed* form. Cognition is not viewed as symbol manipulation but *dynamic patterns of activity* in a multi-layered network of nodes with weighted positive and negative interconnections. Network constraints govern how input activations spread through the network. An important issue of the connectionist or distributed approach is *learning*. It is achieved by adjusting the connection strengths of the network during a training phase. For the different existing algorithms, see [Hay98].

[1] There are approximately about 10^{11} neurons, and each neuron is connected to a large number of other neurons via several hundreds to a few thousands synapses [Ste98].

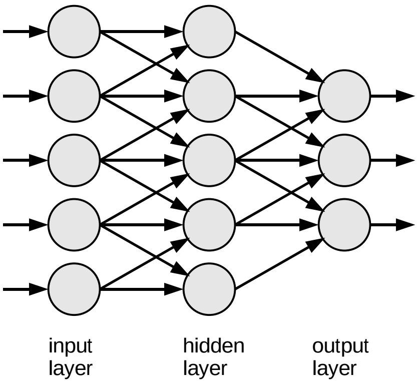


Figure 2.1: A classical artificial neural network

The mathematical tools used in the distributed approach mainly come from either *statistics* (for Bayesian networks, Markov Models, and the like) or from *dynamical systems theory* (which is based on differential equations). In fact, from a mathematical point of view, a neural network simply *is* a special kind of dynamical system (in the narrow mathematical sense), the defining differential equation being the learning rule governing the update of the connection strengths (e.g. delta rule, Hebbian rule, etc.).

Problems and Advantages

A learning organism does not need a complete description of the domain before attempting a solution. Instead, *information is extracted from the statistical properties* of the environment. Thus, neural networks and other methods using statistical regularities of the environment (e.g. Markov models, etc.) are able to generalize well to unseen cases [PS99, p. 176]. They can be *robust to exceptions, noise, and incomplete data*, and they can automatically model the hidden influences of apparently unrelated phenomena. For these reasons, they are often used for AI systems that deal with the tasks of processing and categorizing 'sensory' input, like vision or speech signals.

The problem when designing a connectionist system for a specific task is that successful performance often depends crucially on a careful pre-processing of the data sets used for training the system [KK92]. Usually, there are very many parameters that have to be determined which is not an easy exercise.

Another problem when trying to build intelligent systems with the connectionist approach is that their pattern recognition and categorization capabilities are exclusively based on *implicit representations* (see e.g. [Min91]). This is also connected with an advantage: They can occur bottom-up without having to be predefined. There is however a big disadvantage of nets. Unlike symbols, the patterns of interconnections *do not decompose, combine and recombine according to a formal syntax* that can be given a systematic semantic interpretation. This is a serious limitation of connectionist systems, especially when trying to build bigger systems that have to fulfill a variety of hierarchically structured tasks [FP88].

Finally, a negative aspect (in which classical symbolic systems, by the way, are not better than connectionist systems) is that mind is still treated as passive resource that files in information but that is not *intrinsically geared to take action.* This point of criticism is mainly issued by proponents of a more behavior-oriented view of cognition described in the next section.

2.1.3 Embodied or Behavior-Based Cognitive Science

In the early days of artificial intelligence, expectations that computers were only a small step away from producing intelligent behavior were widespread. Computers even became a metaphor for the brain. By the end of the 1980s, some researchers from artificial intelligence, as well as brain and cognitive science, realized that this view was maybe misguided or at least too limited. The brain was not designed to 'run programs' for specialized purposes, like performing logic or playing chess. Instead, as evolutionary theory tells us, the brain has evolved *to control our behavior* such that we can *survive in highly dynamic environments.*

One of the key elements of the new perspective is that intelligence always has to manifest itself in *behavior.* Rodney Brooks, one of the founders of the field suggested to *'do away with thinking and reasoning'* [Bro91a] and focus on the *interaction* of the organism or artifact with the real world. He called the new view *embodied intelligence* as the role models are now active *agents* and not passive programs any more.

A fundamental aspect of embodiment is that agents have to be situated in their environment. They can be biological, or robots, or pieces of software, but, in any case, agents have *to be subjected to the influences* of their environment – whether real or virtual – and they have to interact with their environment. This implies that they must have sensors via which they can acquire relevant information about their environment, and a motor system to act on their environment.

Problems, Advantages, and Prototypical Agents

One advantage of the new approach is that the heavy use of the system-environment interaction *minimizes* the required amount of *world modeling.* In the mid-1980s, R. Brooks argued that the *sense-model-plan-act* paradigm of traditional AI (Figure 2.2.a) was less appropriate for autonomous robots as the process of building world models and reasoning using explicit symbolic representational knowledge often took too long to produce timely responses. As an alternative, he developed a layered control system with a new functional decomposition into task-achieving modules called the **subsumption architecture** (Figure 2.2.b). Each of these modules can produce a specific behavior, for example wandering around, relatively independently of the others. Therefore, when extending the system, new modules can be incrementally added on top of the others without having to alter the existing ones. However, the modules are connected to each other by connections that can suppress input to modules or inhibit output. From an engineering point of view, the subsumption architecture is attractive because it is *robust* and *easily extendable.* From a cognitive science perspective, it contributes to the idea that *intelligence can arise or emerge* from a large number of simple, loosely coupled parallel processes.

Another famous category of agents that is ideally suited to study the relationship between implemented mechanisms and arising behavior are **Braitenberg vehicles**, named after their inventor, the neuroscientist Valentino Braitenberg [Bra84]. Vehicles are very simple machines or robots situated in an environment containing heat or light sources (Figure 2.3). The architecture of a vehicle mainly consists of almost direct connections between sensors and motors. By making small variations of how to connect sensors and motors – either laterally or counter-laterally – and how to relate sensor intensity to motor intensity – either positively or negatively – different kinds of behavior arise. Thus, vehicles show behavioral patterns which are not directly programmed. These behaviors can look quite complex. Note that Braitenberg vehicles have no internal representation at all.

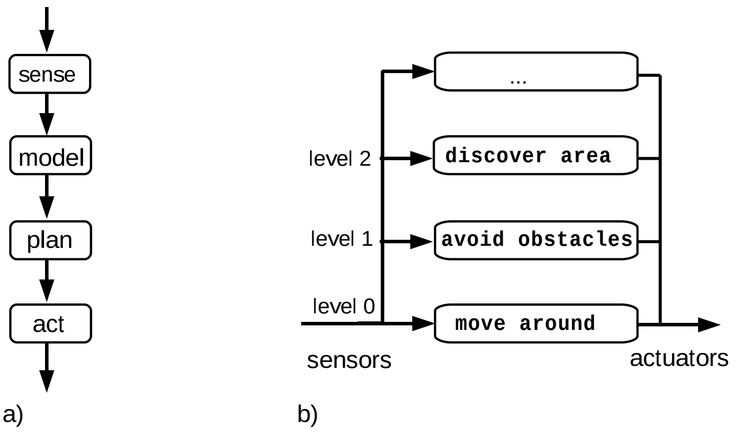


Figure 2.2: a) The traditional decomposition of an AI control architecture versus b) the 'task-achieving' decomposition of the subsumption approach (based on [Bro86]).

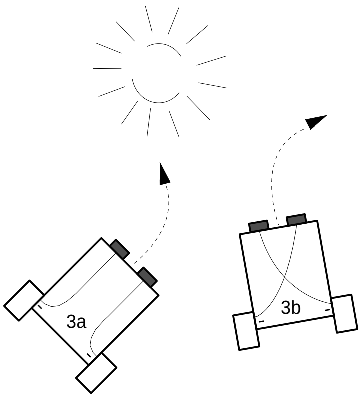


Figure 2.3: Braitenberg vehicles of type 3 (after [Bra84]). The sensors of these vehicles exert an inhibitory influence on the motors: The more the sensors are activated, the less power is delivered to the motors (indicated by a minus sign). Vehicle 3a turns towards the light source (and stops somewhere close), whereas 3b turns away. Both of them slow down in the surroundings of the light source, and, thus stay there for some time.

The two presented archetypes demonstrate a common deficit of behavior-based approaches, whether in AI or robotics: The capacities of the systems mostly stay limited to lower-level, sensorimotor capabilities. Higher-level cognitive abilities, like planning or language skills, are not considered. Moreover, often (although not always) only implicit, non-symbolic representations are used (which has the disadvantages as discussed in Section 2.1.2).

Another famous embodied agent comes from the Japanese psychologist Toda [Tod82]. The agents described by him are the earliest role models for *complete agents* (for a definition see 3.1.1). Already in the 1950's, Toda was discontent that cognitive psychology was only focusing on complex planning and decision making strategies in order to analyze intelligent behavior. As an alternative, he designed an autonomous agent, the **fungus eater** robot, together with an artificial, science-fictionally inspired environment. Sent to a distant planet, the robot has the task to collect uranium ore. Any activity of the robot requires energy. In order to survive, it has to eat wild fungi growing on the surface of the planet. Both resources, ore and fungi, are distributed randomly over the foreign planet. The agent is equipped with a relatively simple but *complete* set of rules – including everything it needs to survive, e.g. sensory, locomotion, decision making, and collecting capabilities. Additionally, Toda provides the robots with ‘urges’. Urges are small, predefined subroutines that are activated once a situation has been identified as being relevant to some vital concern. Each urge is linked to a specific action. Toda describes a whole set of those urges and calls them ‘fear’, ‘anxiety’, ‘love’, etc. He claims that urges would make the fungus eaters ‘emotional’ and that being emotional would make their behavior more intelligent. With his concept of the fungus eater, Toda was one of the first to argue that emotions can contribute to intelligent behavior. This will be explored in more depth in the rest of the thesis.

2.2 On the Matter-Specifity of Mind

Natural intelligence has not only inspired artificial intelligence. Understanding has also passed in the opposite direction. The attempts to artificially build an intelligent artifact have changed the way we think about our natural cognitive capabilities.

The brain is an organ like any other – except that it is the seat of the mind. For hundreds of years, philosophers and scientists have tried to understand the relationship between *matter* and *mind*, whether they are two quite distinct things, or whether one can arise out of the other – and in this case, which would be the more fundamental. One of the most prominent proponents of the dualistic view of matter and mind was Renè Descartes [Des75]. Over hundreds of years, his view has proved very seductive which is no wonder given the fact that we constantly *experience* ourselves *in a first-person form* as being there, reasoning about what is going on around us, and deliberatively setting actions. All these capabilities *subjectively* feel as existing independently from the material world around us – and clearly distinguishable from it. At first sight, our thinking human mind appears to owe nothing to our body and our surrounding environment. Correspondingly, the sciences studying the subjective aspects of the mind (the ‘soul’ or the ‘psyche’) have been almost completely separated from the sciences studying the mind – or better, the brain – as a natural object for hundreds of years.

With his early work dating back to the early 1890s (e.g. [Fre91]), S. Freud was one of the first to investigate how the mind (psyche) may be derived from the underlying physical processes of the brain. It was only from the 1950’s onwards (after the dominant period of the behaviorists) that various sciences formed a loose coalition under the name *cognitive science* and again seriously took up the goal of understanding how the mind may arise out of a ‘material machine’, the brain. Originally, researchers were mainly from psychology, linguistics, philosophy, neuroscience, and computer science, whereas, only more recently, biologists, engineers, and others have joined the interdisciplinary effort [PS99, p. 39]. This corresponds to the fact that, in the beginning, the information processing metaphor was the predominant paradigm of the field.

> **Information Processing Metaphor**: The view that cognition is in essence computation. The central processes of the brain – finally producing mind – are considered as being analog to the processes occurring in a computer: information storage, copying, matching, retrieval, building up internal knowledge states, drawing logical inferences.
>
> The brain is acknowledged to be the ‘mechanistic’ underpinning of the mind, but its *specific* material embodiment is assumed to be exchangeable and therefore of no importance. The prevailing assumptions are (compare with [Har02, pp. 297–301]):
>
> - Mental states are computational states.

- Computational states are *completely* implementation independent.
- Passing the Turing Test (which completely abstracts from any bodily features) is the only and therefore decisive criterion for intelligence.

The objects of study are arbitrarily assigned, abstract symbols which represent events and characteristics of the external environment. They can be manipulated independently from the actual aspects of the environment they represent, and without taking recourse to their physical carrier substance.

2.2.1 Cognition as Computation: What is Missing

What is left out of the picture is the question of the *origin* of these symbols. There are no considerations of how representations come to represent the environment in the first place, or how symbols come to have *meaning* (symbol-grounding problem). Thus, the old opposition between matter and mind persists. The brain is recognized as the material carrier of the mind, but it is given no attention. The starting point of research is abstract information processing. The focus lies on creating intelligence, and not *behavior*. Thereby, intelligence is modeled as *symbolically encoded solutions to symbolically encoded puzzles*.

The problem with this is not the usage of symbols, but how the symbols are derived, and the fact that the body proper of an organism and the environment where it is embedded are not taken thoroughly into account. No attention is given to how the body and the local environment are *literally* built into the processing loops. In Section 3.2 of the next chapter, exactly this specific topic will be investigated in more detail, while in the following of the current chapter, the analysis of meaning and how it can be created will be pursued in a more general manner.

In the previous section, it was mentioned that the classical approach to AI, referred to as symbolic or computational, typically leads to systems with a narrow field of expertise which are hard to extend. Such systems are not able to perform well in highly dynamic, permanently changing environments.

2.2.2 Towards a Remedy?

As has already been mentioned, in the last years, a new direction of research appeared within AI and cognitive science, called embodied cognitive science. Researchers in the field realized that they had thoroughly underestimated the complexity of *everyday behavior*. They had built computers (like Deep Blue) that could play world champion-level chess (which, by the way, did not require those computers to be particularly intelligent), but they could not get a real robot to navigate successfully a crowded room.

Instead of trying to model human intelligence, researchers adhering to the new paradigm of embodied intelligence often try to model at least animal intelligence. In any case, there is an increased focus on the following aspects:

- Sensorimotor capacities
- Solutions that do not require huge knowledge bases
- Replacement of detailled representations by interactions

The nature of the interactions influences the formation of the emerging cognitive capacities, and it specifies some of their properties. For the most extreme proponents of the new view, mind is a dynamical process that cannot be separated from the environment at all (see [Cla97, pp. 148] for a list of references). It just emerges from the (nonlinear) interactions of many material components. Information processing constructs, such as representations and symbols, are completely rejected, opting instead for an exclusively dynamical systems description. This means that mind is viewed as a material process following the laws of the dynamics of matter. All there is to do, is to describe the general dynamics of many material components.

Thus, the old philosophical matter-mind debate has been transformed within cognitive science into a *computation versus dynamics* debate. It is very much linked to questions about the nature and the origin of representations and symbols. This topic will be discussed in more detail below.

2.3 From Matter to Meaning

Biological organisms are able to reproduce, *to initiate* actions, and to select responses, instead of *simply being pushed around* by the physical forces acting on them as it is the case for inorganic matter. This achievement was only possible with the ability *to acquire, process, and use information.* As information is very hard to define – and no one has managed to do it so far in a satisfying manner – the term will be discussed rather than given an exact definition.

The following questions will be addressed:

- What *types* and *forms* of information are there?
- How can information be *acquired, manipulated, stored, and transmitted*?
- What is the relationship between information and *meaning*?

2.3.1 Matter and Information

Information seems to be a fundamental property of nature similar to matter or energy [Sto97]. To characterize information, I would like to start with the following, very general description:

> **Information**: *'any type of pattern that influences the formation or transformation of other patterns.'* Unknown

In [EF94, pp. 55–58], the authors distinguish between *'bound'* and *'free'* information.

Bound Information – Principally, every physical system contains bound information. For example, the color of the sunlight carries information about the chemical components the sun is made of, or the geologic layers of the surface of the earth contain information about the history of our planet. This kind of information *is just there*, it represents itself.

Free Information – Free information has a different quality. It is always part of a *relation* between two systems. Therefore, it is not a property of one system alone. Free information has a *purpose.* After it has been extracted out of bound information, it can be processed, manipulated, exchanged, and stored. The storage requires the back transformation of free information into some kind of bound information (the storage medium), however, free information is relatively independent from its physical carrier. For example, a message can be delivered as a letter, a telephone call, an email, or a file on a memory stick. The specific material carrier is not relevant for the contents of the message (although it makes a difference in terms of speed, reliability, accessibility, etc).

The most important requirement for a physical system to act as a storage or transport medium for information is that it has a high degree of stability. Consider the example of sound waves produced during a concert or a speech. They are a very transient phenomenon but they can be stored, for instance, on a magnetic tape which is a more permanent form of storage.[2]

As has been stated, free information can be extracted from bound information. In [EF94, p. 57], this is viewed as a phase transition and called *ritualization* or *symbolization.* Originally, the term 'ritualization' is used (in ethology) *to characterize the transformation of an everyday action into an action with a special meaning.* The process is fundamental for human culture, but it can already be found in the animal regime. For example, many animals have very ritualized mating behaviors. These are more or less fixed sequences of actions which are performed before the eyes of potential partners. The displayed actions originally served other, more direct purposes like drinking water, hunting food, or building a nest. After the ritualization, as part of the mating sequence, the new, more abstract purpose of these behaviors is that the courting partner wants *to signal* to the wooed partner the qualities it has, i.e. what a good choice it would be. Essential for the process is that *both* partners *interpret* the displayed behavior in the same way. This signaling behavior to communicate information has certainly not the quality of human language which is one of the most powerful symbol systems, but it already shows some characteristics that are important for a symbolic relationship.

2.3.2 Signals and Symbols

A sign or a symbol is often described as *something that stands for something else.* Later on, the necessity for a symbol of being part of a symbol system, and the possibility of a symbol *just to initiate something else* (and not to stand for it completely) will be stressed. Whether standing for or just initiating, as every sign or symbol refers to something else, there are always two things or domains involved in a symbolic relationship, the thing itself, and the thing it refers to. There is also a third element: Both domains have to be connected via some kind of *code*, i.e. a convention or mechanism of how to map the two domains to each other. The *existence or establishment of a code* is thus a necessary requirement for the process of symbolization. Free information is always symbolic information insofar as its most characteristic property is its invariance related to its physical carrier: The exact form of the encoding does not matter as long as 'sender' and 'receiver' of a symbol understand or *interpret* it in the same way. The science that studies the relationships between information, signs, symbols, and meaning is called semiotics.

[2] The given example is one where information is stored by using a – more or less – stable state of a *conservative structure.* Physically, another possibility would be to use a *local attractor of a dissipative structure* as a storage medium. The first possibility has a more 'localist' character, the second possibility a more 'distributed' character. In either case, the once stored pattern can be 'replayed' at some time later. It is for some time decoupled from immediate physical interaction.

2.3.3 Symbol Systems and Semiotics

The research topic of *semiotics* are individual signs and symbols, as well as sign and symbol systems [Cha02]. It includes an investigation of how meaning is constructed and understood. One of the most important insights of semiotics is the famous triadic structure of any sign or symbol system. C. S. Peirce (1839 – 1914), one of the founders of semiotics, defined the process of semiosis as follows:

> **Semiosis**: *'[any] action, or influence, which is, or involves, a cooperation of three subjects, such as a sign, its object, and its interpretant'* [Pei06, p. 282]

Some 2-place sub-relations of this triadic relation have special names.

Semantics – This term addresses the relationship between signs and the world external to the sign system (the world of objects). Thus, semantics deals with the meaning expressed in a language, code, or other form of representation.

Syntax – This term refers to rule-based operations between signs within the sign system, that is, the construction of complex signs from simpler signs.

Pragmatics – This term refers to the practical use of signs or sign systems, that is, their interpretation in particular circumstances and contexts.[3]

According to semiotics, no sign or symbol has an absolute meaning. Meaning is always dependent on the *interpretation* of the code that maps the sign or symbol with the object it refers to. Actually, this is another way of saying that isolated chunks cannot be symbolic because symbols always need to be part of a symbol *system*. These systems can be designed by a designer, or emerge bottom-up by self-organizing and selectionist processes, or they are mixtures. In the emergent case, an existing feedback selection mechanism favors successful symbol-object mappings. In combination with such a feedback mechanism, it are exactly pragmatic relationships that contribute to the *formation* of the system in the first place (and also to its continuous *adaptation* if necessary).[4]

2.3.4 Codes and Meaning

Codes can have highly variable appearances. A code can be the rules to form words out of sounds, body movements to show an attitude or emotion, or even something as general as the clothes people wear. Thus, codes are not restricted to human language-dependent symbol manipulating capabilities, nor to the human regime at all – as has already been indicated with the animal example above (at the end of Section 2.3.1 when discussing behavioral codes as a result of *ritualization* processes). In fact, during the course of evolution, nature has evolved bottom-up, without a designer, several *sign/symbol systems* at various levels of complexity [Cla04, Car95].

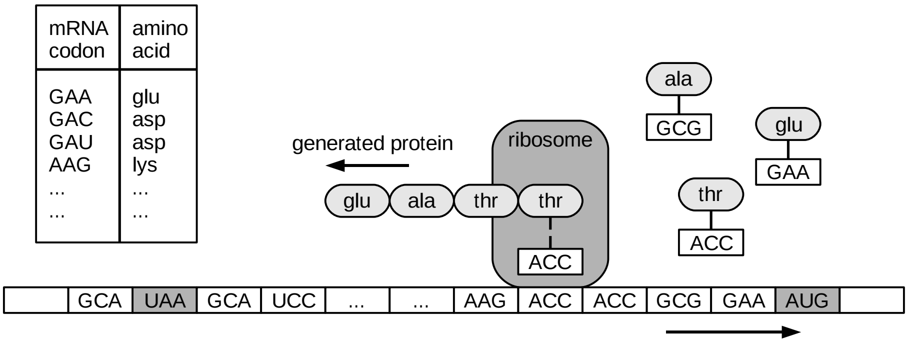
A basic sign system in biology is the genetic code (Figure 2.4). It shall be shortly discussed because it is a striking example of the possibility of an *emergent material symbol system* [Roc01]

[3] In Chapter 3, pragmatic relationships will appear in the form of emotional mechanisms.

[4] Examples of feedback mechanisms are non-linear terms in a differential equation of a self-organizing dynamical system, or a fitness function in the case of an evolutionary process.

that possesses real informational power showing all the dimensions of the semiotic triad [VTR92]. It also demonstrates that emergent, arbitrary coding relations can already exist on the biochemical level.

The idea is that simple components are connected in such a way that they show global, systematic properties which exhibit symbol-like characteristics. The simple components (acting as building blocks of the symbol system) are not just formal concepts, but real, materially implemented structures (for instance, molecules). Their dynamic behavior is governed by the laws that apply to the specific material implementation. In the case of the genetic code, sequences of DNA triplets are mapped to sequences of amino acids. Because proteins are determined by their characteristic amino acid sequence (they are produced by folding sequential amino acid chains into a three-dimensional structure), DNA molecules can be viewed as *encoded instructions of how to build proteins*. This gives the semantic dimension of the genetic code: The meaning of a specific DNA sequence is the protein it results in when decoded. The process of decoding only functions when both types of molecules, DNA and amino acids, are embedded in the cell metabolism, that is, in thousands of enzymatic regulations in a complex chemical network. This can be viewed as the process of interpretation of the code. *Only within the chemical network of the cell, DNA triplets can be treated as symbolic units that code for amino acids.* A syntactic dimension is given via copying, mutation, editing, and recombination processes. In this role, the encoded descriptions are manipulated, changed, and copied *without any recourse to their content* or meaning [Roc01, p. 13]. The diverse 'syntactic' operations serve to disseminate existing or produce new representations (via variations).



mRNA codon	amino acid
GAA	glu
GAC	asp
GAU	asp
AAG	lys
...	...
...	...

Figure 2.4: The genetic code maps DNA triplets onto amino acids. By a reading out and construction process, supported by ribosomes and transfer-RNAs, the coded sequence of triplets leads to the production of amino acid chain molecules (that is, proteins). This is referred to as *translation* of the code. There are other operations on the DNA sequence (not depicted) that copy and manipulate (parts of) the sequence, thereby contributing to the creation of novelty.

An important aspect of the genetic code is the *inertness* of the DNA molecule [Roc01, p. 5 and 14]. This means that it is not constantly involved in the dynamics of the cell, but rather has to be accessed explicitly. Thus, it really acts as a storage medium that can be used at any time, a sort of localized, explicitly accessible memory. Although the genetic code operates at the molecular level where the dynamics of the building blocks is given by chemical-physical laws, there is clearly an informational level where the components act as symbols that are – at least to a great extent – independent from their physical implementation. DNA triplets are like the 'letters' of an alphabet. They can be manipulated by syntactic operations and combined to new symbols ('words'). This is a potentially infinite process whereby ever more complex proteins are encoded. Here, a second order pragmatic relationship comes into play. The newly encoded proteins, i.e. the products of the symbol system, have to prove that they are *of value* for the phenotype, that is, for the organism as a whole. This is done by natural *selection* which is a feedback evaluation mechanism between an organism – its entire set of genetic representations – and the environment. This process could be referred to as genetic learning. Later in this work, emotions will be discussed as inherently evaluative processes that provide another kind of organism/environment feedback, situated on a level higher than the genetic. Correspondingly, higher kinds of learning are supported (as illustrated in Figure 3.1).

Note that the example of the genetic code also shows that symbolic representations need not encode every detail of what they represent or stand for. What is stored are just *initial conditions* acting like switches that can entail whole chains of follow-up processes. These switches can be accessed in a controlled way, allowing the *selection of various alternatives.* Once initialized, a selected representation deterministically leads to a specific outcome whereby 'detail knowledge' is contributed from the follow-up processes.[5]

2.4 Knowledge Representations

When building an intelligent system there is always the big question how to represent information and knowledge. The idea that brains and computer models 'house' internal representations of external objects and events is widespread, although not undisputed (for an overview of the discussion see e.g. [Cla97, pp. 143–175]). The sheer possibility of such representations hinges on the fact that the world, although only predictable in a limited way, is still full of regularities. This is almost trivial in its generality and omnipresence. Without such stable, repeating conditions, it would not have been possible for more complex forms of material organization – such as atoms or living beings – to evolve in the first place.

Regularities appearing in different physical instances can be related or mapped to each other, thereby allowing for the transfer of information from one medium to another (its 're-presentation'). Intuitively, the more an organism is able to incorporate knowledge about the correlations of its environment, the more it can predict what will happen next.

Over the years, as has been already indicated in the discussion of the history and paradigms of artificial intelligence and cognitive science, rather different ideas about the look of such internal knowledge structures (both for natural and artificial organisms and systems) have been formulated. The views span from symbolic representations through connectionist representations to the rejection of dedicated representation structures at all (for references see [Cla97, p. 148] or [Roc01, p. 2]).

In [Roc01, p. 2], the author identifies two extreme positions researchers can have on the issue of internal representation in cognitive systems in particular (and biological systems in general):

Symbolists – For symbolists – especially in the classical AI framework – cognition is identical with information processing ('computations') (see e.g. [Pyl84]). Only the informational value of a representation counts (see Section 2.2). Speed, timing, and other characteristics related to the *specific kind* or *matter* of representation are considered to be irrelevant. In the

[5] Note that, in the case of *material* symbol systems (in contrast to purely logical ones) part of the knowledge is always contained in the laws that govern the dynamics of the information-storing building blocks, that is f.i., the molecules in the case of the genetic code.

extreme case, cognitive states are viewed as computational states, and computational states are viewed as completely implementation independent (see e.g. [Har02, pp. 297–298] for a discussion). Thus, according to this view, the essential nature of mind is implementation independent, which results in the assumption that mind can be equally instantiated in different embodiments. Research is devoted to rules according to which symbols (standing in lieu of the actual aspects) are manipulated, and not to the question how *'representations come to represent the environment to begin with, or how symbols come to have meaning, or in the end, how matter becomes mind.'* [Roc01, p. 2].

Dynamicists – For dynamicists, cognition is not viewed as symbol manipulation but dynamic, distributed patterns of activity in a network with many interconnected nodes. Mind – which according to this view cannot be separated from its substrate – should be studied by dynamical systems theory alone. For radical dynamicists, the problem of symbol reference does not exist because there are no information-processing constructs qualifying as symbols that would need to be coordinated with the world.

Below, first there will be a discussion of (eventual) differences between symbolic and connectionist types of knowledge representations. The issue is not independent from how *the nature of symbols* is conceptualized in the first place. This comprises the question of how to assign meaning to a set of given symbols, but also the problem of how material symbol systems can be established (inclusively an emergent syntax, and, in particular, in an evolutionary way without a designer). From here, it follows a discussion of this second problem, the origin of symbol systems.

2.4.1 Symbolic versus Connectionist Knowledge Representations

Symbols are chunks of information. They are discretized, local pieces of information that can be *explicitly* accessed and manipulated. Compared to this, connectionist representations are the result of dynamic processes within a networked structure and not the result of a combination of a variety of well-understood components. As they usually cannot be decomposed into sequences of simpler and simpler components, they can only be *implicitly* addressed – or better *initiated* – as a whole by applying the respective inputs. To illustrate this point, consider a pile of unlabeled CDs. The pieces of music on them might fall into different categories, like jazz, punk, or opera (whereby the categories most probably will not be disjoint). The classification of the music depends on musical patterns and might be achieved by an implicit solution, e.g. a neural network. If now labels are assigned to the different pieces of music according to the implicitly derived classification, completely new possibilities arise of addressing whole sets of pieces of music without having again *to introspect* the contents of each piece.[6] I will refer to the labeled representation of the categorized pieces of music as more explicitly accessible as the neural-net based representation. Also, I will ascribe it a higher degree of implementation independence. Note that the introduction of labels alone does not result in a *symbol system.* What is additionally required is a syntax that allows to *systematically assign* meaning to symbol tokens (labels) as well as to composites of such tokens, to composites of composites, and so on.

[6] Note that the labels need not be realized as completely abstract symbol tokens. They can also be realized as 'inert arrangements of dynamical components'. Then, *'symbol manipulation would be governed not just by the arbitrary shapes of the symbol tokens, but by the nonarbitrary shapes of the icons and category invariants in which they are grounded.'* [Har90, p. 335]

The distinction between 'implicit' and 'explicit knowledge' is also standard in cognitive neuroscience when describing the different types of memory systems humans possess [Tul85, Sch97], and the different learning processes involved in skill acquisition [RKLC80].

The question that directly arises out of the existence of implicit and explicit knowledge is how they develop alongside one another. In this respect one can distinguish between top-down models – starting with explicit knowledge and analyzing how it can be related with, or turned into, implicit knowledge – and bottom-up models – starting with implicit knowledge and trying to turn it into explicit knowledge.

The top-down versus bottom-up direction is also of relevance for the understanding of the relationship between the syntax and semantics of representations.

The most severe problem of traditional symbolic AI is that all the knowledge an intelligent system has about its environment has to be built-in by the programmer of the system (see Section 2.1.1). Symbols in the classical realm are entities without internal structure and history, defined in a purely syntactic way by how they relate to other symbols. As they do not arise out of interactions with the environment, one cannot explain meaning with them because the assigned meaning always stays arbitrary. Thus, when starting with a formal symbol system, the question that arises is:

How to get from rule-based symbol manipulations (syntax) to meaning (semantics)? [Roc01, p. 104]

Distributed representations are suited to automatically capture implicit knowledge. The problem with distributed representations is that they lack syntactic manipulability and systematicity [Har90, pp. 337–338] (see also Section 2.1.2). They cannot be recombined easily in different ways. Thus, when starting with a connectionist system, the question that arises is:

How to get from a direct, dynamically given meaning (semantics) to rule-based symbol manipulations (syntax)?

An answer to the above question has to include a *decoupling* of the representations from the dynamic processes that produce them as this is the prerequisite to step beyond and exert control on the constraints that limit the own dynamics [Jua99]. This is exactly what biological organisms have managed to achieve during the course of evolution in various degrees. Although they are subject to the universal physical laws of matter, living beings have gained *control over context-dependent, selective aspects* that have significance for their individual life in a local environment. As illustration, consider the following examples:

- Instead of always reacting immediately, organisms have evolved the ability to take in information and use it later.
- Organisms can make reactions dependent not only on particular sensed external stimuli but also on internally sensed states of need.
- Organisms can trigger more than one reaction and use stored information to influence the choice between them.
- Organisms can make generalizations, infer information from other information, and anticipate what another organism is likely to do next.
- etc.

Especially the capability *to explore branching sets of possibilities* – a decisive element of all higher forms of cognition – requires descriptions that can be accessed explicitly (see e.g. [Roc95, pp. 5 and 10]). However, those descriptions should be thought of not as complete representations capable of producing universal specification, but rather as material representations that produce matter-specific dynamic configurations (compare with the example of the genetic code in Section 2.3.4; it can produce whatsoever proteins but nothing else). Still, these descriptions are informational and symbolic, not dynamic (as they are sufficiently decoupled from the ongoing dynamics). What distinguishes them from the purely syntactic symbols of classical AI is that their semantics is grounded in the self-organizing dynamics of the system and its emergent properties. But also their syntax is such an emergent property.

2.4.2 Origin of Symbol Systems: Grounding Meaning *and* Syntax

A big question when trying to understand the mind and when artificially designing a cognitive system is how to get from dynamically determined physical processes to symbolic pieces of information that obey syntactic rules (compare with Figure 2.5). This is the question of *the origin of symbol systems*.

The discussion of semiotics has indicated the necessary ingredients to answer the question. Semiotics leads one to think of symbols not simply as abstract memory tokens which are defined from the outside and which can be arbitrarily syntactically manipulated, but as materially grounded, functional tools used in the situated and context-specific interactions between agents and their environments. The important thing is not to forget any of the three aspects that make up a semiotic relationship, neither syntax, nor semantics, nor pragmatics.

Especially, pragmatics must not be neglected as it has been done in the beginning of AI, be it in the symbolic or in the connectionist approach. In both approaches, information is just processed in a uni-directional way, without appropriate feedback or evaluation. *Evaluational mechanisms* (belonging to pragmatics), however, are the integrating factor between syntax and semantics. Without them, there are no closed control loops – the most important prerequisite for systems able to adapt to non-static environments: Only feedback allows for the necessary adjustments between a system and its environment such that the system can mirror or represent the changing environment in an adaptive fashion. One of the big merits of embodied cognitive science is that it has recognized the importance the evaluational aspect has for any intelligent system. Consequently, this aspect has been put at the center of embodied cognitive science by focusing on a system's interactions with its environment. However, often proponents of the approach have overlooked or even denied the need to get from non-symbolic representations to symbolic representations, a step which seems to be necessary at least for some higher-level cognitive abilities (for instance such that require extensive counter-factual reasoning). The exclusion may be to some extent related to the difficulty of such a step.

In his famous paper '*The symbol grounding problem*', S. Harnad sketches a way of how to make semantic interpretations of a formal symbol system *intrinsic* to the system [Har90]. He stresses that symbolic representations have to be grounded bottom-up and identifies two kinds of non-symbolic representations that act as constituent parts of symbolic representations:

Iconic Representations – These are *'analog copies of the sensory projection [of objects or events] preserving its 'shape' faithfully'* [Har90, p. 342].

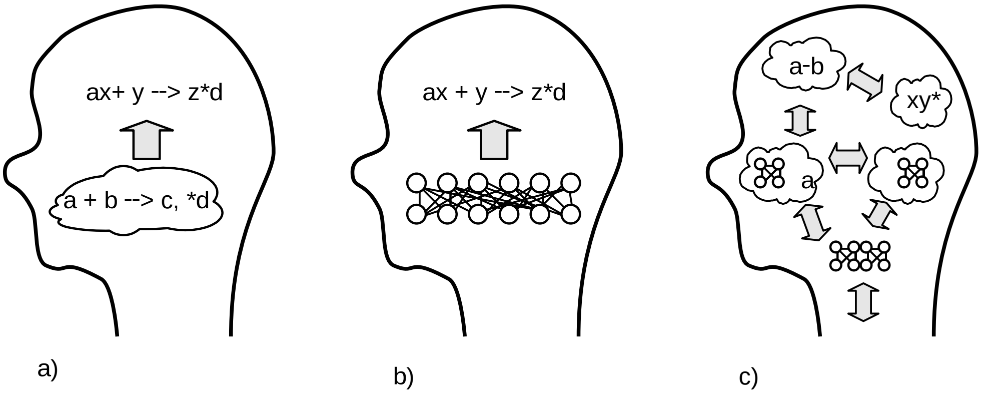


Figure 2.5: Possible relations of symbolic thought and underlying representations (inspired by [TS96]). Figure a) depicts the traditional computational view where there are at all levels in the brain just distinct, mathematical-like symbols that are manipulated according to syntactic rules. Figure b) depicts the view that at the lower levels there are dynamically interacting networks whose global stabilities then get a symbolic redescription at the higher levels. What still can be criticized on this picture is the unidirectional flow of information from the bottom to the top. Figure c) depicts a view where there are also symbols emerging out of dynamically interacting networks, but here there are many heterogenous systems influencing each other in different directions (bottom-up, top-down, and within the same level.

Categorical Representations – These are icons that have been *'selectively filtered to preserve only some of the features of the shape of the sensory projection: those that reliably distinguish members from non-members of a category'* [Har90, p. 342]. These representations are the result of a *selection* to the *invariant features* of the category performed by *'learned and innate feature detectors'* [Har90, p. 335].

Both kinds of representations are connected to the objects they pick out by a causal (and not predefined formal) relationship. They are a result of behavioral interactions between the objects and the system.

However, although there has to be a grounding of symbols in interactions, once a genuine symbolic level has established itself, including a syntax whose interpretation is mediated by the symbol system as a whole, not every single symbol must be grounded bottom-up. After all, it is the essence of syntactic rules to allow the production of new symbols without direct reference to their physical grounding. In such cases, the meaning or 'aboutness' of symbols is only indirectly grounded. In fact, in a fully established symbol system, meaning is as much given by the syntactically supported formal constraints of the symbolic level than by bottom-up links (compare with Figure 2.6).

The above point is in particular stressed by T. Deacon in [Dea98]. After starting with an account on how symbols can be grounded bottom-up, he elaborates on the top-down determined aspects of meaning in symbolic thinking as it is performed by humans. He examines how human symbolic thinking is connected with human language, and why human language makes human intelligence so particular among all other species (as f.i. only humans write poems or perform mathematics). Deacon starts his analysis with C. S. Peirce's distinction between iconic, indexical, and symbolic relationships as the three different forms of referential associations possible between a sign token and the physical object represented.

Icons – Icons are mediated by some similarity between sign and object, indicating some kind of *resemblance*. For example, pictures of objects are iconic to what they depict.

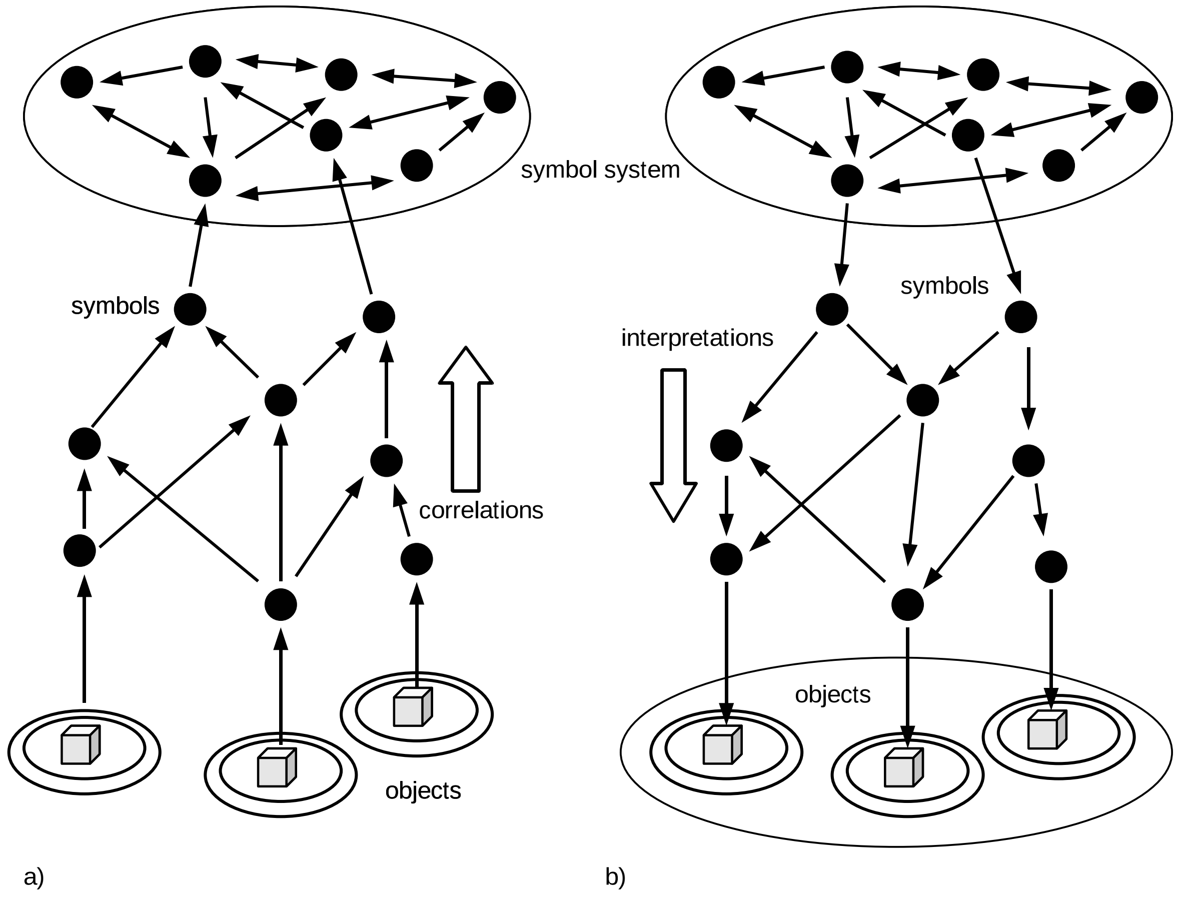


Figure 2.6: The relation between a) bottom-up symbol grounding, and b) top-down symbol attachment. Case a) illustrates that first, based on object-stimulus correlations, individual indices are learned, second, more and more systematic relationships between index tokens (signs/symbols) are established, leading to the creation of higher-order signs/symbols, and, finally, to a whole system of symbolic relationships. Case b) shows that, once, a real system of symbolic references is represented, there is a shift in semantic strategy insofar as the meaning assigned to objects is now more top-down determined by the formal (syntactic) relationships on the symbol system level than by bottom-up object-stimulus correlations. Thus, objects in the world are mainly interpreted according to a complex symbolic structure ('world model'). Perceptual stimuli mainly serve as detail information to reduce possible ambiguities of the interpretation processes.

Indices – Indices are mediated by some kind of physical, temporal connection between sign and object (*correlation*). For example, a disagreeable odor might indicate the presence of a skunk, a thermometer indicates the temperature.

Symbols – Symbols are mediated by some formal or merely agreed-upon link irrespective of any physical characteristics of either sign or object (*convention or code*).

See Figure 2.7 for an illustration of the three different kinds of relationships. No particular object is intrinsically an icon, an index, or a symbol, they can only be *interpreted* as icon, index, or symbol. Reference is hierarchic. Different modes of reference can be understood in terms of *levels of interpretation.* Complex relationships are analyzable to simpler forms. A sign can be interpreted on an iconic level, it can be interpreted indexically, or it can be interpreted symbolically – of

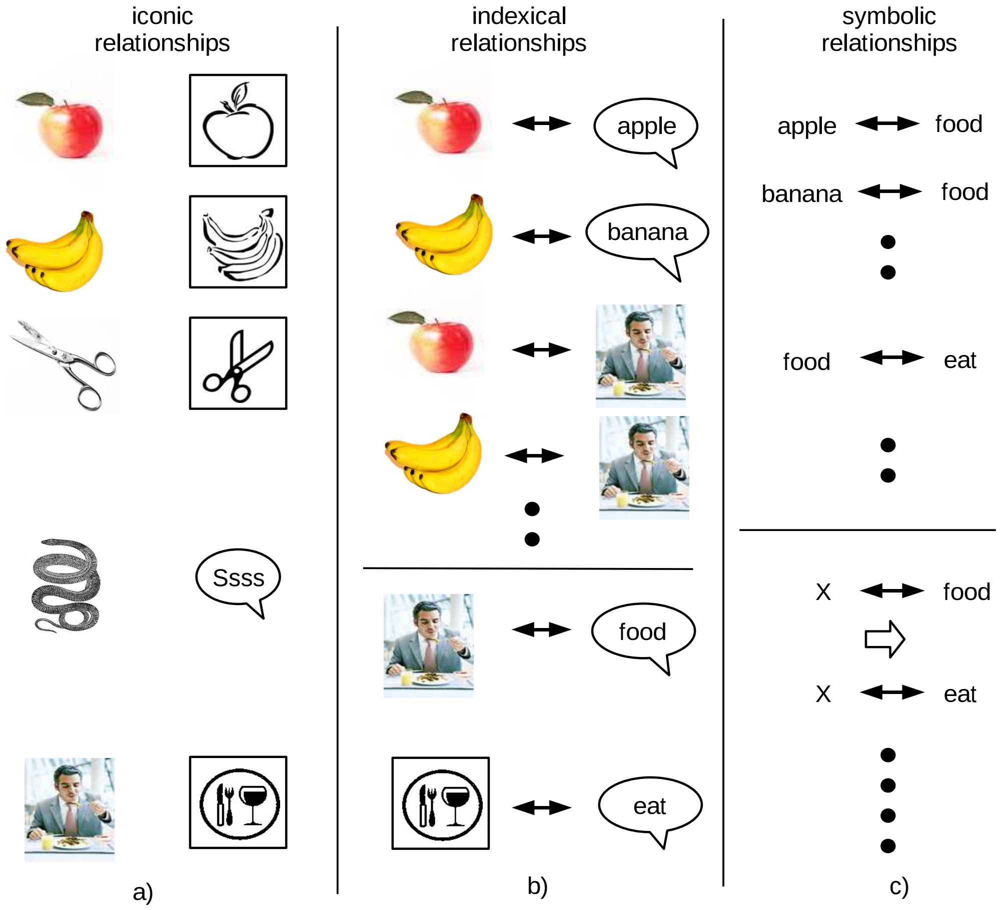


Figure 2.7: Symbolic relationships are based on indexical ones which are based on iconic ones. a) Iconic relations arise out of similarity (whether optic, acoustic, etc.). b) Indexical ones are due to correlations as indicated by a double-sided arrow: The hearing of the (arbitrary) token 'apple' can indicate a real apple if token and apple coincide. A real apple (always) indicates the possibility of eating. If the occurrence of a group of tokens like 'apple', 'banana', etc. repeatedly correlates with one and the same thing, e.g. eating, this represents a higher order kind of indexical relationship, leading also to a higher-order kind of categorization that summarizes the individual indices under a new token (e.g. 'food'). c) On the level of symbolic relations, inferences can be made completely independent from the actual occurrence of real-world correlations.

course, given the existence of each of the respective levels of association. Symbolic relationships are the most complex ones, composed of indexical relationships between sets of indices. Indexical relationships are composed of iconic relationships between sets of icons, and iconic relationships are the ones on which this whole building of semiotic reductionism is built.

Deacon argues that there is a threshold between indexical and symbolic associations [Dea98, pp. 79–101]. An indexical association is established, for example, by the repeated correlation between the smelling of smoke and the presence of flames. Learning to pair a sound or a typed string with an object in the world does, in Deacon's terms, not constitute a symbolic association (only an indexical one). A rat may be trained to correlate the hearing of the word 'food' with the dropping of food into a tray. It gets conditioned on the stimulus but it does not understand the word 'food'. Understanding the symbolic meaning of a word requires the ability to correctly use it out of the learned context which, in turn, requires finding the common features of the word, the ones which are independent from the respective context. This is similar to the categorization process addressed by S. Harnad. Still, the transfer of an association from one context to another is necessary but not sufficient to make a symbolic relationship. In contrast to indexical relationships which break down when the constituting correlation is broken, symbolic relationships still can endure. This is because *symbolic associations are not independent from each other.* Words not only point to objects, they also point to each other. Thus, what determines the pairing between a symbol like a word and some object or event is not the probability of co-occurrence, but rather a complex function of the relationship that the symbol has to other symbols. For a symbolic interpretation of the world, the system of higher-order relationships on the symbolic level is usually more decisive than the bottom-up correlations between perceived stimuli (compare with Figure 2.6). Thus, apart from stimuli generalizations (which give rise to the transfer of indexical relationships), there are also generalizations on the symbolic level – and in fact many more of them. Because the abstract symbols on the symbolic level do not only encode objects but also *ways in which objects can be related*, new symbols can easily be incorporated and combined with others. Thus, the way symbolic relationships work – and also their power – essentially derives from *combinatorial* possibilities and impossibilities, in other words, from a rule-based syntax.

2.5 Summary

Deacon's analysis of the symbolic, language-based thinking of humans supports the view that the existence of some kind of symbolic world model is necessary for intelligence, at least for its higher forms, including for instance counter-factual reasoning. On the other hand, the necessity of large world models has been challenged first by R. Brooks [Bro91b] and later on by other embodied cognitive scientists as well as, more recently, by some 'dynamicists' who deny the existence and necessity of any form of symbolic representation within the mind at all (for a discussion and references see [Cla97, pp. 143–175]).

I have tried to outline that there need not be a contradiction, that minds may be embodied and embedded and still depend crucially on brains which represent and compute. The brain (as a part of the body) certainly is a physical, dynamical system. On the other hand, cognition certainly is a kind of information processing. In my opinion, a first step towards a more profound understanding of the relationship between the hardware 'brain' and the mental states which arise out of its processing is to turn the focus of analysis from *information processing* to the *creation and transfer of meaning* whereby meaning should be understood as inherently systemic property.[7]

Concepts of (bio-)semiotics have been used to ground the emergence of meaning in a biological framework. In nature, an important element in the (emergent) process of meaning creation is the reduction of redundant information by the finding of higher-order regularities in the mess of associations that are constantly perceived by the sensors. This categorizational process (which is a kind of abstraction), however, is not achieved by perception alone. Whatever a living being senses – food, predators, sexual reproduction, etc. – means something to it with respect to some potential actions. Von Uexkuell, a pioneer in biosemiotics, has put this as follows: *'Every action [..] that consists of perception and operation imprints its meaning on the meaningless object and thereby makes it into a subject-related meaning-carrier in the respective Umwelt'* [J.v82]. Thus, in

[7] Note that understanding meaning as systemic property in a hierarchy of part-whole relationships implies that meaning is not restricted to the human level, and might even be extendable beyond the level of biological organisms.

nature, ‘aboutness’, values, and meaning arise out of the combination of perceptions with actions, executed in a feedback giving environment by embodied organisms.[8]

In the course of semiotic processes, evolution has brought about mental states, intentionality, and a whole psychic world. Although humans are now characterized by the great amount by which they can uncouple mental states from bodily action (key word ‘inhibition’), one must not forget that ’mental aboutness’ grew out of ’bodily aboutness’. This is a central aspect of the proposed approach based on a combined neurological/psychoanalytic view.

To summarize, in designing artificial systems, the best available guidelines seem to come from biological systems, where embodiment in an environment and actions play a crucial role. It is therefore sensible to instantiate such embodiment in artificial systems as well, at least functionally. Additionally, the following design principles can be extracted from the argumentation given in this chapter:

- The grounding of symbols is a hierarchical process. For building a system of symbolic relationships, the formation of iconic representations (‘images’) and categorizational mechanisms that filter out invariant features are constitutive elements (see 2.4.2 and e.g. [Har90, p. 342], or [Dea98, pp. 69–79]).

- The meanings of symbols can be naturally grounded bottom-up in sensory values [Har90, p. 343]. However, not necessarily every individual symbol must be grounded bottom-up (see [SC05] and [Dea98, pp. 79–101]). Once, a full net of symbolic relationships is established, there is partly a reversal in the direction of how meaning is determined (see e.g. [Dea98, p. 87] and Figure 2.6), from largely bottom-up to largely top-down. *We see what we expect to see, that is, what fits into our already internalized categories of knowledge.* This is an insight already stressed by the Russian neuropsychologist A. Luria [Lur73, pp. 73–75].

- *Emergent* symbolic representations depend crucially on *evaluational feedback*. Their establishment is based on the ‘ritualization’ of interactions (see Section 2.3.1 and [EF94, p. 57]) leading finally to abstract codes with *syntactic* properties that support the building of new representations out of existing chunks in an open-ended manner (see Section 2.3.2 and e.g. [Roc95]).

- To implement higher forms of reasoning most probably symbolic representations – which, however, can arise out of hierarchically grounded emergent symbol systems – are needed (see e.g. [Cla05]). World models seem to be necessary, but they also have to be constructed such that they are dynamically adapted according to the system’s ongoing experiences with its environment.

[8] The decisive point is that representations start as material affairs with a ‘natural’ meaning based on causal relations that ground and ‘put forward’ meaning in a bottom-up direction. But meaning also possesses a top-down component which appears when answering the question, what a representation is good for, that is, in which way it contributes to the functioning (stability, survival) of the system as a whole. The question can be decided by looking at the feedback-mediated consequences of the outputs the system produces (based on its representations). In this sense, the meaning (or purpose) of a leg is given by the fact that it serves a person to walk.

They may forget what you said, but they will never forget how you made them feel.
Carl W. Buechner

The tendency of aggression is an innate, independent, instinctual disposition in man.. it constitutes the most powerful obstacle to culture.
Sigmund Freud

3 Embodiment, Drives, and Emotions

The most distinguishing inspiration for the proposed ARS-PA architecture are psychoanalytic conceptions and insights (presented in Chapter 5). Throughout the work, I will try to outline how psychoanalytic conceptions can be *brought into coherence* with modern neuroscience and with an evolutionary account of cognition. Last but not least, I will try to work out how they can be of value for the technical design of an intelligent autonomous system. Central to this task are considerations about the body, its needs, and the value-providing function of emotions. In this chapter, a functional analysis of these concepts is delivered.

In the previous chapter, it was proposed that the way organisms are embodied is a decisive aspect for the development of their intelligence, and that the same also applies to autonomous systems or agents. This stands in contrast to the classicist/cognitivist research program that maintains that cognition can be understood by focusing primarily on an organism's internal cognitive processes. This view of cognitive processes cuts organisms off from their body and from their interactions with the environment. Thus, the classical stance overlooks the following claims which are central for the embodied approach to cognition [VTR92, Cla97]:

Bodily form shapes actions – It is exactly the particular form of embodiment that simultaneously prescribes and constrains the manner in which a system can interact with its environment. The kinds of sensors and motors a system has – for example, eyes, ears, legs, fins, etc. – as well as the system's internal, 'bodily' states exert the major influence to determine the look of the actions a system can perform in the world.

Actions (via feedback) shape cognitions – The actions performed by a system in the world always receive feedback from the environment, and this feedback is the ultimate source to shape an organism's cognitive capabilities. More specifically, actions entail consequences which, during the course of evolution, have been related to particular *sensorimotor and emotional experiences*. The claim now is that these experiences serve as the basis for the formation of cognitive categories and concepts. Think of the well-known saying: 'If the only tool you have is a hammer, then everything looks like a nail.'

To summarize, the argument goes that the way in which an organism is embodied and embedded in its environment determines the properties of its interactions, and these interactions, in turn, lead to the development of particular cognitive capacities and determine the precise nature of those capacities.

In the first part of the current chapter, the most important design principles derived from an embodied view on cognition are outlined. One of them, the requirement to introduce *value systems*, leads to the second part of this chapter which discusses emotions as *functional concept that provides evaluations.* Evolutionary theory, and also today's neurobiology, give a hierarchical picture of emotional mechanisms. This picture starts with implicit, automatic evaluations of different behavioral alternatives (e.g. taxes, drives, instincts) and ends with conscious appraisals of emotionally afflicted situations using inner speech. Note that the psychoanalytic view on drives and emotions ('affects' in psychoanalytic parlance) is presented at the end of Chapter 5. Here I just want to draw the reader's attention to some aspects: first, Freud described human behavior as strongly influenced by (unconscious) drives and affects (that is, not just as product of rational thought), second, he acknowledged (in contrast to some of his followers) the connection of drives and affects with the body, third, he assumed that affects are *'reproductions of ancient, survival relevant events'* [Fre26, pp. 163f], and fourth, he particularly studied the subjectively experienced aspects of affects (the 'feeling' part of an emotion in today's neurobiological parlance).

3.1 Principles of Embodied Cognitive Science

Although there are slightly different conceptions of embodied cognitive science depending on the field from which the researchers come (psychology, linguistics, robotics, etc.), all of them maintain that one essential condition for cognition is embodiment. Developmental psychologist E. Thelen describes this central notion of the embodied approach as follows:

> *'To say that cognition is embodied means that it arises from bodily interactions with the world. From this point of view, cognition depends on the kinds of experiences that come from having a body with particular perceptual and motor capacities that are inseparably linked and that together form the matrix within which memory, emotion, language, and all other aspects of life are meshed.'* [TSSS01, p. 1]

The common goal of embodied cognitive research is to develop explanations that capture the manner in which body, mind, and world mutually interact and influence one another to ensure an organism's adaptive success in its environmental niche. Thus, the isolationist view of intelligence pursued by classical cognitive scientists is replaced by a *systemic view* that focuses on the relations between organisms, their actions, and the environment. Making the way the world appears to an organism relational does not imply that there is no objective, external reality and that everything is subjective. An objective, observer-independent world can still be assumed. Just the way this world is understood by an organism depends on its embodiment and its experiences.[1]

In the following, some of the most important design principles derived from embodied accounts of cognition are briefly discussed. They present important insights on the creation of artificial intelligent systems.

[1] Of course, for the organism itself, its own, subjective interpretation of the world is the one that counts. Actually, the subjective perspective is the only one an organism can have, because no organism has direct access to the things-in-themselves. In this sense, one could say that each individual has its own 'reality'.

3.1.1 System-Environment Coupling

One of the defining theoretical assumptions of embodied cognition is the importance of system-environment interactions unfolding in real time. In AI and robotics, this is reflected in a change of the prevailing role model from passive computer programs to active *agents*. Thus, the essential shift of focus is from inert computer code to entities that can *act* on or interact with their environment.[2] Agents can have several properties. Some important distinctions are for example the following [PS99, pp. 82–95]:

Situated Agents – Agents that are able to acquire information about their current situation via interactions.

Autonomous Agents – Agents that are able to control their actions as well as internal states on their own without the direct intervention of humans or other agents.

Self-sufficient Agents – Agents that are able to perform not only a single task, but a variety of them, all they need to survive in their respective environment.

Adaptive Agents – Agents that are able to sustain themselves under changing environmental conditions.

Complete Agents – Agents that are situated, embodied (see Section 3.1.2 below), autonomous, adaptive, and self-sufficient.

All this properties have not been in focus in the era of classical symbolic AI, but they immediately appear in a bionic approach. For example, self-sufficiency requires the management of internal resources, a central aspect of any organism (see Section 3.1.2 below). Adaptiveness, in turn, is a necessary consequence of self-sufficiency.

The coupling of autonomous systems with their environment through *sensors* and *actuators* gives agents the possibility to adapt to their environment. Being able to act on the environment and to observe feedback is an important source of learning, and thus a necessary prerequisite for intelligence.

Moreover, making use of the system-environment interaction minimizes the amount of *world modeling* required. According to the classical paradigm, programmers would have to guess all the conditions the autonomous system could probably encounter and then to spell out all the relevant information that is needed to generate an appropriate response. However, anticipating all this information is very difficult, and most often even impossible. Especially, large sensor spaces are only manageable by exploiting the constraints one gets from system-environment coupling. Thus, as put by R. Brooks, one can *'use the world as its own best model"* [Bro91b, p. 139]. For example, instead of solving a jigsaw puzzle by 'pure thought', one can take a candidate piece, hold it above a roughly assessed potential location, rotate it, and finally try out whether it fits in or not. This is a hybrid strategy, using a mixture of acting and thinking.

[2] From a psychoanalytic point of view, *agency* is important because it is a basic feature of subjective mental experience: 'I shall do this'. This experience of active agency is synonymous with the sense of self.

3.1.2 Embodiment

Being embodied is related to how an autonomous system or agent is situated or embedded in its environment, that is, with its capabilities to interact. Those capabilities, first depend on the kinds of sensors and actuators (motors) a system has, and second on its internal needs.

Sensors and Actuators: Interfaces between Body and Environment

A system which can see – whether using eyes or cameras – will *construct a different representation* of its surroundings than a system which can, e.g. smell. Thereby, as has already been stressed above, perception is no end in itself, it always is or has to be related to the actions the system must perform. For example, the way a useful representation has to look also depends on whether the system has legs or wheels to move around.

Note that the interaction-dependent view of representations presented above also means that embodied cognition thinks of representations as *active constructions* and not as observer-independent collections of facts that just 'mirror the world as it is'. Thereby, the process of constructing representations is controlled by feedback. In the simplest case, interaction-derived representations are formed by, or consist of, sensorimotor feedback. For example, in order to learn to grasp an object, a system can just stretch out its arm approximately in the right direction, observe how much it has missed it, adjust the applied forces, angles, etc., stretch out again, observe again, and so on, until it succeeds. All this feedback-guided activity results in a representation that is based on those environmental features that are directly relevant to the goal-directed action the system is currently performing.

Internal Needs

The representations a system builds up and the behaviors it performs do not solely depend on its sensors and motors which are in direct interaction with the environment. There is another big factor that has to be taken into account when determining the actions of an autonomous system: its *internal state*. In biology, organisms have a physical body whose essential physiological variables must be kept within certain, usually limited, ranges. The necessity to fulfill these bodily needs is the real driving force for any organism to do something at all, and the reason why its behavior gets the quality of being goal-oriented. The fundamental, bodily-related problems of life are: to provide the body with energy, to keep its inner chemical milieu within certain ranges compatible with being alive, to defend and protect the body against external causes of hurt and destruction, to reproduce the body – first, as a remedy against decay, and second, to give it the chance to keep track with an ever-changing environment. The 'solutions' nature has found to these problems (further discussed in Section 3.2) are organized in a hierarchical way, reaching from simple bodily reflexes to cognitive reflections. It follows from what has just been said that to avoid environmental influences having a negative effect on the body is the primary task of an embodied being. To do so, not only external, but also internal sensors capable of receiving the values of physiological variables are required. Additionally, the actions carried out also have to be selected with respect to how they contribute to the protection of the body's integrity and well-being.

Technical Interpretation

Although artificial autonomous systems do not have the same kind of physiological body as an organism, they still do have internal states and resources which need to be managed. Therefore, it will be required to equip autonomous systems, such as robots, not only with *external sensors and actuators*, but also with *internal sensors and actuators*, and with *internal ('bodily') needs.* Whether the body has to be a physical one or whether it can also be a virtual one – as for example in the case of software agents or simulations – in this work the following position is taken:

> **Embodiment (functional view)**: Essential requirements of embodiment are
>
> - mutual perturbation capacity (the autonomous system can influence its environment and vice versa) [QDNR99], and
> - the existence of internal resources that are not unbounded and consequently need to be managed.

3.1.3 Exploratory Behavior

One of the most basic and common activities performed by brains is the instigation of motion. Even very simple animals move around searching for food or other resources[3]. In general, mobility is a means *to explore* the environment and thus *a source of gathering experience.*

Usually, brains must make their motions (and also other activities) fast. Organisms depend on quick and fluent real-world interactions, they cannot sit back and take their time. The reason for this, as has been explained above, is the necessity to constantly protect the body's integrity by keeping it from harm and by fulfilling its internal needs. These needs are the ultimate impulse for any goal-directed action of an organism, the force driving it to perform motions and other, 'higher-level', kinds of activities, the latter being only indirectly related to the fulfillment of needs, but still. In general, the higher the cognitive level involved in an activity, the lesser and more indirect these activities can be brought into connection with the fulfillment of a bodily need. Thus, at the lower levels of behavior, actions are directly motivated by bodily needs. At the higher levels, there are also other 'sources of needs' (for example social or psychological ones).

A special form of exploration is play. In [Bat79, p. 151], the following definition of play is suggested: *'the establishment and exploration of relationship'.* This definition is opposed to the one of ritual: *'the affirmation of relationship'.* Superficially, play has no purpose. However, although there may be no direct one, there is a huge indirect one: Play is the open-ended, non-settled trying out of relationships. Thereby, a system can learn to better adapt to its environment.

3.1.4 Values

Organisms have needs, but in nature, resources are always limited. This makes it necessary for creatures to concentrate their senses – in a world filled with an overabundant amount of information – on features which are *relevant* for the fulfillment of their needs, manifesting itself in a

[3] Note that the nervous system is an 'invention' of animals. Plants which are usually rooted and therefore immobile (apart from their seeds) obviously did not need to evolve such a complex and powerful tool as a brain to ensure their survival.

selective perception of the environment. There is a similar problem on the action side. If an artificial agent shall be autonomous and situated, it has to make a decision between several possible actions by *judging what is good for reaching its current goal and what is not.* Thus, intelligence is not the problem of how to process all the potentially available information but how to pick out, in a timely manner, small but significant pieces of information. This is achieved through a *value system.* Having emerged from adaptations based on agent-environment interactions, value systems could be described as follows:

> **Value system**: *'A system where behavioral goals are defined in terms of their recognizable consequences. Value systems modulate or bias learning processes and decision making.'* [PS99, p. 499]

The above characterization of values

- connects values with goals, and
- states that values only makes sense in a systemic, feedback-providing context (because consequences are required).

Sometimes, people think of values and goals as categories legitimately used only in the human realm. But this is not the case. Every selectionist, feedback-driven process is related to the concept of value. In [EF94], the authors suggest that *values express the relationships which exist between parts of a system and the stability or 'well-being' of the system as a whole*, for example the relation of one behavioral act regarding the survival of the whole organism.

From a systemic perspective, one sees that goals are similar to needs, drives, desires, wishes, intentions, or tasks, in so far as all of these concepts have one thing in common: They constrain actions in a top-down oriented way, meaning that the system as a whole (more general, the higher levels of a system) put requirements on the behavior of its parts (that is, the lower levels of a system). For example, my intention (as a whole person) to eat a fruit from that tree over there, is the reason for my feet (which are a part of me) to move in that direction.

On the other hand, as has been discussed as central claim of the embodied view, the structure and the functional capacities of the parts of any system (in the above example, my feet) significantly shape, in a bottom-up way, the behavior of the whole (in the above example, my movements). Thus, parts and whole are intrinsically connected. They mutually construct and limit each other.

Note that all systems consist of at least two organizational levels – the level of the parts and the level of the global system. Systems with *several nested levels* are called hierarchies. Usually, these levels are of increasing complexity. As an example, take a multi-cellular organism with its cells, tissues, organs, and so forth, or a building consisting of bricks, walls, rooms, or a big company consisting of many management levels, or a text consisting of letters, syllables, words, sentences, ideas, etc.

It has already been said that the purpose of value systems is to pick out relevant, that is *significant* pieces of information. In other words, value systems contribute to the construction and transfer of meaning. In semiotic parlance, evaluations provide the pragmatic part of the semiotic relationship (compare with Section 2.3.2). They link internal and external conditions (in general semiotic terms, the referent and what it refers to) in a way that is based on the success or failure of the linkages (which may get adjusted in the process). In biology, this continuous adjustment is performed by evolution based on trial-and-error. As a result of feedback, organisms build up representations ('world models') that link internal and external states in a meaningful way. From bacteria to humans, these world models are more and more complex, consisting of more and more levels of representations put upon one another. On each of the levels, the semantics of the representational constructs has a bottom-up aspect (given by causal interrelations), but it is the top-down perspective that reveals the functional significance (the meaning proper) of a representation (the relevance of the representational construct regarding the output of the level as a whole).

From what has been said so far, it should be clear that *evaluations* are a central element for the construction of world models. Depending on the level of complexity, evaluations, however, take different forms:

- On the lower levels, the evaluational part is given by *sensorimotor feedback* which is used to control the movements of an organism, e.g. swimming, walking, reaching etc.
- On the middle levels, to achieve control that goes beyond the guidance of current movements, nature has evolved *emotional mechanisms* in the widest sense. They classify what is good and what is bad for an organism in a given situation in a relatively rough way, that is without performing an exhaustive analysis. Still, this classification helps to categorize the environment in a way meaningful for the organism. By memorizing emotionally classified experiences, an organism can build up a knowledge base with high predictive power for a variety of typical situations occurring repeatedly in the life of an organism.
- On the top levels, evaluations take the form of *cognitions* that resemble computations.

The above list also indicates that human mental experience is a result of many processes and functions, running concurrently on different levels and influencing one another. Each of the levels of the above list consists again of several levels.

In Section 3.2 below, emotions proper will be discussed as just one level of complexity in the hierarchy of *emotional mechanisms*. Biological emotions are value systems which have evolved in evolutionary history long before the more intellectual forms of judging situations entered the scene. Emotions have started to interest researchers in AI and robotics as soon as their functional relevance as value system has come into focus. Generally, a key function of emotion is to communicate simplified but high impact information. It can determine what makes a piece of information relevant in one situation and not in another. Emotions will be explored in more depth in the following part of the chapter.

3.2 Emotions as Evaluations

For a long time, emotions have been considered solely as disturbing influence to 'rational' thinking. In that time, AI researchers in their effort to create intelligent artifacts have completely neglected emotions. Instead they have exclusively focused on higher-level cognitive capabilities such as thinking and planning. They developed knowledge bases, logical inference systems, algorithms for case-based reasoning, data mining, learning, categorization, language understanding and imitation, etc. All these systems model capabilities that are certainly ingredients of intelligence. But there are some elements missing. These are related to the fact that the brain was not designed to 'run programs' for isolated, very specialized purposes. According to evolutionary theory, the brain has evolved to control our behavior in order to ensure our survival in highly demanding and constantly changing environments. Thus, perceptual and motor abilities are also essential for intelligent *behavior*, as well as emotions which, in an informal sense, can be seen as mechanisms to enhance acquired information with values.

Freud is famous for revealing the strong influence drives and emotions ('affects') have on our human behavior. However, intelligence is not a concept appearing in his work. The combination of emotions and intelligence, and in particular, the insight that emotions may be an important facet of intelligent behavior was accepted only recently. By now, several of today's neuroscientists have argued convincingly that emotions are crucially intertwined with cognitive problem solving and decision making [Dam03, Pan98, Led03]. Besides, as for example elaborated by the psychologist Frijda [Fri04], emotions are considered to be an essential part for the establishment of social behavior. In order to be able to use emotions successfully in a technical implementation, it is necessary to understand them *functionally.* To do so, it is helpful to look at the history of the phenomenon, to see how it has evolved in all its broadness. In nature, living beings have evolved a cascade of more and more complex control levels for behavior selection, all of them aimed at ensuring homeostatic balance of the body. Great parts of this hierarchy can be referred to as 'emotional mechanisms' in the widest sense. Accounts supporting this claim have been worked out by many researchers coming from very different fields. Below, some of these accounts will be described in more detail. For the psychoanalytic view, see Section 5.2.2.

The common view is that each newly emerged control level is built up of parts of the lower levels. However, the different levels are not clearly separated, but rather very much intertwined, with much back-and-forth communication going on over the linkages connecting the various levels. This makes it so hard to analyze their functionality. Moreover, as each level is crucially based on a feedback loop, it cannot be decomposed in its parts without losing some of its emergent features.

3.2.1 Account from Ethology

Feedback is essential for evaluation. In [Rie81, pp. 104–106], R. Riedl exemplifies that feedback can always be described as a loop between *experience* and *expectation* where both terms are used in a very general way. Experience refers to the input the system receives from the outside world as a response to its own actions. External perceptions are interpreted (recognized) based on given knowledge. Expectation refers to the process of building up inner knowledge representations and projecting them ahead. Outer world and inner representation have to be related to one another, ideally in an adaptive, dynamic way. The better this relation is achieved, the more the organism (or autonomous system) will be to sustain itself even in highly demanding environments. However, the technical design and implementation of such a flexible association mechanism is quite a difficult task. When looking at nature, one can find a hierarchic picture of 'solutions' to the problem (Figure 3.1).

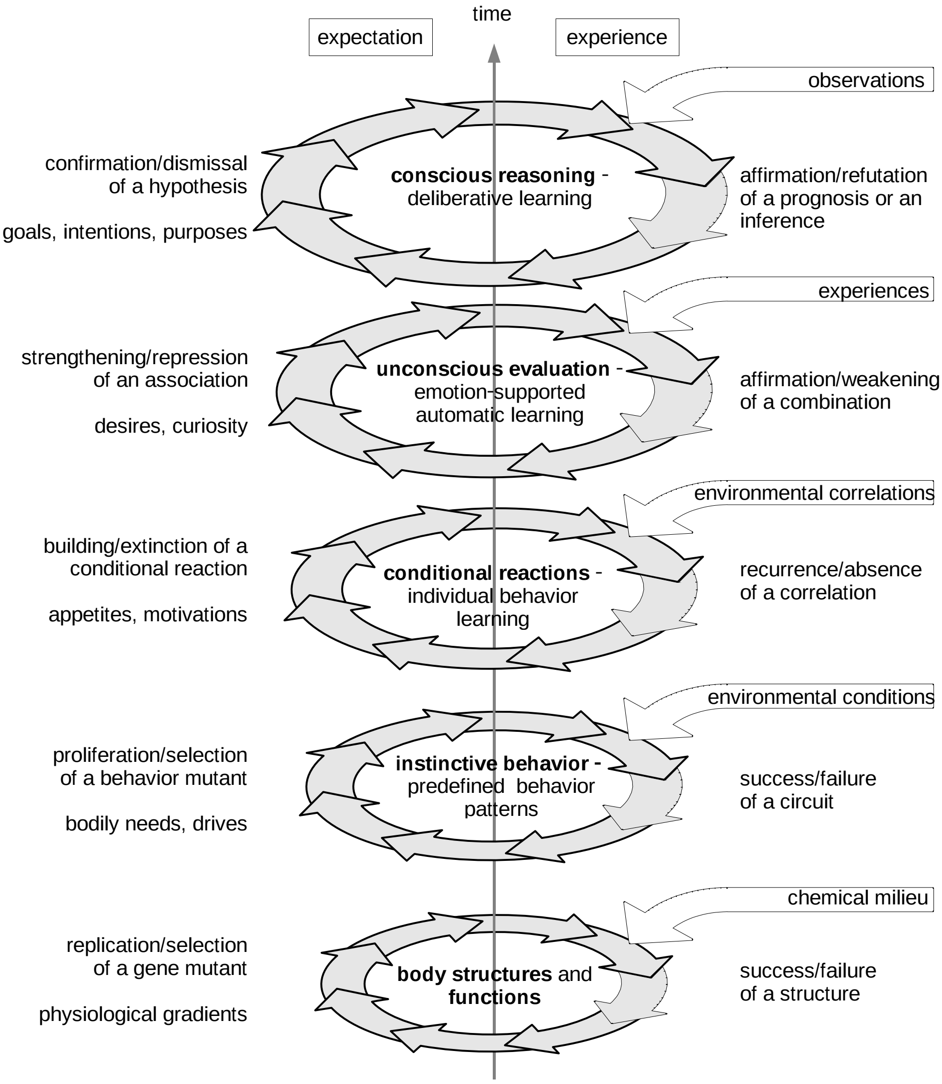

time
expectation
experience
observations
confirmation/dismissal of a hypothesis
goals, intentions, purposes
conscious reasoning - deliberative learning
affirmation/refutation of a prognosis or an inference
experiences
strengthening/repression of an association
desires, curiosity
unconscious evaluation - emotion-supported automatic learning
affirmation/weakening of a combination
environmental correlations
building/extinction of a conditional reaction
appetites, motivations
conditional reactions - individual behavior learning
recurrence/absence of a correlation
environmental conditions
proliferation/selection of a behavior mutant
bodily needs, drives
instinctive behavior - predefined behavior patterns
success/failure of a circuit
chemical milieu
replication/selection of a gene mutant
physiological gradients
body structures and **functions**
success/failure of a structure

Figure 3.1: The evolution of learning algorithms (after [Rie81, p. 106 and p. 178]). On the right, there are the parts related to experience, and on the left, the parts related to expectation. Experience is the process of deducing meaning out of perception. Expectation is the anticipatory part of inducing a connection between (future) perception and meaning. From layer to layer, the forms and contents of experience and expectation vary, but the principle of the algorithm stays the same. The arrows coming from the right indicate the kind of information coming from the outside (i.e. the environment) into the system. The text on the right side identifies the kinds, and the effects and contents of experience of each layer. The text on the left side names the motivations, and the contents and effects of expectation of each layer.

In the most simple case, perception and action are just connected by hard-wired rules. These rules can be viewed as external regularities having been internalized during evolution by the interplay of mutation and selection. With time, the incorporated processes between perception and action have become more and more complex. For example, first there were solely unconditioned reactions, later on, by opening simple control loops, conditioned reactions came into being, the latter possessing a higher degree of flexibility and context-sensitivity.

Mechanisms such as reflexes, drives to initiate active behavior, instincts to perform behavioral patterns, conditioned reactions mediated by lust and pain, or emotions-proper to react to a variety of dangerous or advantageous situations in the individual as well as social life of an organism, are all elements of a list of distinguished solutions having been accomplished by evolution to tackle the problem of homeostasis.

In accordance with M. Minsky's writings in his latest book [Min06], one could term them as 'ways to think', just as the more 'intellectual sorts of thinking' which are on top of the behavioral control pyramid. Examples of this traditional category of intelligence are conscious reasoning, planning, and decision making. On the expectation side, these processes are guided by hypothesis-producing mechanisms such as ideas, theories, and even general world view.

3.2.2 Account from Evolutionary Psychology

In [Fre26, pp. 163f], Freud tried to explain the somatic correlates of emotions with Darwinian concepts (see Section 5.2.2). The psychologist W. Mc Dougall ties up to this view when drafting his theory of emotions evolutionary-rooted [McD69]. Mc Dougall generally believes in an inherited basis of human thinking, feeling, and acting. The starting point of his considerations are instincts, which he believes to provide the motivation for all action and thought. According to him, each *instinct mechanism* is comprised of three aspects, being

- the perception of an instinct-relevant *stimulus*, be it an object or an event, whereby the stimulus can be innate or learned, and
- an *emotional excitement* as central component, producing visceral changes to support the third aspect, namely
- an instinct-specific *impulse* to execute a certain behavioral pattern ('action tendency').

Thus, in Mc Dougall's theory there are emotions occurring during an instinct process. He terms these emotions *primary emotions*. For him, the purpose of an emotion is to signal the conscious subject its own kind of excitement and action impulse [McD69, p.326]. It follows that an emotion helps an organism to recognize the state which it is in and the action tendency by which it is driven, thereby enabling it to regulate both of them to some extent. Concerning the possibility to regulate instincts, Mc Dougall thinks, that the part of the instinct-eliciting stimuli as well as the instinct actions can be modified (by experience), whereas the central emotional component of an instinct process, and also the action impulse itself are not modifiable [McD60, pp. 30–33].

Mc Dougall also introduces *secondary emotions*. He distinguishes between these which are *combinations* of primary emotions (see e.g. [McD60, pp. 140–142]), and these which are *derivations of primary emotions*. The latter, derived emotions, are viewed as degrees or mixtures of pleasure and unpleasure [McD28]. They depend on an assessment of the probability of success or failure of strivings and desires [McD69, p. 350]. Examples are, among others, hope, desperation, disappointment, or regret. Their purpose is to strengthen or weaken currently active behavioral impulses. Note that derived emotions depend on specific cognitive capacities (e.g. mental representations of the goals of instinctive action impulses, an assessment of the probability of future events). Therefore, Mc Dougall speculates that they are probably restricted to humans and higher animals [McD28].

3.2.3 Account from Neurobiology

In [Dam99, pp. 39–40], the contemporary neuroscientist A. Damasio bases his analysis of brain and mind upon three assumptions. He considers these assumptions to be fundamental for the understanding of the working of brain and mind and wonders why they have been neglected only until recently. The three assumptions are:

- Evolutionary perspective
- Integrative view of brain and body
- Concept of homeostasis

Damasio further argues that all the above three aspects are related to the concept of emotion, understood in its broadest sense. This broad understanding of the term emotion, referred to as *emotional mechanism* in the following, can be defined as follows:

> **Emotional Mechanism**: Any mechanism that serves the homeostasis of the body, that is, the regulation of its integrity and well-being.

Damasio gives a rough overview of the various *levels of homeostatic control* [Dam03, pp. 31–37], ordered approximately as they arose in the course of evolution (Figure 3.2). The idea is that simpler regulative reactions reappear as parts of the more complex reactions, leading to a strong entanglement of the different levels [Dam03, pp. 37–38]. Between the various levels, there are bottom-up as well as top-down relations. As has been said, all the mechanisms listed in this hierarchy aim at providing the organism with reactions that support its self-preservation. Note that by embedding emotions-proper in this hierarchy, Damasio emphasizes the biological (evolutionary, functional, adaptive) foundation of emotions while, at the same time, he also introduces the social (learned, constructed, cognitive) aspects of emotions.

Emotions and Feelings

Damasio makes a distinction between emotions and feelings, although conceding that, in humans, emotions are usually immediately followed by feelings such that the idea of different levels of emotional quality may not be so striking at first glance[Dam03, p. 29]. He argues that emotions are prior to feelings simply because evolution has brought them about first. As has been just described, emotions are viewed as internally directed sensory modalities with a long evolutionary history, to a great extent biologically determined, and dependent on specific, innate brain structures and processes. Emotions are usually initiated automatically, i.e. unconsciously. Feelings, for Damasio, are the conscious part of emotions. He writes: *'[..] you can observe a feeling in yourself*

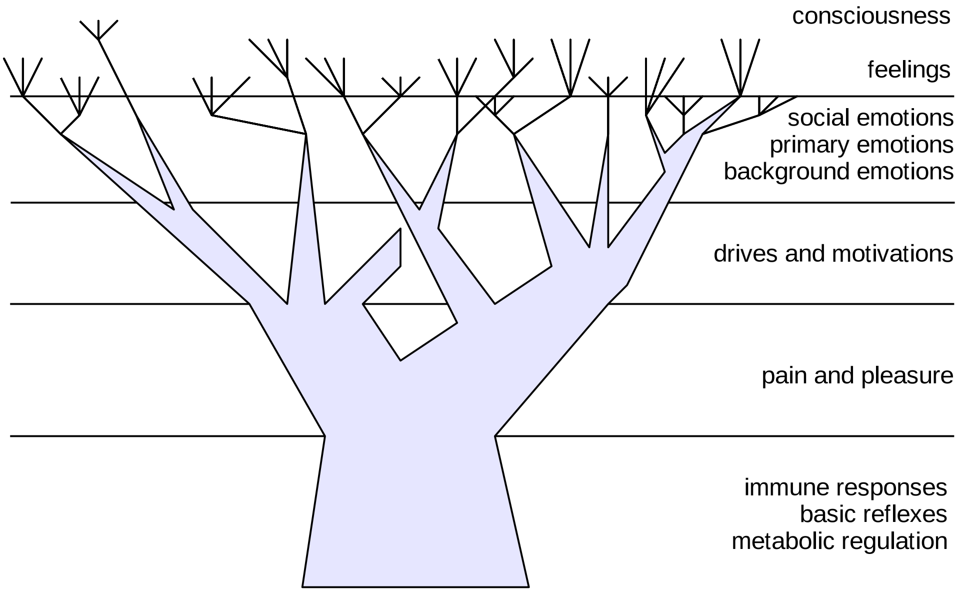


Figure 3.2: Damasio's tree of emotional mechanisms to maintain homeostasis (after [Dam03, p. 40]). Throughout the tree, the nesting principle applies. For example, social emotions incorporate mechanisms that are part of background and primary mechanisms.

when, as a conscious being, you perceive your own emotional states' and *'The term feeling should be reserved for private, mental experience of an emotion, while the term emotion should be used to designate the collection of responses, many of which are publicly observable.'* [Dam99, p. 42]

Although declaring feelings to belong to the conscious and thus subjective realm, Damasio still makes an attempt to describe how the brain might produce them. He characterizes feelings as collections of second-order processes based on the first-order processes that make up the corresponding emotion. Note that according to this, any feeling first and essentially also is an emotion:

Emotions – When an emotionally competent stimulus is sensed it gets (subconsciously) 'appraised' in particular circuits of the brain resulting in an emotional state that consists of a bundle of inwardly and outwardly directed responses (see [Dam99, pp. 59–70]). Inwardly, this bundle includes physiological changes (like the release of specific hormones, the bumping of more blood to the muscles, etc.). Outwardly, emotion manifests itself in the execution of specific behaviors (like shouting, fighting, fleeing, etc.), and in a physical expression of the emotional state (like facial expressions, baring the teeth, flushing, etc.) Together, these processes make up the *emotional apparatus* which enables organisms to react effectively to stereotypical objects or situations.

Feelings – At least in human brains, evolution has brought forth a higher-order representation of the above processes. Damasio claims that apart from the representation of the emotion generating object or event, there is also (somewhere else in the brain) a representation of the bodily changes occurring during the emotional state, and finally, a second-order representation, that combines these two first-order representations [Dam99, pp. 168–171]. The second-order representation is argued to be the precondition which lets us (subjectively) experience, that is *feel*, our bodily changes *in connection* with the occurring external changes.

Feelings are hence mental perceptions which are first and foremost about the body. However, the body is not the only source of emotions (and thus feelings) any more because the representation of the body can also act as a source, in this case an emotional state would have a virtual (or mental) source.

Damasio argues that feelings are the basis for conscious thoughts, that is, the starting point of self-awareness [Dam99, p. 172]. Note that originally feelings are non-verbal experiences, but humans can integrate them into language-based thinking [Dam99, p. 185]. By interacting with memory, imagination, and thoughts, feelings contribute to the production of non-stereotypical reactions. With 'sufficient integration of the now, the past, and the anticipated future' feelings can enable humans to achieve a more effective plan for survival and well-being [Dam03, p. 178].

Classification of Emotions

Damasio provisionally divides emotions into the following three classes [Dam03, pp. 43–46]:

Background Emotions – These emotions are not so pronounced. They are the result of the constant monitoring within the brain of what goes on in the body. Normally, this monitoring stays on the subconscious level.

Primary Emotions – These are easily recognizable emotions, including fear, anger, happiness, sadness, and disgust. They are also referred to as *basic emotions.*

Secondary Emotions – These are emotions that mostly serve to regulate social relationships. Therefore, they are also referred to as 'social emotions'. Important categories are shame, reproach, pride, appreciation, and empathy (see Section 3.1). Secondary emotions usually have one or more underlying primary emotions on which they are based. They need more support by experience and learning than primary emotions.

In contrast to background emotions, primary and secondary emotions are usually immediately followed by corresponding feelings, at least in humans. For example, when the emotion sadness is evoked by some external cause, shortly afterward our brain produces thoughts which also usually make us sad intensifying the original sadness. We now clearly feel sad, and it may take us some time until some positive thoughts are able to 'pop up' again. The reason for this is that by associative learning, emotions and thoughts have been mutually linked. Certain emotions evoke certain thoughts and vice versa. Thereby, the emotional and the cognitive level are in constant connection.

Drives

As can be seen in Figure 3.2, Damasio views drives and motivations as different (and in fact more basic) than the emotions proper. However, drives strongly influence emotions and vice versa. As examples for drives, Damasio names hunger, thirst, curiosity, exploration, play, and sexuality. Again, as with emotions and feelings, he thinks that there is a difference between 'just occurring' drives and the state of being consciously aware of one's own drives.

Technical remark – It is not so clear how Damasio's description of feelings could be translated into a technical context. In principal, second-order processes would be no problem for a technical implementation, in fact they are rather common. However, what no one so far has managed

	Pride	**Shame, Guilt**
Stimulus:	recognition of a contribution to cooperation in self	weakness/failure/violation on the individual's own behavior
Basis:	joy, dominance	fear, sadness, submissive tendencies
Consequence:	reinforcing, cooperation	prevent punishment by others, enforcing of social conventions
	Gratitude	**Contempt, Indignation**
Stimulus:	recognition in others of a contribution to cooperation	another individual's violation of norms
Basis:	joy, submission	disgust, anger
Consequence:	reward for cooperation, reinforcing of tendency towards cooperation	punishment of violation, enforcing of social rules
	Sympathy, Compassion	
Stimulus:	another individual in suffering/need	
Basis:	attachment, sadness	
Consequence:	comfort, restoration of balance in group	

Table 3.1: Some of the main social emotions (after [Dam03, 156]). For each group, the stimulus triggering the emotion, the basis emotion, and the main consequences are identified.

is to build a machine which is in the least sense conscious of itself, with a real strong *sense of self* where the machine deeply, comprehensively, and autonomously monitors and reflects on its own (bodily) state and its relationships to the world in a first-person manner. What one can do rather straight-forwardly is to functionally implement emotional mechanisms, the 'mechanical' underpinning of emotions, and consequently also of feelings. Time will tell which kind of emotional quality (and maybe beyond in the direction of machine consciousness) can ever arise out of such efforts. As a consequence, in connection with the cognitive architecture presented in this work, I will only speak of emotions and not of feelings. The discussion has been included to completely cover the subject, especially as in psychoanalysis emotions (which are called 'affects' by Feud, see Section 5.2.2) also are conceptualized as having an objective, physiological part, and a subjectively experienced, that is, feeling part.

3.2.4 Account from Neurochemistry and Brain Organization

In [Pan05], J. Panksepp claims that a detailed neuroscientific understanding of human emotions may depend critically on understanding comparable animal emotions. Studying the neurochemistry and the brain infrastructure of emotions in mammals for many years, he has come to the conviction of the existence of basic emotions, an issue not undisputed within emotion theory. To settle the question, Panksepp maintains that one cannot do without brain research [Pan98, p. 34], thereby opposing himself to researchers who theorize on emotions in a purely conceptual way completely leaving out neurophysiological facts.

Basic and Complex Emotions

According to Panksepp, there are various subcortically situated emotive circuits, shared, more or less, by all mammals [Pan98, p. 34]. They give rise to rapid emotional responses. By intensively studying the organization and also the neurochemistry of the command transmitters on which these basic emotional circuits depend, Panksepp has come to believe that these basic emotional command systems are limited in number (four to seven) and that they *'arose from earlier reflexive-instinctual abilities possessed by simpler ancestral creatures in our evolutionary lineage'* [Pan98, p. 50]. Panksepp has suggested the following names for the basic emotional systems: SEEKING, LUST, FEAR, ANGER, PANIC/LOSS, PLAY, and CARE system.

Basic Emotion	**Cause**	**Associated Behaviors**	**Related Complex Emotions**
SEEKING	positive incentives	locomotion, exploration	anticipation, hope
LUST		repeat	pleasure, desire
FEAR	threat of destruction	flight, freeze	alarm, worry, anxiety
ANGER/ RAGE	invasion, frustration, limits	attack, fight	indignation, contempt, hate
PANIC/ LOSS	social loss, loneliness	distress vocalization, social attachment seeking	grief, sadness, separation distress, panic attack
CARE		parental care	tenderness, love
PLAY	only when other instincts are met	learning of physical and social skills, symbolic experimentation	joy, laughter

Table 3.2: Basic emotion command systems (after [Pan98, p. 50])

Furthermore, Panksepp thinks that through mixtures of basic emotions plus social learning a great number of complex emotions can be achieved. Those complex emotions are the result of evolutionary elaborations and interactions of the more basic systems with higher brain functions [Pan98]. Still, complex emotions are not necessarily conscious, although their purpose is to inform the higher cognitive apparatus how world events relate to intrinsic needs. This is the same purpose as that of basic emotions, but done in a more refined way. Of course, the higher the brain functions included, the higher the interpretations and 'appraisals' that come to (eventually) surround an emotional state. Thus, additionally to rapid emotional responses, the human mind with its capabilities to think, plan ahead, and speak can also produce slow, deliberative reflections on emotionally challenging situations. Nevertheless, the affective power comes from the more primitive neural circuits and, *'in case of emotional turmoil, the upward influences of subcortical emotional circuits on the higher reaches of the brain are stronger than the top-down controls.'* [Pan98, p. 301]

Interactions between Body, Emotions, and Cognitions

Panksepp's picture of emotions is a hierarchical one with various descending and ascending interactions between environment, body, lower brain areas, and higher brain areas [Pan98, p. 48]. On the one hand, the fundamental emotive circuits – and the physiological changes they induce –

interact with the brain mechanisms of consciousness. On the other hand, the cognitive apparatus can greatly shorten, prolong, or otherwise modify the more hardwired emotional tendencies we share with other animals [Pan98, p. 34]. Figure 3.3 illustrates the various interactions.

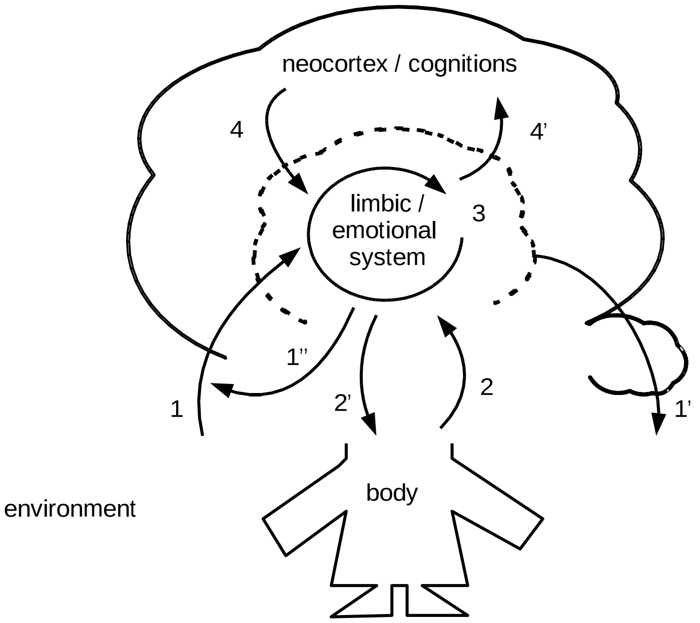


Figure 3.3: Ascending and descending interactions between body, emotional systems, and higher cognitive areas (inspired by [Pan98, p. 48]): (1) various external, and (2) internal sensory stimuli can unconditionally access the emotional systems, (1') emotional systems generate instinctual motor outputs, and modulate (1") external and (2') internal sensor inputs. (3) Emotional arousal can be sustained after the evoking events have passed, and (4) cognitive inputs can modulate the working of the emotional systems, and (4') the emotional systems can modify and channel cognitive activities.

Feelings and Consciousness

Panksepp uses the term *emotion* as umbrella concept that includes *affective*, cognitive, behavioral, expressive, and physiological changes [Pan05, p. 32]. The term *affect* includes a reference to the subjective feeling component *'that is very hard to describe verbally'*. This is very similar to how Freud conceptualized affects as possessing an objective, physiological and a subjective, psychic component (see Section 5.2.2).

For Panksepp, as for Damasio, consciousness is a graded concept with different levels, and there is a connection between consciousness and the ability to internally *feel* emotions [Pan03]. In contrast to other researchers, Panksepp already ascribes to animals (minor) forms of affective consciousness, and also internal emotional feelings, although not the ability to cognitively reflect on such feelings [Pan05]. The difference for him is that animals do not extend feelings in time, as humans can do with their rich imaginations. Internal feelings may directly mediate learning by coding behavioral strategies for future use, or perhaps they do this indirectly by interacting with the self-representational system within the brain.

Although Panksepp concentrates on the neuroscientific foundations of emotions, he does not deny the importance of introspection, that is, the possibility that the conscious mind can see the dynamics of its subcortical heritage. He votes for a comprehensive discussion of emotions considering the operation of neural circuits, behavioral/bodily changes, and affective experience *concurrently* [Pan98, p. 34].

3.3 Summary

This chapter has stressed the importance of an integrative view of body and mind. It has introduced emotions as *inherently evaluative* mechanisms that provide, in a feedback manner, a link between the body and the environment.

The focus on emotions will stay throughout the rest of the work. It will be elaborated how emotions can act as bridge between lower-order capacities of the human brain (the ones we share with some other mammals) and the higher-order, cognitive abilities which are most developed in humans. The latter are usually carried out in human language, which is a shared, symbolic communication system.

For the design of the cognitive architecture proposed in this work, the basic assumptions and the main claims of embodied cognitive science are adopted. However, most of the existing architectures based on the embodied approach stay very ‘flat’, that is, concentrate only on low-level tasks such as the sensorimotor control of movements. In this aspect, the proposed architecture deviates significantly from many other cognitive models within the embodied approach. In contrast to these architectures, the emphasis of the present work lies on a description of *how to integrate bodily phenomena with higher-level cognitive and finally psychic* phenomena into *one* model. The proposed key for doing so will be the neuro-psychoanalytic picture of the human mind. Before describing this picture (including the psychoanalytic conception of emotions) in Chapter 5, the following chapter will give a short overview of existing computational models of emotions, demonstrating their achievements but also their deficits towards the creation of a comprehensive, unified model.

I laugh, I love, I hope, I try, I hurt, I need, I fear,
I cry. And I know you do the same things too,
So we're really not that different, me and you.
Colin Raye

Our similarities are different.
Dale Berra

4 State of the Art

Emotions are a key element of the proposed ARS-PA architecture which will be presented in Chapter 6. The task of defining emotions in technical terms has often been accounted as infeasible. The main reason for this lies in the difficulty of reaching a profound comprehension of emotional behavior. Nevertheless, an increasing number of researchers in the software agent and robotic community believe that computational models of emotions will be needed for the design of intelligent autonomous agents in general, and for the creation of a new generation of robots able to socially interact with each other or people in particular.

During the last years, various kinds of systems trying to computationally model, implement, and investigate emotions have been developed. The systems differ significantly regarding their aims and assumptions. They refer in various degrees to existing theories of emotions. An overview of the many different emotion theories in psychology can be found in [Str03]. The concept of emotion is very broad and covers various aspects including physiological, motivational, and expressive ones, as well as the subjective experience of emotional states in the form of feelings, and the ability to cognitively reason about emotions. So far, most of the computational emotion systems focus *only on a subset* of the above aspects. The existing work can be roughly divided into *communication-driven* approaches that focus on the surface manifestation of emotions and their influence on human-computer interaction, and *process-driven* approaches that attempt to model and simulate the mechanisms of emotion as they unfold. This distinction, however, is not clear and mutually exclusive. Some of the systems address both perspectives.

In the following, a small number of key projects, mainly focusing on the process perspective of emotions, will be discussed. This is done because these systems come closest to the newly proposed psycho-analytically inspired cognitive architecture such that comparisons between them and the new ARS-PA architecture can be made. Related work directly referring to psychoanalysis is almost non-existent. The chapter closes with a presentation of one of the rare exceptions.

4.1 Low-level Emotional Approaches

Some of the early efforts concerning the usage of emotions in computational systems were devoted to the design of emotion-based architectures for *adaptive autonomous* agents. As defined in Section 3.1.1, a characteristic of autonomous agents is that they can fulfill their tasks without the help of other agents and especially without the intervention of human operators. It is further defined that agents are called adaptive if they are able to change their behavior according to the current state of the environment. Adaptations can happen on different time scales. Models that deal with adaptations to momentary, smaller changes are referred to as *action selection models*. Here, emotions support rapid, context-sensitive decisions. Adaptations to bigger changes involve the capability to change and improve behavior with time based on experience, and, thus, *learning models*.[1]. All of the low-level approaches presented in the following section have in common, that they are based on neurobiology or ethology, but not on personality-oriented, psychological theories of emotion. Neural net-based models or such that make use of some other form of sub-symbolic representation are common, but not a defining feature.

4.1.1 Action Selection Models – Case Study: Cañamero's Ethology-Based System

Emotions clearly possess bodily aspects. From an evolutionary perspective, to protect the body's integrity and well-being most probably is the primary task for which emotional mechanisms have been evolved in the first place (see Section 3.1.2). The view of emotions as control mechanisms that have been efficiently governing behavior long before higher-level cognitive mechanisms showed up (compare Section 3.2) is often utilized when constructing selection architectures for autonomous adaptive agents. This approach leads to ethology-inspired architectures for which agents with a physically embedded body – whether real or simulated – and mechanisms to prioritize the usage of limited resources are important design aspects. As a prototypical example, the architecture proposed by Cañamero will be presented.

In [Cn97], Cañamero designs a behavior selection architecture for robots together with an environment containing various types of resources, obstacles, and predators. In this environment, the robots have the task to survive as long as possible. In order to maintain their well-being (internal 'milieu'), they have to carry out different activities whereby the behavior selection process is influenced by the following components:

- *Synthetic physiology* (consisting of survival-related variables (e.g. energy, blood sugar, etc.) and hormones)
- *Motivations* (e.g. hunger, aggression, curiosity, fatigue, self-protection, etc.)
- *Behaviors* (e.g. attack, withdraw, eat, drink, play, rest)
- *Basic emotions* (interest, happiness, fear, anger, sadness, and boredom)

Motivations (needs) are activated when the survival-related variables depart from their homeostatic regime. For example, when a robot is too warm, its motivation to decrease its temperature

[1] There are also purely emergent models of emotions for autonomous agents. In such models, the emotional mechanisms are not an explicit part of the agent architecture (an example being, for instance, the Braitenberg vehicles described in Section 2.1.3). Such models will not be discussed further. Emergent processes are assumed to exist, but they are also considered *not* to be *sufficient by themselves* to produce intelligence. Emergent processes may be dominant at the lower levels. However, it is the very essence of dynamic systems theory, that smaller parts self-organize into bigger wholes on a higher level, whereby the emerged higher-level entities cannot be completely reduced to the lower-level entities. Intelligence/mind is a phenomenon on top of a system consisting of a hierarchy of ever higher levels of organization and complexity. The higher the level, the more symbolic the representative entities get and the more decoupled the ongoing processes become from the laws that govern the dynamics of the smaller entities on the lower levels. The ongoing processes also get less 'causally' (that is, bottom-up) determined, but more informational.

is invoked. Each motivation has an intensity, and the one with the highest intensity controls the behavior of the robot. Motivation intensity (and therefore behaviors) is influenced by the emotions of the robots. Emotions can be triggered by the presence of external objects, or by the occurrence of internal changes or patterns. Such an internal pattern could be for example when one or more of the motivations are continuously too high making the robot angry. Activated emotions release hormones that have an effect on the robot's physiology, attention, and perception.

Discussion – Cañamero concedes that the described architecture incorporates only very simple and low-level mechanisms, and that issues such as for example learning or the development of strategies (thinking, reasoning, planning, etc.) are not addressed. Thus, a cognitive level is not reached. Psychological or social aspects of intelligence are also not addressed. What can be accommodated into the model are different reward and punishment mechanisms. Cañamero considers such mechanisms as very important for learning, although they are not yet implemented in her model.

Models that have already incorporated learning mechanisms based on emotions will be discussed in more depth in the following section.

4.1.2 Emotion-Based Learning Models

Emotions can be viewed as evaluations that signal what is 'good' or 'bad' for the well-being of an autonomous agent. Therefore, emotions can be used for learning purposes. Learning based on positive or negative feedback is widely used in the machine learning community (labeled as *reinforcement learning*), however without referring to the concept of emotions.

Compared to the traditional reinforcement learning approach, the introduction of explicit emotional mechanisms provides additional flexibility that cannot be found in simple stimulus-response models of learning. For example, in [Mow60], it is demonstrated that emotions enable a two step learning process. In a first step, the agent learns to respond to a special stimulus with a special emotional state (*classical conditioning*), for example to 'fear' a certain sound if it is always coupled with pain. In a second step, it learns to associate *a behavior* with its influence on the emotional state (*operant conditioning*), for example, that going away from the sound reduces the fear. Simultaneously, the introduction of an emotional state such as fear *motivates* the agent to actively seek for different means of behavior if the originally conditioned one is not useful any more. Another advantage of explicit emotional states is that a singular emotion can influence several processes at the same time. For example, frustration can direct the focus of attention, trigger reassessment of a situation, activate predictions about how to improve the situation, etc.

Gadanho's and Hallam's Neural Net-Based Learning Robot

An attempt to analyze how the usage of emotions can improve traditional reinforcement learning in a neural network based architecture is presented by Gadanho and Hallam [GH01]. A robot with the task to navigate in an environment containing energy resources and obstacles is equipped with a set of emotions that directly map to its external and internal state. More specifically, emotions (joy, sadness, fear, anger) are the result of the bodily state (e.g. hunger, pain, temperature, eating, warmth, proximity).

The central part of the system is a neural net based, adaptive controller that learns to associate the robot's behaviors with the robot's bodily state *as the latter is perceived by the robot.* A difference between actual and perceived state can result because of hormones that can hide the real value of sensations from how the robot 'feels' its body. The hormones are produced by the emotions which thereby try to sustain themselves in a feed-back manner. As a consequence of this construction, emotions can still persist even if the emotion-inducing situation has happened some time ago.

The authors emphasize that a crucial problem of learning is how to associate smaller or bigger changes which are more or less continuously happening with specific, previously performed actions. The central idea of the model concerning this problem is that emotions get the role of signalling important state transitions, that is, transitions where there is a relevant and not just accidental connection between a perceived reward (or punishment) and the previously executed behavior. Thus, in the model learning only takes place *after* the detection of an *event that has significantly changed* the currently dominant emotion.

Discussion – The addressed issue of event detection and determination when to learn and when *not* to learn is a very important one, although the model itself is far from being comprehensive. Another criticism is that the relationship of emotions and bodily state is completely hard-wired.

Velásquez Connectionist Learning System

A more complex approach in which most of the stimulus-emotion-behavior links are not hard-wired was developed by Juan Velásquez [Vel98]. His connectionist model of emotion synthesis is called **Cathexis**. One of its implementations is done in a pet robot called Yuppy (a Yamaha puppy).

The architecture of the robot consists of several computational subsystems, among them a perceptual, a drive, an emotion, a behavior, and a motor system. *Releasers* filter data and identify special conditions according to which they then send excitatory or inhibitory signals to connected subsystems. Releasers substantially contribute to any evaluation procedures happening in the architecture. The emotion subsystem of the robot is based on the concept of basic emotions. This means that each of the following six emotions – anger, fear, sadness, happiness, disgust, and surprise – is modeled as a separate, discrete emotional subsystem. Emotions can result from interactions with the drive system, the environment, or people. For example, people can stroke or punish Yuppy, either producing happiness or anger.

For each of the basic emotional subsystems, the defined releasers can be categorized into four different types:

- *Neural releasers* (neurotransmitters, etc.)
- *Sensorimotor releasers* (facial expressions, postures, muscular tensions, etc.)
- *Motivational releasers* (drives, pain and pleasure, other emotions)
- *Cognitive releasers* (appraisal-based reasoning, attribution, memory)

Each emotion is calculated separately, but – in contrast to the OCC model discussed below – according to an update-rule that takes the same form for each type of emotion. However, what can vary are parameters. The new intensity of each emotion is a function of its releasers, its decayed previous value, and influences from other emotion intensities. The resulting intensity is compared to an emotion-specific activation threshold. Only when this is passed, the corresponding emotion influences the behavior system as well as other emotions. The behavioral repertoire also includes *communicative* emotional behaviors such as 'smile', 'wag the tail', etc.

As a special ability, Yuppy is able to learn 'secondary' emotions when rewarded or disciplined by a person. It can for example learn to 'fear' the sound of a flute when the occurrence of this sound is paired with the occurrence of a releaser that causes 'pain'. The learned associations are stored in the form of (new or modified) cognitive releasers.

Discussion – An achievement of the model of Velásquez is that it is one of the first that incorporates at least an approximation of all the types of influences known to be involved in human emotion synthesis. However, apart from the ability to learn associations and the usage of some cognitive elicitors, the model remains still rather low-level, more sophisticated cognitive capacities are not modeled. The focus is on basic control circuits implemented via connectionist networks. No explicit symbolic representations are used.

4.2 High-level Emotional or Appraisal-Based Approaches

While acknowledging the importance of physiology and behavior on emotional processes, appraisal-based models of emotions largely ignore these factors and focus instead solely on the *cognitive* structures and mechanisms that are involved in the generation of emotions. In these models, affective reactions are the *result* of cognitive appraisal processes that map the features of a situation onto a set of output emotions. To do so, high-level rule-based representations of goals, preferences, and situations are used.

4.2.1 The OCC Appraisal Model

One early appraisal-based model that has served as the basis for the implementation of several computational models of emotions is the OCC (Ortony, Clore, Collins) cognitive model of emotions [OCC88]. Originally, the model was not intended for emotion synthesis but to enable AI systems *to reason about* emotions, a capability thought to be useful especially for applications such as natural language understanding and dialog systems. The model does not use basic emotions, but groups emotions according to a scheme of cognitive eliciting conditions. Depending on the stimuli that cause the emotion, three classes of emotions are distinguished:

- Emotions induced by events
- Emotions induced by agents
- Emotions induced by objects

Using this structure, 22 emotion types are specified (Table 4.1). A complex set of rules couples the features of a situation with the agent's beliefs and goals. More specifically, events are coupled with goals, (moral) standards are coupled with agents, and objects are coupled with preferences (tastes). To do so, intervening structures and variables are used. To illustrate this by an example, if $D(a, e, t)$ is the desirability that agent a assigns to event e at time t then the potential for generating a state of joy P_j is given by a joy-specific function f_j that depends on the assigned desirability of the event and a combination of some global intensity variables (e.g., expectedness, reality, proximity) represented by $I_g(a, e, t)$:

$$IF\ D(a, e, t) \geq 0\ THEN\ set\ P_j(a, e, t) = f_j(D(a, e, t), I_g(a, e, t))$$

The rule above does not directly cause a state of joy. There is another rule that activates joy with a certain intensity I_j only if a joy-specific threshold $T_j(a, t)$ is exceeded.

	Positive Reactions	*Negative Reactions*	
Event	because something good happened (joy, happy-for, gloating)	because something bad happened (distress, sorry-for, envy)	**Goal**
	about the possibility of something good happening (hope)	about the possibility of something bad happening (fear)	
	because a feared bad thing did not happen (relief)	because a hoped-for good thing did not happen (disappointment, sadness)	
Action	about a self-initiated praiseworthy act (pride, gratification = pride + joy)	about a self-initiated blameworthy act (shame, remorse = shame + distress)	**Standard**
	about an other-initiated praiseworthy act (admiration, gratitude = admiration + joy)	about an other-initiated blameworthy act (reproach, anger = reproach + distress)	
Object	because one finds something appealing (love, liking)	because one finds someone/-thing unappealing (hate, dislike)	**Taste**

Table 4.1: The OCC structure of valenced reactions [Ort03, p. 194]. Emotions are rated as positive or negative, and, concerning their source, divided into three classes: those induced by events which promote/hinder the achievement of a goal, those induced by self/other-initiated actions which obey/violate (moral) standards, and those induced by objects which are liked/disliked.

Discussion – Concerning details like what values to use for the thresholds, or how emotions interact, mix, and change their intensity, the model is not very specific. Basically, the OCC model is a knowledge-based system to generate different types of emotions. Hence, it is not the most flexible one. Moreover, it has a limited capability for emulating 'hot', that is, emotionally affected, cognitions as it focuses on how cognitions influence emotions and not vice versa. Feedback effects from induced emotions to the cognitive system are (almost) not included. Moreover, the bodily aspects of emotions are not addressed at all.

4.2.2 OCC-Based Appraisal Systems

The OCC model of emotion generation has been given several trial implementations in computers, some of them will be presented below.

Em

An emotion generating computational system based on a subset of the OCC appraisal theory of emotion is the system Em by S. Reilly and J. Bates [Rei96]. Em is part of a larger project called 'Oz' whose goal is it to create *believable* agents that *appear* to be emotional. These agents are synthetic characters inspired by Disney figures that live in virtual worlds as for example the simulated house cat named Lyotard or the ball-like creatures called Woggles. The full architecture of these characters integrates not just emotions, but also rudimentary perception, goal-directed behavior, and language. The importance of goals influences the intensities of the generated emotions. There are thresholds for the emotions as well as functions that model emotion decay. Most importantly, Em's emotions are separated into positive and negative ones. By combining all the positive emotions, e.g., joy, hope, relief, etc., and all the negative emotions, e.g., distress, hate, shame, etc., respectively, a state called 'mood' can be determined that is either good or bad. The summing of the intensities within a group is done using a logarithmic formula. Emotions generated by Em can influence behavior and perception as well as some cognitive activities like the generation of new goals. For example, a character might be so angry that it generates a goal to get revenge.

Discussion – Most of Em's rules are hard-coded including social rules of which some researchers think that they should be more flexible. However, Em is part of a drama-inspired project in which the creators of characters do not want to give up deliberative control over their creatures.

The Affective Reasoner

Another implementation of (an extended version of) the OCC model is the Affective Reasoner by Elliott [Ell92]. It focuses on the generation of emotions among characters with social relationships, and on the *deduction* of other characters affective states. Again, the conditions to synthesize each of the 26 emotion types of the model are implemented as rules. Additionally, characters or agents are given a personality in the form of a set of symbolic appraisal frames that contain the agent's goals, preferences, principles, as well as current moods. The personality of an agent exerts a two-fold influence. First, it addresses which emotional state is derived based on each agent's individual appraisal frame. Second, it influences how an agent will express its emotional state. For example, an agent with an outgoing personality might express its joy verbally, a more inward type might simply enjoy an internal feeling of happiness. Concerning the derivation of other agents' emotions, an agent maintains an internal representation of the presumed ways in which others appraise the world. Based on this, it can perform forward logic-based reasoning from presumed appraisals, and events, to guesses about the emotions of others, and backward, case-based, reasoning from facts about the situation and the other agents' expressions to their presumed emotions. Generally, an agent can have three kinds of social relationships with other agents:

Friendship – The agent generates similarly valenced emotions in response to another agent's emotions.

Animosity – The agent generates oppositely valenced emotions in response to another agent's emotions.

Empathy – The agent temporarily substitutes another agent's presumed goals, standards, and preferences for its own.

Discussion – The Affective Reasoner concentrates on how to map appraisal variables with behavior. What is largely ignored is how to derive the value of these appraisal variables. Thus, it requires a huge number of very domain-specific rules to appraise events. Being reliant on such very specific, predefined rules does not give the system much insight in how to derive evaluations in general, nor in how to integrate evaluations with appropriate actions other than the ones explicitly provided for in one of the rules. What is good about the Affective Reasoner is that it appraises one and the same event from multiple perspectives (the agent's own and that of others).

4.2.3 Systems Based on Other Appraisal Models

There are other appraisal theories than the OCC theory that have been adopted as basis for computational systems of emotions, for example the model of Frijda [Fri04]. This model is the basis for a number of systems, including WILL and EMA.

Will

WILL [MF95] tries to address the problem of motivation using Frijda's concept of *concerns* thereby concerns determine what is important for a system. WILL's architecture consists of several modules (Perceiver, Executor, Emotor, Planner, and Predictor) running in parallel, all of them connected to a central workspace called the *Memory* acting as a blackboard. Before reaching the Memory, all data must pass through the concerns layer where it gets evaluated for relevance. As a result, charge values are attached to data items, and the one data item with the highest emotional charge is chosen to be the *focal element*. Only this element receives further processing. Thus, perceptions (and more generally also inferences) are continuously checked for their intrinsic significance by matching them against the system's (predefined) concerns. Thereafter, only relevant pieces of data can become the source of the creation of new beliefs, plans, and predictions.

Discussion – Will is designed to be as simple as possible and to use as much existing AI technology as possible. What is valuable of the system is that it experiments with the route from perception to execution via a kind of working memory. Based on a judgment of emotional significance, a focus of attention is realized which changes with a changing context. However, the charge formula to determine emotional significance could certainly be improved (arbitrary thresholds, the focus element loses charge every time a cognitive module fails to process it, etc.). Moreover, Will has only a small number of emotions implemented, for instance, it has no social emotions which could address the problem of blame/credit attribution. Correspondingly, WILL's awareness of others as social being is non-existent. Also, the psychological issue of coping mechanisms is not addressed. A further big deficit is that the system has not included learning mechanisms. Finally, like almost all appraisal-based approaches, is a completely classical symbolic system built on rule-based representations and algorithms.

EMA

The Emotion and Adaptation (EMA) model of Gratch and Marsella [GM04] focuses on how subjective appraisals can alter plans, goals, beliefs, etc. The model is implemented in Soar [New90] and encoded using the same knowledge representation schemes such as many autonomous agent systems, like STRIPS, or utility functions. The problem of modeling emotions is approached from a symbolic AI perspective. Cognitive processes serve to build an agent's interpretation of how external events relate to its goals and desires. Planning and dialog processes, and not sensor values, deliver the information for the following causal interpretation processes. A distinctive feature of the EMA model is that it goes beyond using appraisals or emotions to just guide action selection. It also includes coping mechanisms, like positive reinterpretation, denial, blame shift, etc., as alternative strategies to resolve conflicts. Thereby, original appraisals are modified, e.g. conflicting intentions may be dropped, uncomfortable beliefs changed, etc., thus resulting in a new interpretation of the environment/agent relationship.

Discussion – The attempt to model the wider range of human coping mechanisms is a big achievement of the EMA model and its most distinctive feature compared to the majority of other computational models of emotions. A drawback with regard to this capacity is that appraisal and coping are exclusively tied to causal interpretations produced by cognitive operations. A positive feature of the model is that appraisal frames include past and potential future developments. Again, there is a drawback insofar as no learning is implemented. There is also no episodic memory that memorizes positive/negative outcomes of action/event sequences. Correspondingly, there is no ego-development. The biggest deficit of EMA is that just the cognitive components of emotions are modeled and the bodily sources and effects of emotions are neglected. There are goals, but no 'hot' motivational concepts like basic needs or drives. There are also no emotional states enduring over some time (moods). On the whole, its a classical computational system, built on standard AI algorithms and representations.

4.3 Architecture Models – Case Study: (H-)CogAff

Apart from emotional computational systems that deal either with low-level mechanisms of emotion or with schemes of how to cognitively elicit emotions, there are also some ambitious models that try to combine all these aspects. As a prominent prototype of such an architecture-level model the approach to emotion and cognition of A. Sloman will be described in detail.

Sloman's CogAff (Cognition and Affect) Project

With his 1981 article '*Why robots will have emotions*', A. Sloman was one of the first to write to the artificial intelligence community about computers having emotions [SC81]. Regardless of this fact, and some other papers he wrote on emotions, he is not particularly interested in emotions, at least not in creating artifacts that just *simulate the effects* of emotional behavior. What he is interested in is the construction of a general intelligent system. Within this system, emotions would be just one part among others embedded in an overall architecture. Apart from the generation and management of emotions and motives, such a global architecture of the mind also has to include mechanisms for perception, learning, making plans, drawing inferences, deciding actions etc. All these mechanisms have to be implemented within a resource-bounded agent. For Sloman it is more important to design a *complete* system that may be shallow than isolated modules with great depth. Sloman refers to this view as *architecture-based approach* (sometimes he calls it *design-based*) [SCS04]. According to him, there is not just one 'right' kind of architecture, but a wide variety of architectures. To clarify this view, he defines the terms *design space* – the space of possible architectures – and *niche space* – the space of sets of requirements for architectures. Design and niche space are not independent from each other but connected via a diverse set of relationships. The collection of requirements to be satisfied by a functional system (its 'niche') determines a range of architectures that can be possibly used for its implementation. A given architecture may fit a given niche more or less well. For example, whether a robot needs or should have emotions depends on the tasks and environment the robot is intended for. Thus, the requirements to be fulfilled by the robot determine the kinds of 'emotional' mechanisms necessary or useful for the robot, and the structure of the architecture determines whether these mechanisms can be met by the architecture. The same interdependence of design and niche space not only applies to emotions but also to other cognitive abilities. Consequently, different information-processing architectures not only support different classes of emotions, but also different varieties of perception, mental reasoning, consciousness, etc.

The CogAff Schema

During the last years, Sloman and his colleagues have proposed various drafts for a cognitive architecture. All their drafts are based on an architectural schema that is very generic, called the CogAff schema [Slo01]. This schema is able to integrate various kinds of emotional and non-emotional mechanisms (Figure 4.1). So far, only some of the mechanisms that principally fit into the scheme have been explicitly elaborated, actually implemented in software and thoroughly evaluated. One fundamental conjecture for the generic framework is that it has to be multi-layered whereby each layer provides a different level of abstraction. Sloman thinks that a cognitive architecture for the human mind needs at least three layers: a *reactive layer*, a *deliberative layer*, and a *self-monitoring or meta-management* layer.

In the following, these layers are described ordered according their complexity, starting with the simplest one.

Reactive Layer – The reactive layer is the oldest one in evolutionary age. Systems with just a reactive layer can only produce relatively simple and predictable behavior. The reactive layer is based on the detection of characteristic stimuli in the environment. In turn, rather automatic behavioral responses are elicited. Reflexes based on direct connections from sensors to motors can be counted as the most basic form of reactive mechanisms. Although most of the reactive mechanisms are hard-wired, some simple form of conditional learning is also possible on this layer. However, reactive agents cannot make plans or modify their behavior to a larger extent. Thus, they are not very flexible. But they are very fast as the underlying mechanisms can be implemented in a highly parallel way using a mixture of analog and simple rule-based algorithms.

Deliberative Layer – The deliberative layer introduces 'what-if' mechanisms. It contains formalisms for combining existing behaviors in new ways, making plans, describing alternative possibilities and evaluating them before execution. The deliberative layer is also capable of learning generalizations. The plans created on this layer need some form of long-term memory for the storage of the various behavioral sequences and all the consequences of the alternative behavioral patterns. The step-by-step, knowledge-based nature of the construction of plans makes the involved processes serial and slow. Therefore, the question of how to allocate limited resources becomes an important issue. However, serial processing also presents some advantages especially when it comes to the task of associating rewards with previous actions.

Meta-Management Layer – The meta-management layer provides mechanisms for monitoring and evaluating internal states and processes. It can control, reject, modify, and generalize the processes of the lower layers in order to keep them from interfering with each other and to increase their efficiency. The categories and procedures on this layer are largely culturally determined. As the meta-management layer has incomplete access to all the internal states and processes a perfect self-evaluation is impossible.

The 'vertical' order just explained is combined with a 'horizontal' perspective decomposing cognitive processing into the stages of perception, central processing, and action. The combination of both ordering schemes leads to a grid-like structure.

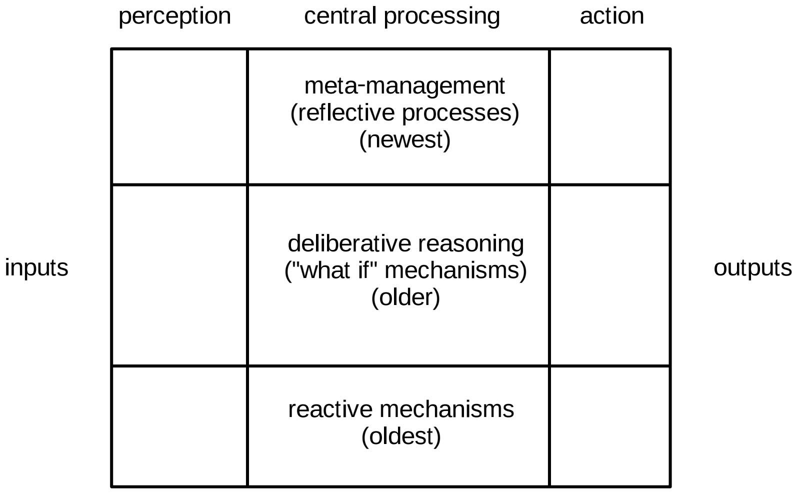


Figure 4.1: The CogAff scheme [SCS04, p. 20] defines a crude, first-draft division of mechanisms by superimposing two three-fold distinctions. Many information-flow paths between boxes are possible.

Concerning the question of representation, the suggested architecture is intended to be implemented by a hybrid symbolic/sub-symbolic modeling approach. The automatic, pre-attentive modules that correspond to the very old parts of the brain are thought to make use of sub-symbolic, distributed representations that enable highly parallel processing of incoming sensor data. However, higher modules are argued to require explicit symbol manipulating capabilities.

All three layers are thought to work in parallel. Between the processes on the different layers, lots of interactions take place. Thus, there are concurrently occurring processes. To manage them efficiently, a variety of *control states* are required. Control states can be based on different realizations, ranging from simple physical signals to complex processes acting on an informational level. Some of them are very short, and some of them endure over a longer period of time. Control states can be found on each layer of the architecture, they can circulate within the hierarchy, gain or lose influence, split up etc. Usually different control states are mutually dependent. The connections between control states can be supportive or suppressive – even leading to dead-locks. According to Sloman, emotions are one useful form of such control states.

Emotions, Affects, and H-CogAff

On several occasions, Sloman has criticized the confusion arising from the lack of generally agreed definitions of concepts like 'emotions' or 'affects'. In Sloman's terminology, 'affect' is used as an umbrella term, and the class of affective phenomena includes 'ordinary' emotions, like fear or anger, but also such concepts as desires, pleasures, pains, goals, values, attitudes, preferences, and moods[2]. In [SCS04], Sloman makes an attempt to clarify the distinction between affective and non-affective states. Therefore, he introduces the notions of *desire-like* and *belief-like* states. Both notions are functionally defined, that is, referring to the needs of an information-processing system or architecture. *Affective states are now suggested to be identified with desire-like states*, and the latter are informally defined as

> *'those [states] which have the function of detecting needs of the system so that the state can act as an initiator of action designed to produce changes or prevent changes in a manner that serves the need.'* [SCS04, p. 11]

Examples would be pleasures, pains, preferences, attitudes, goals, intentions, moods, and emotions. *Belief-like states* 'just' provide information to enable the desire-like states to fulfill their function. Examples include percepts, fact-sensor states, and memories.

Emotions as a subset of the above defined desire-like states are also given an architecture-based definition. Starting point is the idea that a system, while involved in some sophisticated processing, suddenly encounters a situation that needs fast handling. Therefore, *alarms* are necessary. This leads to the following attempt to very generally define an *emotional state* as state where

> *'some part of [an organism] whose biological function is to detect and respond to 'abnormal' states has detected something and is either actually interrupting, preventing, disturbing, or modulating one or more processes [..], or disposed [to do so], but currently suppressed by a filter or priority mechanism.'* [SCS04, p. 25]

Different types of emotions can be distinguished by the sources of their alarm triggers, the components which they can modulate, the time-scales on which they are operating, etc.

In principle, emotions can arise on each level of the architecture. A full three level architecture leads to at least three classes of emotions, namely a) *primary emotions* initiated in the reactive layer, b) *secondary emotions* initiated in the deliberative layer, and c) *tertiary emotions* initiated in the meta-management layer. An example of a full three level architecture is the H-CogAff architecture by Sloman et al. [Slo01] whose structure can be seen in Figure 4.2 (where 'H' stands for 'human').

Primary Emotions – These emotions of the reactive layer are triggered by simple stimuli and based on innate, largely hard-wired mechanisms. As they arise without prior cognitive appraisal, they are very fast. Examples are being startled by a loud noise, or being disgusted by some vile food.

Secondary Emotions – These emotions are triggered by events in the deliberative layer. They usually require some reasoning about goals, objects, and events. Unlike primary emotions which can only be triggered by actual occurrences, they can also be initiated by thinking what might have happened or what did not happen, etc. Examples are being relieved that something bad did not happen, or being worried about not reaching a goal.

[2] Defined in this way, the term 'affect' has a broader meaning for Sloman than for Freud. In particular, Freud did not use the word 'affect' as umbrella term that also includes drives and desires, instead he used it more or less as a synonym for emotions in the narrow sense (see Section 5.2.2 for a discussion of the topic). Then again, like Sloman, Freud considered affects as well as drives and desires to possess a certain kind of urgency, an energetic quality distinguishing them from other, more rational and not so impulsive mechanisms.

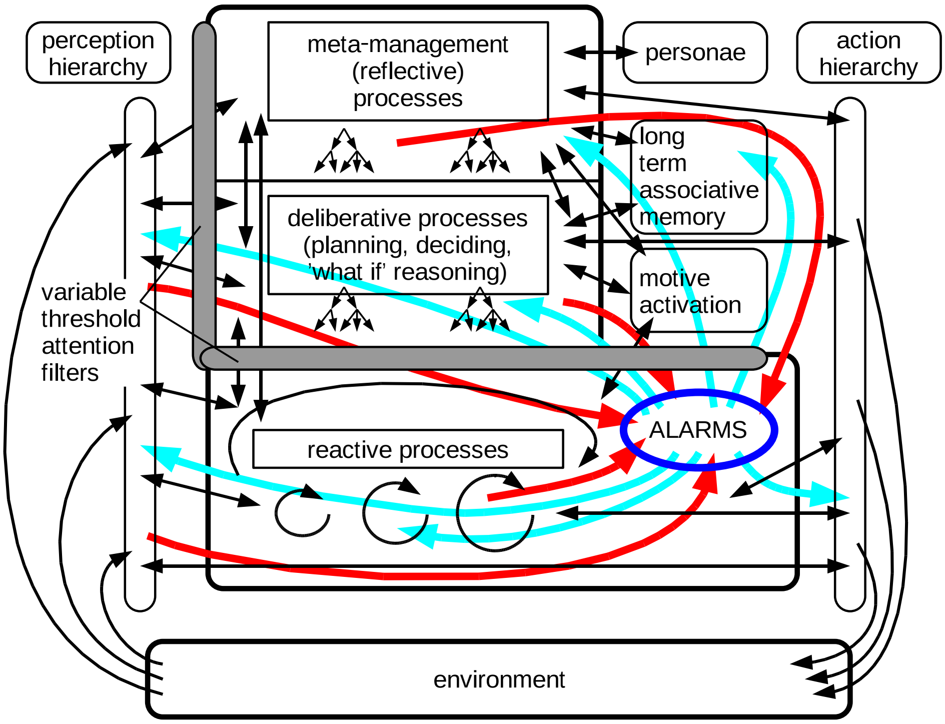

perception hierarchy
meta-management (reflective) processes
personae
action hierarchy
long term associative memory
deliberative processes (planning, deciding, 'what if' reasoning)
motive activation
variable threshold attention filters
ALARMS
reactive processes
environment

Figure 4.2: The H-CogAff architecture [SCS04, p. 23]. It is a special realization of the CogAff schema. All the boxes include many mechanisms performing different sorts of tasks concurrently, with complex interactions between them.

Tertiary Emotions – These emotions are situated on the meta-management layer. They are the ones for which the notion of 'self' is relevant. They disrupt high-level *self*-monitoring and control mechanisms, and can thus perturb thought processes. Sloman suggests that emotions associated with this layer include shame, grief, jealousy, humiliation, and the like.

In complex situations, all three kinds of emotions can coexist. In humans, primary emotions often immediately trigger some higher order emotions. Therefore, emotions cannot easily be attached to a certain level. The whole architecture does not contain a dedicated emotions module. Instead, emotions of various types are thought to emerge on all levels of the architecture from various types of interactions between many mechanisms that serve different purposes.

Apart from emotions, Sloman introduces *motivators* as another form of control states. Motivations can only be generated when there are goals. For Sloman, *goals* are representations that try to initiate behavior that adapts reality to its representation in contrast to *beliefs* that are representations that try to adapt to reality. Motivators have a structure that consists of several attributes (e.g., importance value, urgency, commitment status, beliefs about possible states, plans, etc.), and they are generated by a mechanism called motivator generator. Motivators compete for attention. There are filters that define a certain threshold to be passed to be able to recruit attentional capacities. These thresholds are variable. They can be modified by learning. As more than one motivator can pass the filter at one time, motivators need to be managed which includes processes like adoption assessment, scheduling, expansion, and meta-management.

Demonstration Implementations

During the last years, Sloman and his colleagues have developed a generic cognitive framework that is able to accommodate a wide variety of types of organisms and machines. However, at present, there is no complete implementation of the hypothesized architecture, although some of the aspects of the architecture have been implemented in specific test beds. One of the oldest is the minder scenario [Slo97]. In a nursery, a minder or nursemaid has to look after baby-robots keeping them out of trouble until they are old enough to leave. Possible sources of troubles are for example ditches the babies could fall in, or the danger that they are not recharged regularly.

Discussion – Although it has not been thoroughly implemented and evaluated in computers, Sloman's approach is very important for affective computing and especially for the problem of emotion categorization and synthesis. Still, Sloman does *not* believe emotions to be necessary for intelligence. However, he concedes that, because emotions have evolved in resource-constrained environments, they might prove useful for certain AI applications even though there is no physiological body. Quite generally, Sloman argues that there should not be put too much focus on physiological processes. Yet, to which degree bodily aspects of emotions in particular, and the concept of embodiment in general, can be neglected without abandoning something important is debatable (see 3.1). The biggest deficit of Sloman's architecture is that it stays too vague. This is related to the architecture's high degree of abstractness and to the permanent effort of Sloman to point out what else might also be possible and not to narrow down at least some concepts and mechanisms.

4.4 Psychoanalysis-Based Models – Case Study: Buller's Machine Psychodynamics

The computational cognitive systems using emotions presented so far have been inspired by ethology, and/or by various psychological theories about emotions and/or cognitions. Such systems are quite common. In contrast, the psychoanalytic perspective on the human mind has practically never been debated seriously and only very rarely consulted when it comes to artificially create human-like intelligent behavior. Correspondingly, one can hardly find systems which make use of psychoanalytic concepts for a technical design. One exception is the work of A. Buller which will be presented below.

In [Bul05], Buller tries to establish *machine psychodynamics* as a new paradigm of how to build brains for robots intended to achieve human-level intelligence. The approach is based on a central assumption of psychoanalysis: the view that mental life is a continuous battle of conflicting psychic forces such as wishes, fears, and intentions. To model the dynamics of the mental forces and the thereof resulting behavior, Buller introduces as his key elements tensions and defense mechanisms. The most fundamental operation within a psychodynamic agent is *the discharge of a tension, giving rise to pleasure.* In his work, Buller tries to elaborate the specific details of how pleasure has to be generated in order to enable intelligent systems or robots to (self-)develop and grow their cognitive capacities (their world model) via pleasure-oriented interactions. According to Buller, the following points are essential for pleasure generation (they are not completely independent of each other): a) it is essential that not a state of low tension, but a *move towards* it induces pleasure; b) correspondingly, pleasure is experienced not when being within a homeostatic regime but *in the process of approaching* such; c) the occurrence of a pleasure signal requires a previous deviation from the homeostatic equilibrum; d) the speed of the return to equilibrum, that is the shape of the pleasure signal following the discharge of a tension, is decisive for pleasure-induced action selection and learning to work properly; e) a pleasure signal should, with a small time delay, rise sharply after the detection of a discharge of a a tension, and afterward decay slowly.

In [Bul02], Buller makes a suggestion for a cognitive structure/architecture of a psychodynamic robot motivated by tensions. The structure is called Volitron and its central element is a working memory, called *MemeStorms*, modeled as a theater where populations of identical pieces of information ('memes') compete for dominance by outnumbering rival meme populations. Memes are the general data structure of the Volitron concept, defined as ordered pairs of propositions, and involved in practically all of its processes. Memes can represent beliefs, wishes, feelings, etc. An example meme could look as follows: <`agree` [to date a person who is] | `nice`>. Volitron uses some of its meme processes to internally create models of the external reality. Thereby, four different kinds of reality are built up: a model of perceived reality, a model of desired reality, a model of anticipated reality, and a model of ideal reality. Out of a comparison of memes belonging to the model of perceived reality with memes belonging to the model of desired reality, memes of satisfaction or dissatisfaction are produced. Based on their numerical balance, a level of tension is calculated which can activate the production of a physical action. Apart from model comparison and action-drive production, there are several other kinds of meme processing, called e.g. categorization, imitation-drive production, hunger-for-knowledge, candidate-plan generation, anticipated-reality creation, judgment of plans, and defense mechanisms. For example, action-drive production can stimulate the generation of multiple candidate plans which can be evaluated in a virtual execution process simulating the expected changes of the intended actions. As a result, a model of anticipated reality is created and compared with a model of ideal reality which contains 'moral' concepts of good and bad. The contents of the model of ideal reality can be acquired by 'young robots' following the instructions of their teachers. Defense mechanisms come into play when no satisfactory action can be executed for a long time .

Buller has tested some of its psychodynamic ideas on three robots. The first one, *Neko* is a physically constructed robot performing the task of obstacle avoidance. It is equipped with wheels and several sensors, including a camera, and a speaker. *Neko's* brain, partly running on an on-board module and partly on an external PC, has several tensions implemented, namely fear, excitation, boredom, and anxiety. *Miao* is a simulated robot equipped with the *MemeStorms* model. It has got the same tensions as *Neko* plus hunger. The tensions compete with each other whereby a dominating tension quickly suppresses its rivals. *Miao-V*, like its predecessor, is also a simulated creature, equipped with sensors, actuators, and tensions. However, in the case of Miao-V, its tensions represent a desire for red, green, or yellow objects. Miao-V's brain, realized as a neural network, can develop and grow in a literal sense: Each pleasure-related experience can add new cells and connections. *Mia-V* learns by interacting with a caregiver. As a 'newborn', *Mia-V* can only produce random moves and sounds as a consequence of its increasing tensions. The caregiver randomly gives it objects, and if by chance, an item accidentally discharges a dominating tension, the connection between the given object and the previously uttered sound is strengthened. In this way, *Miao-V* and its caregiver gradually develop a common, mutually-understandable proto-language.

Discussion – Buller's work drawing on psychoanalysis introduces some very interesting ideas, like the Freudian idea of a continuous battle between conflicting psychic forces, the idea of really taking the dynamics of pleasure generation and decay seriously, the idea of a 'young robot' learning via interactions with a caregiver (specifically leading to a co-evolvement of a proto-language). So far, however, Buller has not presented a comprehensive model of the human mind.

4.5 Summary

In recent years, an increased interest in modeling emotion within cognitive models of human intelligence and behavior-based agent architectures – whether software or robotic – can be witnessed. As attempts to computationally model intelligence move beyond simple, isolated, non-adaptive, and non-social problem solving algorithms, the challenge of how to focus and allocate mental resources has to be faced – especially when considering competing goals, parallel, asynchronous mental functions, and a highly demanding, constantly changing environment. Recent findings from affective neuroscience, contemporary psychology, and evolutionary biology indicate that emotions as well as drives and other kinds of non-rational behaviors service such resource allocation needs for biological organisms, including humans. A century ago, S. Freud already stressed the fact that human decision making is not only based on deliberate, conscious judgments, but also on non-conscious mechanisms such as drives and 'affects' and that these mechanisms inform the mind of bodily needs.

Out of the significant number of emotion-based computational architectures and applications, some prototypical approaches have been described. On the whole, they all deal with one or more key factors of emotional intelligence. However, often the work is carried out in a rather 'ad hoc' manner. Although most of the developed applications or systems somehow refer to an existing theory of emotion, in most cases only some elements of the theories are accepted, whereas, on the other hand, additional components are introduced without either explicitly stating so, nor arguing the choice. Of course, there is the problem that there is not *one* generally agreed on theory of emotion, but a great variety of them, most of them highlighting some aspects of emotions more than others. In general, emotion theories that are good for operationalization are preferred as a basis for computational models of emotion.

Most of the existing models and architectures cover either the low-level bodily aspects of emotions or the high-level cognitive ones. There are some ambitious approaches that try to integrate both aspects, but they are either not general enough, or too vague concerning the details of all the emotional and cognitive mechanisms that may be principally accommodated within the architecture. What is missing is a *comprehensive model* that nevertheless makes *clear statements concerning the included mental mechanisms and modules* and how they work together. The goal of this work is to present such a model based on the neuro-psychoanalytic view of the human mind. This view will be described in the following chapter.

The conscious mind may be compared to a fountain playing in the sun and falling back into the great subterranean pool of subconscious from which it rises.

Sigmund Freud

5 Neuro-Psychoanalyis

Over a century ago, the subjective approach to the human mind split off from the objective approach. Freud's principal method of investigation was not controlled experimentation but *'simple observation of patients in clinical settings, interwoven with theoretical inferences'* [Sol04, p. 84]. Being a medical doctor and having started his career as neurologist with controlled medical experiments, he turned to this method because, at his time, he did not have the science and technology to objectively study the difference in brain organization between normal and neurotic persons.

Meanwhile, surprisingly to some of the critics of psychoanalysis, support for Freud has come by a number of today's best neuroscientists, among them E. Kandel, nobel laureate in medicine for his research on the physiology of memory processes. Writing on the relationship between biology and psychoanalysis, Kandel notes that *'psychoanalysis is still the most coherent and intellectually satisfying view of the mind'* [Kan99, p. 505].

Since it has turned out that neuroscientific findings can validate some of Freud's most central assumptions [ST02, Dam03, Sch97, Pan98, Rot04], the gap between the often antagonistic fields of neuroscience and psychoanalysis has narrowed considerably, manifesting itself in the foundation of the International Neuro-Psychoanalysis Society in the year 2000 [nps07]. This event has a defining role in the formation of the new scientific discipline neuro-psychoanalysis. Previously, all began with an interdisciplinary study group in New York where neuroscientific findings and a psychoanalytic perspective on a topic at issue were presented one after another followed by a discussion between scientists of the two fields [ST02, pp. 309f]. Soon thereafter, several similar discussion groups were formed all over the world, a journal (*Neuro-Psychoanalysis*) was created, and an annual congress established (where, at the first congress in London, the formation of the Neuro-Psychoanalysis Society took place).

Both methods of approaching the mind, the objective neuroscientific and the subjective psychoanalytic one, have advantages and disadvantages. As they proceed in different directions, bottom-up and top-down, they can complement each other in a fruitful way. M. Solms writes: *'[..] for the integration of psychoanalytic knowledge with the neuroscientific equivalent is by no means tantamount to a replacement or reduction of psychoanalytic knowledge to neuroscience. Nobody gains anything by jettisoning the subjective perspective of psychoanalysis; our goal is only to strengthen it by coupling it to another, parallel perspective, which has a different set of weaknesses, so that the two perspectives may serve as mutual correctives for viewpoint-dependent errors.'* [ST02, p. 315]

Studying the details of neuronal organization, the contributions of specific neural systems, and the complex interactions among them delivers important facts and insights to understand the way our mind works. Still, an isolationist analysis of bottom-up, causal, physiological, and neurochemical processes alone does not deliver a satisfying model of the mind. First and most importantly, this is because neuronal processes and mental experiences are phenomena that take place *on different levels* of a complex, dynamical system with *many levels of complexity* – the brain. Second, the adaptive properties of the brain emerge *only relative to a crucial background of bodily and environmental structures and processes* (see Chapter 3). Third, the psyche, although a product of the neural infrastructure of the brain, also influences, top-down, the very same structure from which it emerges (compare with Figure 3.3).

In modeling the psyche, Freud's work plays the important role of a framework on which newly discovered details can be arranged coherently.

5.1 Neurobiology

In the following, some aspects of the brain will be very briefly discussed, already with the perspective in mind of what will be *functionally relevant* for the computational architecture proposed in this work. Due to this, the discussion will not be very comprehensive, and it will not go very much into neurophysiological details.

5.1.1 Brain Organization

The brain is an organ consisting of millions of neurons linked together by an even much greater number of synapses [Ste98]. More specifically, there are approximately about 10^{11} neurons, and each neuron is connected to a large number of other neurons via several hundreds to a few thousands synapses. The small chemicals which pass the synaptic gap are called neurotransmitters. This linkage permits the transmission of 'information' from one cell to another. Compared to other cells of the body – which of course also interact with each other in various ways – the capacity of brain tissue to transfer information is outstanding.

Neuroanatomically, a basic distinction is that between brainstem, diencephalon, and forebrain [ST02, p. 14]. Each of these structures contains various substructures. The brainstem is a direct extension of the spinal chord. In evolutionary terms, it is the most ancient part of the brain. Hanging behind the brainstem, there is the cerebellum ('little brain') which is involved in motor control. The diencephalon, coming on top of the brainstem, mainly consists of two parts, the thalamus and the hypothalamus. The forebrain, which is phylogenetically the youngest part, principally consists of the two main cerebral hemispheres. The outer surface of these hemipheres is the neocortex. There are also various forebrain nuclei lying subcortically, like the basal ganglia, and the amygdala. And there is the hippocampus which is not a nucleus but a structure consisting of a phylogenetically old kind of cortex.

The often used term 'limbic system' refers more to a theoretical concept than to an anatomical structure and is therefore vaguely defined [ST02, p. 17]. Mostly, thalamus, hypothalamus, amygdala, hippocampus, and the orbitofrontal cortex are included to the limbic system. All these structures play a very prominent role with emotions and memory.

The overall role of the *thalamus* seems to be a way station between cortical and subcortical structures [Kel91]. Most sensory information (including somasensory, auditory, and visual information, is relayed from the peripheral sensory systems to the sensory cortices through various paths of the thalamus. The thalamus also relays motor signals from the motor cortex. The *hypothalamus*, lying behind the thalamus, is involved in the regulation of the autonomous nervous system, the endocrine system, and with primary behavioral functions which are survival-related such as hunger, thirst, and sex drive [Sch70]. The *hippocampus* plays an essential role in the formation of new memories about experienced events [ST02, p. 162]. Apart from long-term memory formation, it is also supposed to be responsible for spacial navigation, and the formation of contextual representations [O'K90]. Next to the hippocampus, in the same subcortical area, lies the *amygdala*. It is connected with the hippocampus and besides receives input from all levels of sensory processing, including olfactory areas, and also transmits information back to the sensory cortices. Functionally, it performs a primary role in the formation and storage of memories associated with emotional events [Led96].

5.1.2 Brain Development and Learning

Our genes only determine the basic outline of the organization of our brain. The overall structure, including all the details, is strongly influenced by environmental factors during lifetime.

In the first years of our life, there is a selectionist process taking place (see e.g. [ST02, p. 147], [Ede87]). At birth, the brain disposes over billions of (synaptic) connections more than actually are needed. Those potential connections might be needed when constructing internal maps and models of the world. From the demands made by interacting with the environment the brain derives selection-criteria determining which connections are ultimately consolidated and which turn out to be superfluous. The ones which are not strengthened degenerate ('die'). Thus, from the great number of potential patterns of how neurons could connect only a significantly smaller number actually connects depending on the individual experiences of the organism.

It is widely agreed upon among today's neuroscientists that a great deal of our personality and even our ways to think is formed by the post-natal processes indicated above already in early infancy. For example, M. Solms writes that *'it is very likely that the networks that survived the great pruning processes of early childhood serve as templates around which all later memories are organized'* [ST02, p. 148]. The neurobiologist G. Roth also thinks that pre-natal and early post-natal influences determine the basic structure of our personality. He specifies that these processes are especially intensive during the first three years and clearly decrease around the age of ten [Rot03]. All in all, today's neuroscientific findings about the role of early childhood experiences support the psychoanalytic view on this subject which has always attributed them great significance. Of course, as the essential structures for forming conscious (explicit) memories (mainly the hippocampus and the surrounding association cortex) are not functional within the first two years, we cannot consciously remember our earliest experiences, thus giving an elegant neuroscientific explanation of what Freud called infantile amnesia. Nevertheless, those early experiences shape our personality and affect our adult feelings and behavior decisively.

After the period of infantile amnesia, during childhood, our episodic and semantic memories are gradually built up [Sch97, Ede89]. The consequence is that the perceptual process is more and more governed by deeply encoded and abstract knowledge derived from learning experiences. This has already been emphasized by the famous Russian neuropsychologist A. Luria in 1973 when he suggested that, while for a small child almost everything depends on the momentary sensory input and cognition is driven by concrete reality, later on, we see, what we expect to see, and we are surprised or fail to notice when our expectations are contradicted [Lur73, pp. 73–75]. This is a very central concept for the present work as well as for the whole ARS project: *the heavy use of 'abstract images' acting as templates.*

During our whole life, the human brain stays very adaptive with a huge capacity for learning. In principle, we can learn new things during our whole life until we die – although, most probably, there is some decrease in our learning potential the older we get. The discovery of the neurophysiological basis of memory consolidation (and thus learning) goes back to D. Hebb who, in 1949, postulated that some growth process or metabolic change takes place when one neuronal cell repeatedly excites another one [Heb49]. This result is often paraphrased as: *'Neurons that fire together, wire together'.*

What has been said so far in this section about the factors contributing to the development and adaptation of the brain's organization could be summarized as three different kinds of learning:

Phylogenetic Learning – This kind of learning refers to structures, knowledge, and capabilities we start with at birth because they have already been 'learned' by our ancestors. Usually we would not label them as learned but inherited. This kind of learning could also be termed *genetic learning.*

Pruning Processes in Early Infancy – This kind of learning is achieved by neuronal selection processes going on during the very first years of life. Thereby, in particular motor capabilities but also the emotional personality are greatly determined.

Learning – This kind of learning refers to the memorizing of facts, procedures, or skills through repeated rehearsal[1]. Neurophysiologically, it is based on the Hebbian law.

The above list can be compared to a passage of S. Freud [Fre02, p. 44]. There, Freud describes the id of its structural model (see Section 5.2.3) as representing inherited influences of the past, the super-ego as representing past influences of others as acquired during early infancy and the years of childhood, and the ego as mainly determined by individually experiences of present ongoings.

Technical remark – The above classification has relevance for the design of a technical cognitive system because it describes possible sources and modes of knowledge acquisition. Analog to the human case, a technical system can have the following kinds of knowledge:

- predefined knowledge stemming from the designer of the system
- knowledge acquired during an initializing training phase
- knowledge acquired via run-time learning

5.1.3 Brain Control Structures

Another very relevant aspect for the design of any technical system are control structures. In the human brain, there are various *neural control structures*, that is, *neural mechanisms devoted to the regulation of the flow of information* between cortical areas. There exist neural circuits, structures, or processes whose primary role is it only to modulate the activity of other neural

[1] In the case of a significant emotional arousal, a learning effect can already take place after just one exposure to an event.

circuits, structures, or processes, rather than tracking external events or directly controlling bodily actions. Some examples from neuroscience supporting this claim are described in [Cla97]. Clark not only describes these examples from a neuro-physiological perspective, he also discusses them from an information processing standpoint.

Generally, the brain includes many pathways which link distant cortical areas and which lead back from higher to lower brain areas [Ede87]. G. Edelman calls such reciprocal interconnections 'reentrant pathways' and argues that they allow the goings-on at one site to become usefully correlated with goings-on at the other site. By modulating the flow of information between various populations of neurons, certain classes of attentional effects, multi-modal memory recalls, and many other mental effects can be achieved.

5.2 Psychoanalysis

Apart from neurological findings, the construction of the cognitive architecture is based on psychoanalytic knowledge. However, there is no standard psychoanalytical model. But there are elements of the psychoanalytic characterization of the human mind shared by all the various schools of psychoanalysis. In the following those elements are outlined, thereby answering the question what is it that makes an approach inherently psychoanalytic. Thereafter, it follows an introduction to the important psychoanalytic concepts of drives, pleasure, and affects, and a description of Freud's structural model, also known as id-ego-superego model. This model has strongly influenced the design of the cognitive architecture proposed in this work (see Chapter 6). Finally, another structural model of the mind, going back to J. Sandler, is shortly presented. It serves as example of a rather recent psychoanalytic model, that already considers contemporary neurological findings and still stays rather adherent to the model of Freud.

5.2.1 Psychoanalytic Principles

The following five elements psychoanalytically characterizing the functioning of the human mind are commonly agreed upon by all the, sometimes quite distinct, schools of psychoanalysis. They will also be considered when designing the cognitive architecture.

Subjective Experience

The object of psychoanalytic studies are subjective experiences. This is directly linked with the psychoanalytic method of investigating the workings of the mind. The core of the method are *introspective observations* of mental experiences. This delivers data which cannot be obtained in any other way. The produced data can be juxtaposed or completed by data derived from objective observations – which is done within the neuro-psychoanalytic approach. What lifts the psychoanalytic method from pure phenomenology is the presupposition that the events occurring in the discontinuities and gaps of introspective experience, like Freudian slips or thoughts appearing out of nowhere, tell a lot about the underlying natural mechanisms and causes that make up a mind [Fre15d]. From taking introspective accounts of the mind seriously and combining them with observations of the life and behavior of his patients, Freud inferred his central theoretical claim stating that most of our mental functioning operates unconsciously. This was highly controversial at his time but is established neuroscientific knowledge today.

Focus on the Unconscious

The starting point for all psychoanalytic conception is that a large portion of all mental activity happens unconsciously, that is, below the threshold of consciousness. Mental activities, as can be seen by introspection, deal both with the outer world *and* with our subjective inner world. In analog to Kant, who posited that we cannot experience a *thing-in-itself*, but only via the intermediation of the 'perceptual apparatus', Freud postulated that we actually do not just have an outwardly directed perceptual apparatus, but also an inwardly directed one [Fre15d]. It perceives and processes what goes on within ourselves. Thus, psychoanalysis distinguishes between three entities:

- the *brain* as the organ of the mind which can be *objectively* studied,
- the *mind (mental experience)* which can be *subjectively* perceived, and
- the mental apparatus lying in between.

Although the mental apparatus is a concept inferred from objective and subjective data (coming from neurology studying the brain and psychoanalysis studying the mind/psyche), it is assumed to be 'real'. Its performance is not random but causal, and its activities give rise to observable phenomena and behavior.

Dynamic Unconscious

The function of the mental apparatus is to mediate between the outside world and the needs of the organism. This can be seen when looking at how Freud described his concept of 'drive' which he assigned a pivotal role in his theory. According to Freud, a drive is defined as follows:

> **Drive**: *'[..] a psychical representative of stimuli originating within the organism and reaching the mind, as a measure of the demand made upon the mind for work in consequence of its connection with the body'* [Fre15b, p. 122].

Thus, a drive is situated on the border between the somatic and the psychic. With actions directed at the outer world, the demand a drive is posing can be met, thereby linking outside world and internal world. Previous such linkages are registered in memory, evaluated in degrees of success and failure. This evaluation is experienced as degrees of pleasure or unpleasure [Fre20a], and related to the ultimate source of all motivation according to the psychoanalytic view: *the seeking of pleasure*, manifesting itself in a desire to repeat previous experiences of pleasure and to avoid a previous experiences of unpleasure. This is known as the *pleasure principle.* Further, it is maintained that there are inhibitory forces (the *defense mechanisms*) [Fre36] whose purpose is to selectively repress unpleasurable memories from becoming conscious. Thus, mental life is recognized as a continuous battle between conflicting psychic forces.

Multipersonal Approach

The minimal unit in modern psychoanalysis is not the individual functioning in its subjective reality but rather the individual in relation to someone else [Fou90]. Humans live in complex social relationships, which are the source of most of their experiences of pleasure or frustration. While drives originate in directly ensuring bodily well-being, desires or wishes have the same function but do so mostly indirectly by aiming at the stabilization of the individual's social relations. Human behavior greatly arises out of desires or wishes which are relational processes consisting of

- a *need* acting as a motive force,
- an *object of desire* (be it a thing or a person), and
- a representation of
  - *the self* in its role in the relationship, and of
  - *the expected reaction* of the object of desire.

Inseparability of Cognitive and Affective Components

Cognitive representations – that is, internal images of concrete and abstract objects or events – would lose all their meaning if it were not for the system which attributes affective values (motivational or emotional) to them [ST02, pp. 275–281]. Without the affective system, an individual would not be aware of the world, and, on the other hand, without cognitions, the affective system would be blind. Only a combination of both components in the service of action selection and execution brings about that mental experience is *intentional*, that is, always about something.

In the following, the psychoanalytic view about the nature and functional role of affective states will be described in more detail.

5.2.2 Drives, Affects, and Desires

Freud did not use the term 'emotions'. However, he spoke of 'affects' and put them in close vicinity to his important concept of *drives* which has been just defined above in Section 5.2.1 as border phenomenon between the body (its needs) and the psyche. Like drives, Freud also assigned *affects* the role of a link between somatic and psychological circuits. He considered it essential for affects to be subjectively felt, and, at the same time ackknowledged the bodily manifestations of affects (see for example [Fre00, p. 579] and [Fre15a, p. 278]). More recently, Solms and Turnbull present a modern version of the same view by describing affects as *'internally directed sensory modality'* [ST02, p. 106] informing about the current state of the bodily self, as opposed to the 'normal' sensory perception which informs about the outside world.[2]

The most basic kinds or qualities of affects in psychoanalysis are degrees of pleasure and unpleasure, and the most basic process leading to pleasure is the satisfaction of a drive. Freud wrote

[2] Note that, even in the (standard) case that affects are activated by external stimuli, what is perceived when one experiences (i.e. *feels*) an emotion is the own subjective response to an event – not the event itself [ST02, p. 107].

> *'We have decided to relate pleasure and unpleasure to the quantity of excitation that is present in the mind [german: 'Seelenleben'] but is not in any way 'bound' [..] such that unpleasure corresponds to increase in the quantity of excitation and pleasure to diminution [..] the factor that determines the feeling is probably the amount of increase and diminution in the quantity of excitation in a given period of time.'* [Fre20a, p. 4]

Thus, Freud conceptualized pleasure as being evoked by a sudden diminution or discharge of a previous state of excitement. Later, in [Fre89, p. 15] Freud replaced *'quantity of unbound excitation'* with *'tension'*.

A central belief of psychoanalysis is that ultimately all behavior is motivated or *driven* by the seeking of pleasure and the avoidance of unpleasure. Previous experiences of pleasure are *desired* to be repeated, and previous experiences of unpleasure are desired to be avoided. Such experiences of pleasure or unpleasure are memorized throughout life. Each such memory provides a linkage between a drive – more generally a desire – and how it can be successfully satisfied in the external world or how the individual can fail in trying.

Pleasure and unpleasure are just the most basic affective categories. Actually the spectrum of affects is more diverse. Even babies already start with a set of specific, predefined affects (e.g. separation distress). These are aimed to guide their interactions with the world, especially those with their caregiver(s) (see e.g. [Den99, Dor93, MW06]). With time, more and more 'affectively charged' memories linking the self, objects, and the outcomes of possible interactions are stored. Consequently, affects are not always directly related with the satisfaction of a bodily need. They can also come from thoughts, fantasies, or action tendencies. Just think of the anxiety of losing a job, or the pride of receiving an honor. Roughly put, perceptions but also imaginations evoke desires, and any activation, satisfaction/frustration, or only *anticipated* satisfaction/frustration evokes an affect associated either in a predefined way or via experience.

Remark on emotions versus Freud's affects – To summarize, for Freud, affective states are essentially characterized by a) a strong excitement, b) a subjective feeling thereof, and c) physiological symptoms happening in parallel. Thus, to make a final statement concerning the different terminology, the terms 'affects' and 'emotions' can be identified, as long as one keeps in mind that for Freud the subjective, that is *feeling*, component of an emotion is always included in his conceptualizations.

5.2.3 Freud's Structural or Id-Ego-Superego Model

During his lifetime, Freud drafted two models of the psychic apparatus (whereby the first model comes in an earlier version referred to as *affective trauma model* and a later version referred to as *topographic model*). In both versions of his first model, Freud divided the functions of the mind along the line unconscious on the one hand, and conscious on the other hand (in the second version of the first model, he introduced an additional part labeled 'preconscious' which he defined as *'being principally capable of becoming conscious'*).

Later in his life, recognizing that the reality-constrained, executive part of the mind is not necessarily conscious, he drew a new model [ST02, p. 100]. For him, consciousness was no longer the fundamental organizing principle of the functional architecture of the mind. From 1923 onward, in his second and final model referred to as *structural model*, the human personality is conceptualized as a threefold structure, consisting of the id, the ego, and the superego part [Fre23, p. 24].

These psychic entities are in contact and in conflict with each other and with the stimuli coming from outside. The tripartite structure is thought to be dynamic, changing with age and experience. Freud's structural model, the final version of which can be found in [Fre33, p. 78], is the primary source of inspiration from a psychoanalytic point of view for the design of the cognitive architecture (presented in Chapter 6).

Id – The id represents the boiling container of instinctual forces and of repressed experiences. The id is always present, thus, becoming a constant in our personality. Although it is unconscious, it nevertheless strongly influences behavior. The id is governed by the 'pleasure principle' which energizes most motivational behavior. The pleasure principle gives expression to primitive drives humans share with other animals. It is also responsible for wishful thinking that disregards the constraints of reality.

Ego – The ego is governed by the 'reality principle', a pragmatic, logical approach to the world. Freud wrote: *'An ego thus educated has become reasonable; it no longer lets itself be governed by the pleasure principle, but obeys the reality principle, which also at bottom seeks to obtain pleasure, but pleasure which is assured through taking account of reality, even though it is pleasure postponed and diminished'* [Fre16, p. 357].

The ego is not necessarily conscious, in fact it is to a great deal unconscious. Its core capacity is rather *inhibition* of drive energies of the id than consciousness [ST02, p. 99]. The ego also deals with the requests of the superego. Freud drew the picture of consciousness as an entity residing in a small reception room [Fre20b, p. 259]. Before, there is a large anteroom where various mental excitations are crowding. On the threshold between the two rooms, there is a doorkeeper acting as a censor. He examines the various mental excitations and denies them admittance if he disapproves of them. Consciousness cannot see what happens in the anteroom. Thus, when impulses are blocked at an early stage (before getting entrance to the reception room), the subject does not become aware of this. However, ideas repressed in this way, can still persist and may manage to pass the censor in disguised form. This is called *sublimation* by Freud, and it is one of several defense mechanisms. Other examples are *rationalizing*, *regression*, *repudiation*, or *identification*. Apart from its controlling and inhibiting functions, the ego also is the executive agent, urging thinking and planning, and organizing attention, and deliberative action.

Superego – The superego develops in early childhood via social interactions. It is mainly unconscious and contains moral standards to judge, inhibit, but also guide actions. It also contains ego-ideals. The superego tells a person what he/she ought or not ought to do, and if the difference between his/her actions and his/her idealized view of him/her-self becomes too big, the result will be conflict and an anxiety signal. A strong superego serves to inhibit the biological instincts of the id, while a weak superego gives in to the id's urgings. Thus, the superego influences the ongoing struggle between the id and ego for dominance.

5.2.4 Sandler's Structural Model

Although the design of the proposed cognitive architecture (described in Chapter 6) is psychoanalytically guided by Freud's structural model, in the following another structural model of the mind will be shortly presented out of the great variety of clinical theories and practices present in psychoanalysis today.

The model which comes from J. Sandler [SS98] is more recent than Freud's model. Correspondingly, it already tries to match up with data obtained from today's neurosciences – which it achieves rather well. The reason for mentioning this model is not only because it makes a bridge to the hard data of modern neurosciences. More than this, as Sandler's model on the whole remains very adherent to Freud [PLC07], it indirectly demonstrates that it is still scientifically legitimate to take Freud's model as inspiration for a technical design of the human mind/psyche.[3]

In Sandler's model the mind is divided into three compartments, labeled *Past Unconscious*, *Present Unconscious*, and *Conscious*.

The *Past Unconscious* comes from the procedural and habit-like learning processes of infants during their approximately first three years when the structures housing autobiographic memories are not fully formed. When developed, individual episodic memories (which can be equated to the concept of explicit memory) are then making up the personal history of an individual.

In the *Present Unconscious*, every memory is generated or molded in the first instance by the templates of the Past Unconscious. It contains mental representations that are mainly unconscious and that are manipulated according to the (unconscious) homeostasis of the subject. However, it is not excluded that some of the mental representations it contains can have, to some degree, access to consciousness, for example, they can be visualized as internal images, or recalled in words during a psychoanalytic treatment.

The *Conscious* section of the model is the part humans recognize the present world with based on what they have implicitly learned in the other two sections. The implicit templates, engraved as consolidated memories that represent procedures, motor habits, or automatic acts of perception, have a structuring effect on each mental representation, decisively modifying the way how an individual relates with the world.

5.3 Neuro-Psychoanalytic Picture of the Human Mind

In [ST02, pp. 18–36], the authors write on how to combine neurology with the psychoanalytic conception, stating that this results in principle in simple, *unitary* picture. This picture will be sketched below. The starting point is the brain interposed between *the external environment* and *the internal milieu of the body*. Thus, the brain has contact to two worlds, one within ourselves, and one outside ourselves. Quite fundamentally, one of the most important tasks of the brain is to mediate between these two worlds, that is, between our internal bodily needs, and the external environment which critically influences the well-being of the body. To fulfill its intermediate task, the brain is with each of the two worlds *in two ways* in contact: It *receives information* from both worlds, and it *acts* on both worlds.

External information comes in through the sense organs. Within the brain, it flows up a perceptional hierarchy whereby incoming information gets progressively more condensed and, thus, abstract, meaning that, at the lower levels, neurons represent simple structure, and on the higher levels, neurons and groups of them represent more complex structure. In detail, this condensation (or symbolization) runs as follows: In the beginning, information is processed in the posterior parts of the big hemispheres, first, in a channel-like way in the primary cortices *within each*

[3] In this work, the models of Freud and Sandler are just juxtaposed such that their relatedness can 'shine through'. It is the task of psychoanalysts (and not of this work) to analyze in detail how both models relate to one another.

modality (seeing, hearing, etc.), and, later, it is linked in the *association regions* with information from other modalities and *integrated with traces of previous memories.* From there, information is transmitted further to the frontal association cortices in the forebrain. These cortices already belong to the anterior half of the big hemispheres which is, in contrast to the posterior half, not responsible for sensing but for motor processes [ST02, p. 25]. Now, in the frontal parts of the brain, the direction of information processing is reversed, meaning that first abstract contents are handled, and, later on, the more down it goes, more and more concrete contents. In detail, information flows from the *prefrontal cortices* responsible for the *generation and monitoring of potential action plans* (providing the capacity for mentally trying out alternative actions), over the association cortices of the motor half of the big hemispheres, to the primary motor cortices, which are finally connected to the motor organs.

Internal information is registered first and foremost by the hypothalamus, then associated with other information in the limbic system, from where it reaches the forebrain. There it *influences* just like external information *action preparation and execution.* Internal information represents our inner motivation, grounded in the somatic sensations *reflecting the needs* of the changing state of the body. It is communicated in a field-like, global way, meaning that sub-cortical nuclei (of the brainstem or the limbic system) can influence, simultaneously, extremely widely distributed neurons in the forebrain.

On the motor or acting side, the actions that are released can be *willingly executed* actions, but also *stereotyped patterns*, governing the control of the visceral milieu of the body, instinctual behavior, and automatically carried out emotional reactions.

Via the processing of internal information in the brain, the body influences the mental life. Action programs and their regulation are modified. Thereby, the inhibitory control of stereotyped patterns of the visceral systems, the control of emotions, and the control of consciousness (*'focus of attention'*) increases during the years until an individual becomes a mature adult.

The described processes of the brain bring about *mental experience* which, in turn, seems to consist of three basic (but most probably not distinct) properties [Sol07]:

- *Intentionality* – Mental experience is always *about* something, it has objects or meanings.
- *Consciousness* – Mental experience is subjectively perceived.
- *Agency* – The mind experiences itself as entity that makes decisions: 'I shall do this'.

5.4 Summary

The neuro-psychoanalytic view results in a *unified, functional model* of the human mind (or rather the mental apparatus). By combining neurological (objective) and psychoanalytic (subjective) observation, the following aspects can be inferred as essential for the neuro-psychoanalytic model [Sol07]:

Drives – There are bodily needs, and the mental apparatus mediates between these bodily needs of the organism and the outside world, resulting in *units of experience* such as: 'I am experiencing this'. Thereby, 'I am' is the product of internal bodily perception, and 'experiencing this' the product of external perception.

Emotions – Drive states (oscillations in drive tensions) are perceived and related to external objects resulting in states of pleasure/unpleasure, the most fundamental emotional distinction. Furthermore, there are basic emotions, evolutionary elaborations dealing with standard situations of universal significance (usually situations where something specific goes wrong).

Desires/Intentions – Traces of previous linkages of needs and objects are registered in memory, inclusively their outcomes (rated in grades of pleasure/unpleasure). Successful previous experiences are *desired* to be repeated, unsuccessful ones are tried to be avoided. This establishes a connection between perceptions and actions. Furthermore, due to a chronological ordering of the emotionally rated, memorized units of experience, cause-and-effect sequences of events are registered.

Cognitions – As the decision tree increases in complexity, delay is required. This is provided by the power to *inhibit* impulsive actions, contributing to the experience of ownership of action tendencies – *agency*. Memorized cause-and-effect sequences of events together with inhibition render offline simulations of perceptions and actions – cognitions – possible.

Thus, behavior is based on the perception of current outer and inner conditions, and on previous, *emotionally-afflicted experiences* stored in memory. The latter extend feedback in time between actions and perceptions leading to the fact that an individual never perceives an objective, neutral state of the world, but always one that already includes the evaluated consequences of its previous actions.

In the next chapter, it is described how the neuro-psychoanalytic picture described in this chapter, including all the stated psychoanalytic principles, is turned into a *functional technical model* of understanding and behavior generation – the proposed cognitive architecture.

We are threatened with suffering from three directions: from our own body, which is doomed to decay and dissolution and which cannot even do without pain and anxiety as warning signals; from the external world, which may rage against us with overwhelming and merciless forces of destruction; and finally from our relations to other men. The suffering which comes from this last source is perhaps more painful than any other.

Sigmund Freud

6 Cognitive Architecture Design

Within the ARS project [PLD05, Pra06] of the Institute of Computer Technology of the Vienna University of Technology, a new cognitive architecture for automation systems and autonomous agents has been designed by my colleague C. Roesener and me [RLFV06] with the support of psychoanalysts. The architecture, sometimes referred to as ARS-PA (*Artificial Recognition System – PsychoAnalysis*) model, combines low-level, behavior-oriented forms of decision making with higher-level, more reason-oriented forms of behavior generation and selection into one coherent model. At the lower levels, the presented architecture is hierarchically organized, but, at the higher levels, it gets more and more the character of a distributed or multi agent system where several equally powerful elements (realized as functional modules) compete for determining the next action of the system. The integrating factor for the combination of the different levels of the architecture is the psychoanalytic view of the human mind.

6.1 Preliminary Remarks

Before presenting the ARS-PA architecture module by module, some general aspects concerning the design of the architecture are outlined.

6.1.1 Why Using a Psychoanalytic Model?

The architecture is based on Freud's final model of the psychic apparatus, referred to as his structural model (see Section 5.2.3).[1]

From a technical standpoint, the following reasons were important for choosing the id-ego-superego model of Freud among the many other psychoanalytic models as basis for the psychoanalytically inspired cognitive architecture:

a) The id-ego-superego model is a structural model. Consciousness seems to be rather an emerging property of dynamic processes than a specific structure; however, when creating a technical system, in the beginning, a structure has to specified on which the – of course also to be specified – processes can be arranged.

[1] The architecture can also be brought in correspondence with the psychoanalytic model of J. Sandler briefly mentioned in Section 5.2.4. This is not so surprising since the latter is, by and large, a contemporary version of the former [PLC07].

b) The question of machine consciousness is highly controversial. So, trying to technically implement a psychoanalytic model of the mind (as any other model) which is based on a distinction between various forms or stages of consciousness, for instance Freud's first model, may spur unnecessary disputes, diverting attention from the actual functional contents of the cognitive architecture *in a too early state* of technical design. Moreover, as has already been addressed, in his later years, Freud assigned properties previously attributed to the system 'Conscious-Preconscious' to the ego, whereby, however, only a small part of the ego's activities was thought to be conscious (see Section 5.2). Instead, the ego's new core capacity was *inhibition* which Freud considered to be the basis of all the rational, and executive competences of the ego. Note that, nevertheless of what has been just said, the question of consciousness certainly remains a critical issue in any psychoanalytic approach.

The chosen psychoanalytic model will be complemented with established findings from neurology, neurobiology, ethology, and similar sciences, *as long as there are no contradictions.* This is necessary because psychoanalytic models, as derived top-down from analyzing a very complex but already existing structure, naturally lack information about details which *do* have to be specified when artificially constructing and synthesizing an intelligent system bottom-up. Now, that it has turned out that neuroscientific findings can validate some of Freud's most central assumptions, the gap between neuroscience and psychoanalysis seems not so insurmountable any more (compare with Section 5.3. For the technical design, the strategy is to proceed in both directions, bottom-up and top-down, in order to keep white spots in the cognitive architecture as small as possible. However, it is not intended to copy the brain on a structural level. The goal is to create an artificial system that aspires *functional equivalence.*

Although the envisaged autonomous systems, e.g. mobile service robots or building automation systems, do not possess a body of the same kind as a living creature, they possess needs that require to be managed (because of their limited resources and the tasks they have to fulfill), and they will be equipped with emotional assessment mechanisms to motivate and guide their behavior.

Apart from the chosen basic psychoanalytic model, many further psychoanalytic principles and insights will be incorporated in the technical design, shaping the structure, as well as the processes of the architecture. Those principles will be emphasized when they apply in the course of discussing the neuro-psychoanalytically inspired cognitive architecture. The architecture itself is a general one. It introduces various modules and how they are connected, thereby acting as a template upon which a variety of mechanisms and algorithms can be arranged, existing ones as well as still to be developed ones, of course as long as the fundamental psychoanalytic principles on which the architecture is based are not violated. These principles distinguish the presented architectural approach from other suggested general architectures, and provide very important and helpful constraints to guide the design of future elaborations of the architecture.

6.1.2 Inner and Outer World

A fundamental aspect of the proposed architecture is that its functioning shall be rooted in the body of the autonomous system. Consequently, there are two kinds of signals which have to be processed, internal (bodily) signals, and external (world) signals. Similarly, actions do not only influence the outside world, but also internal states. This conception is compatible with psychoanalysis as well as neuroscience (see Section 5.3).

As has been said, Freud postulated analog to the 'perceptual apparatus' responsible for experiencing the outer world, the existence of a similar apparatus designated to perceive and process our inner, bodily needs. Mental experience is a mixture of inner and outer experience. With its theoretical conceptions, psychoanalysis describes mental experience as product of the mental apparatus, whereby the used theoretical conceptions are derived from an introspective, top-down analysis of mental experience. Insofar as the working of the mental apparatus can be correlated with neurophysiological processes and structures, it can also be described objectively.

During the last years, various neuroscientists have stressed that our subjective, psychological inner world is inherently rooted in the monitoring of our body, that is, in the way we *feel* ourselves, and that this feeling gives rise to consciousness [Ede03, Dam99, ST02]. Being aware of our inner state – whether it is good, bad, or something in between – may not be necessary for intelligence but certainly is useful for the integration of external affordances, internal needs, and appropriate actions, the key task of intelligent behavior. In humans, the coupling of the current state of the self with the current state of the world is done in a fluctuating way, from moment to moment, whereby *'each unit of consciousness forges a link between the self and objects'* [ST02, 92]. The storage of correlations of internal and external events leads to the formation of a value-category memory enabling to link the current perception of world signals to *previous* value-laden experiences (e.g. [Ede03]; see also Section 5.3). With this emotionally-afflicted memory, the integration of signals representing bodily needs, with signals from the surrounding outer world is extended beyond *'the current present'* [Ede03].

Given of what has been said about the importance of the inner state of organisms, it is strange that artificial intelligence has completely neglected all aspects related to the embodiment of their systems and machines for such a long time. This has changed with the upcoming of embodied cognitive science in which tradition the present work can be viewed. Many of the basic assumptions and insights of embodied cognitive science are shared, however, the focus of the new approach is much more higher-level, and not so focused on the control of sensorimotor capabilities.

6.1.3 Interaction and Feedback

In recent years, people have realized that the sole purpose of perception – and in fact the origin of intelligence – is the guidance of action [ST02, p. 280]. In Section 3.1.2, it is addressed how actions are motivated by the necessity to fulfill bodily needs. Furthermore, in Section 3.1.1, it is outlined that there is always feedback from the environment, such that organisms are not just acting on, but interacting with the environment.

System/environment interactions did not receive the attention they deserve for a long time. However, for several reasons they are very important when designing an intelligent system. First, they can reduce the amount of world modeling required (see Section 3.1.1). Second, feedback has always an evaluative character (see the argumentation in Section 2.4.2). More specifically, experiences, that is, sensorimotor and emotional feedback, act as the basis for the formation of the cognitive categories and concepts by which a system perceives its environment (see Section 3.1.2). Finally, as described in in Section 3.2, nature has evolved *a hierarchy of more and more complex mechanisms* in between perception and action, all of them aiming to ensure the well-being of the body. The spectrum of mechanisms ranges from simple reflexes, through drives and emotions, to reasoning and planning capacities, each of them *working on different levels*, however, with many reciprocal connections between them, both of activating and inhibiting nature.

Figure 6.1 is an illustration both of the system/environment interactions, which according to the previous section can be divided into interactions between the world, the body, and the brain, and of the hierarchic structure of the mental apparatus.

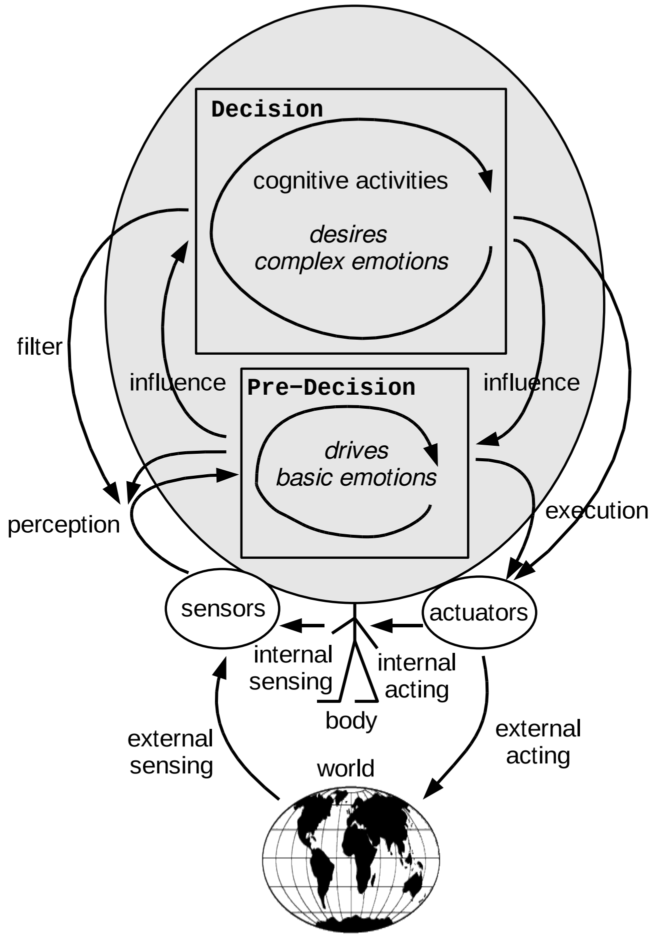


Figure 6.1: Outer world (environment), inner world (body), and brain/'mental apparatus'. Two kinds of information are processed: environmental and bodily data. Sensors and actuators act as interfaces between world, body, and brain/mental apparatus. The grouping of the depicted hierarchic levels into two units, labeled `Pre-Decision` and `Decision`, already refers to the modular structure of the proposed cognitive architecture. Note that throughout the work, the names of modules are in typewriter.

It is clear that improvement in homeostatic control has to go hand in hand with better kinds of representations. So far, no one can exactly tell how complex relationships (especially symbolic ones) are represented in the (human) brain. It is known, that humans possess various kinds of memory [Tul83], and that, during an individual's lifetime, the memory structures of the brain get gradually functional and more and more 'filled' with contents (see Section 6.3). However, neuroscientific findings show that remembering is not just an act of passive retrieval of some fixed set of data, it is an active construction process, a great deal of which happens unconsciously [ST02, p. 155]. This is a similar view as in psychoanalysis. There, some kinds of interactions are given particular attention, like those between an infant and his caretaker, or those between a therapist and her patient. An important element in these interactions are 'transference processes'. Those are processes where, in the course of interacting with each other, personal dispositions are not just passively remembered, but re-experienced [LBP02].

6.2 Overview of the Architecture

The cognitive architecture, technically implementing the neuro-psychoanalytic view of the human mind, has a modular structure, with each of the modules acting as a functional unit [RLFV06, DL+06, RLD+07, BLPV07, Roe07, Gru07]. Throughout the work, the names of the modules are set in `typewriter`. Basically, the architecture consists of two main blocks (labeled `Pre-Decision` and `Decision`), both of them containing various modules (see Figure 6.2). Each of the depicted modules, in turn, contains further sub-structures and processes. There are also different kinds of memory systems (depicted as containers to indicate their different ontological status), as well as information and/or control flows (all of them depicted as arrows) between the functional units and the memory units. Note that a detailed description of the modules of the architecture is presented below, each within an own section. Here just an overview is given and some general remarks about the architecture are made.

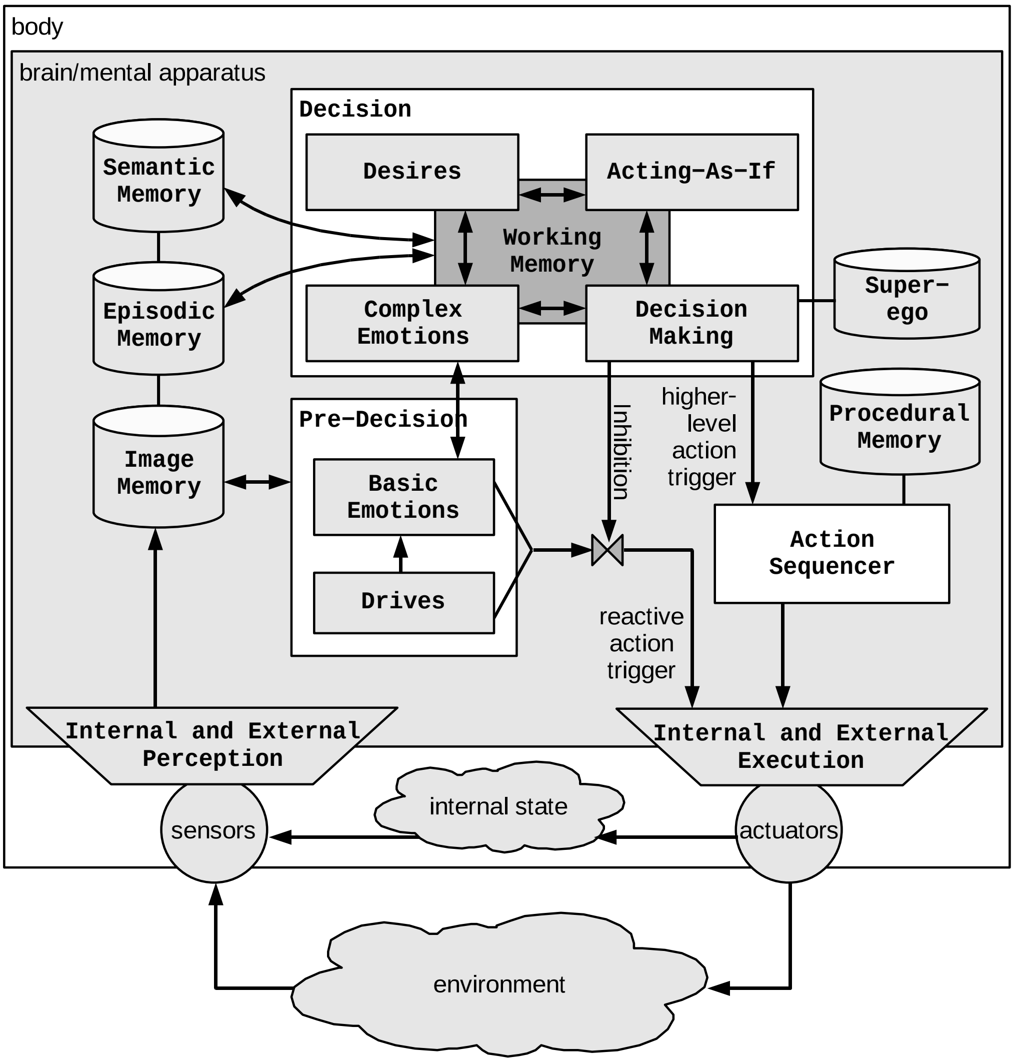


Figure 6.2: The ARS-PA architecture implementing action understanding and behavior generation based on the neuro-psychoanalytic view on the human mind.

In several joint meetings, the architecture has been discussed with psychoanalysts, some of them also being neurologists, to assure its principal compliance with psychoanalytic concepts and principles. From a neurological perspective, the modules of the architecture certainly do *not anatomically correspond* to structures in the brain, and, most probably, most of the functions bundled within a module are in reality much more 'spread out' over the brain (in space and time) and much more intertwined with on another. In particular, this applies to the depicted memory types. Note that a traditional 'store-house metaphor' perspective on memory is *not* intended. Such a view neither corresponds to embodied cognitive science nor to psychoanalysis [LBP02]. However, the presented architecture is understood as a first attempt of a technical implementation, a feasible starting scheme that shall be improved iteratively.

The general scheme of the architecture already shows some of the incorporated key ideas of the neuro-psychoanalytic view on the cognitive capacities of humans.

a) The architecture includes the inner state of the autonomous system (referred to as its 'bodily' state) and the environment as elements that are integrated in the form of feedback loops.

b) The architecture combines several low-level and high-level mechanisms, summarized respectively in the two main functional blocks of the architecture, the `Pre-Decision` unit and the `Decision` unit. The first corresponds to the psychoanalytic id, and the latter to the psychoanalytic ego[2]. Each of these units contains several sub-structures and processes. The various processes within and across the smaller and bigger modules of the architecture give rise to different 'levels of cognitive reasoning', e.g. reactive, deliberative, reflective, self-reflective. Generally, low-level responses are strongly, although not completely, predefined. Therefore, they are quick, but not always accurate. They provide the system with a basic mode of functioning in terms of in-built *goals* and *behavioral responses*. High-level processes take more time, but produce more distinguished forms of cognitive competences.

c) Throughout the architecture, emotions, in combination with drives and desires, are used as important integrative and evaluative elements.

Freud did not use the term emotions but spoke of 'affects'. He saw them as subjectively experienced manifestations of underlying physiological changes. This is similar to Freud's concept of drives which he defined as border phenomenon between the body and the mind (see Sections 5.2.2 and 5.2.1).

Technically speaking, emotions enable information processing systems to learn values along with the information they acquire. Therefore, a very important element of the architecture is the introduction of an 'episodic memory' that contains emotionally evaluated previous experiences.

Before starting with the description of the 'actual' modules of the architecture, the memory units of the model are discussed in the following.

6.3 Memory

Below, first a brief general analysis of memory-related issues is given, followed by a description of the types of memory referred to in the architecture.

[2] Note that in the structural model, the ego is not sharply separated from the id (see [Fre33, p. 78]). Only the repressed is cut off sharply from the ego by defense mechanisms (but can still communicate with the ego through the id). When relating the presented architecture with Sandler's model, the `Pre-Decision` unit could be mapped with the *Past Unconscious*, the `Desires` and `Complex Emotions` module with the *Present Unconscious*, and the `Working Memory` (and in particular the contents of the `Acting-As-If` module) with the *Conscious* component of Sandler's model.

6.3.1 General Remarks

Any cognitive architecture that aspires to result in context-dependent behavior must be able to 'store' information about objects and events in the surroundings of the autonomous system as well as how those relate to the internal state of the system. An abstract structure representing or modeling the outside world internally, that is, a *memory*, ideally is the result of an *adaptive construction process* ([ST02, 155]; compare also with[Ede89, Rie81]). As such, it can match the real world more or less accurately. In fact, as one and the same situation is never experienced a second time, the 'match' will always be a relative one. Also, memory is never a static collection of stored facts. Instead, memory changes with time as the constitutive construction process is practically never ceasing. According to the embodied and also the psychoanalytic perspective, remembering means re-experiencing means reorganization [LBP02]. The way the memory looks and the features it has, depend critically on the embodiment of the system (see Section 3.1.2) and the requirements of the environment. For example, environments that change only slowly (i.e. over generations), have low demands on the organisms' adaptive abilities. In this case, information can be stored genetically. Quickly changing environments (in relation to the lifetime of an organism) require memory types which are more flexible and which can operate in real-time. Information stored in genes has less adaptive value than information learned during lifetime, but it is available from the beginning (from birth). Thus, regarding adaptivity and access time, nature has developed a wide spectrum of 'memory solutions'.

Whether its contents are innate or learned, and how 'storage' may be realized in detail, memory can be characterized as the means by which the past influences the present (and, thus, also the future). The better a cognitive system can take into account *expected consequences* of its actions when performing a decision, the higher the quality of its decisions (in terms of adaptivity and foresight) will be (they will get better, although not faster). Humans outplay animals significantly in the degree by which they are able to evaluate *the consequences of potential alternative behaviors* – without actually carrying them out. Their outstanding *acting-as-if* capability makes their behavior less automatic and predictable than that of animals. The degree of deliberateness that can be attributed to an action corresponds to the range of potential alternatives the autonomous system can 'choose' among.

A big question when technically designing a memory system is how the data structures should look like (for the current solution of the architecture see Section 7.2). The ability to evaluate hypothetical actions makes great demands on the structure and organization of the underlying memory system [MB93]. It requires

- memories of past actions and
- their evaluated results in a form
- such that they can be explicitly (i.e. declaratively) used in planning (acting-as-if) algorithms.

Such an evaluated result could take the form of *how it felt* to do this or that in a specific situation. This alone, however, would be not enough. Important is also the connection of actions with the goals that can be achieved with them (see the discussion in Section 6.6.4). This makes it possible to use past experiences in planning algorithms.

Nowadays, a lot is known about the neurochemical basis of memory in human brains (and those of other animals). Still, nobody knows exactly how experience is really encoded, especially when it comes to the higher-order, more symbolic contents of memory, like mental images. Physiologically, the human brain consists of about 100 billion (10^{11}) neurons, and each neuron is connected via several hundreds to a few thousands synapses to a large number of other neurons [Ste98]. Thus, the brain is a huge network of neurons that shows very complex patterns of activity. There is a gap between the understanding of neuronal activity and the understanding of higher-order mental abilities (which of course are processes on different levels). It is not known which aspect of neuronal activity carries which aspect of information. Therefore, computational models of memory only try to functionally mimic human memory. Two main approaches can be distinguished, the connectionist (sub-symbolic) one and the symbolic one (see the discussion in Sections 2.1 and 2.4.1). Sub-symbolic solutions seem to suit well the fine-grained levels of sensor and motor data processing in the very beginning, but the higher the level in the cognitive hierarchy, a symbolic solution may become more and more necessary. (Though, it may be possible to build the symbolic representation on a sub-symbolic one. Much probably, this should also be the case, after all, statistical approaches can deal much better – and in particular, more adaptively – with huge amounts of data.)

When technically realizing a memory system, the following stages of memory operation have to be considered and provided for by the used datastructures and algorithms [NL04]:

- *Encoding*
    - *Initiation:* Which kinds of events can trigger the recording of a new memory?
    - *Contents determination:* What kind of information is stored?
    - *Feature availability:* What features of a memory unit are available for retrieval?
- *Storage*
    - *Structure:* How should the data structure of the encoded memory units look like?
    - *Dynamics:* Can stored memories change over time, and if yes, how?
- *Retrieval*
    - *Initiation:* How should retrieval be triggered, implicitly and/or explicitly?
    - *Cue determination:* Which data should be used as a key to the retrieval of memories?
- *Usage*
    - *Modus:* How should a memory be used?
    - *Purpose:* For what purpose should a memory be used?

In particular, as has already been mentioned above, the dynamic change of memories in the course of the system's interactions with its environment is a very important issue. In particular, the formation and modification of memory categories has to be modeled in an adaptive way.

6.3.2 Memory System Types

The heavy use of predefined 'mental images' as templates to shape, and also to speed up, cognitive processes is a central feature of the presented neuro-psychoanalytic approach. Perception is not possible without memories. In fact, perception *is* actually a process of re-combining current inputs with previous experiences (projected ahead as expectations onto the world). Thereby, the role of memories is to organize previous experiences into re-cognizable[3] chunks of knowledge. Thus, the processes of the `Perception` units could be viewed as processes that take place within the (perceptual) memory structures, re-organizing their contents. Actually, the same is also true for all the other cognitive operations. However, depending on the cognitive level (high or low), the present task (moving, listening, planning, etc.), and the momentary processing stage (more on the perception or more on the action side), different 'types of memories' are involved.

In general, one and the same experienced situation is stored in more than one type of memory – although in different forms. This is in accordance with neurobiological findings. There, it is almost commonly agreed upon today that humans possess different types of memory systems [Bad97, Tul85, Tul93a, Sch97, ST02]. However, there is more than one categorization scheme, with the different schemes partly overlapping, common distinctions being short-term/long-term, explicit/implicit, procedural/declarative, semantic/episodic memory. In psychoanalysis, the present is conceptualized as being influenced by the past, however, there are no specific considerations about the nature or storage of memories.

Within the architecture, five types of memories are introduced and should be *functionally* realized: *a sensory, a semantic, an episodic, a procedural, and a working memory* (this list is in correspondence with, e.g. the proposal of E. Tulving in [Tul93a]).

`Image Memory`

The most basic form of memory used in the architecture and mainly associated with the hierarchical perception process is `Image Memory`. It contains symbols and image templates, representing objects or events (compare with Sections 6.4 and 7.2.2). It is constantly and automatically accessed, giving it a buffer-like character. When looking for an analog in neurobiology, it may be compared to what is referred to as immediate memory by M. Solms [ST02, p. 143]), or to the perceptual representation system (PRS) of E. Tulving. He characterizes the operation of the PRS a follows: *'A perceptual encounter with an object on one occasion primes or facilitates the perception of the same or a similar object on a subsequent occasion, in the sense that the identification of the object requires less stimulus information or occurs more quickly [..]'* [Tul93a, p. 29].

`Procedural Memory`

One of three types of long-term memory is `Procedural Memory`. Among them, it presumably develops first in infancy [Tul93a]. It is a kind of bodily, action-oriented memory, holding information necessary for the execution of behaviors, for example knowledge how to run, or how to drive a car. Normally, the execution of routine behavior is largely done automatically (implicitly). The contents of the `Procedural Memory` consist of *routines*, that is, *sequences of actions*. Routines are tagged by the goals that can be achieved with them. Besides, they have some other features,

[3] As no situation is experienced exactly twice, recognition is only achievable up to some degree.

like, scheduling and termination parameters, connections to the hardware (actuator) parts of the system and to the resources, etc.

Semantic Memory

Semantic Memory refers to abstract, timeless knowledge of the world. It is a long-term memory containing facts and rules about the world, for example the physical rules of the environment, spatial information, object categories, how objects relate to each other, propositions, etc. In the architecture, semantic knowledge takes the form of algorithms describing relations and constraints between all the data structures of the architecture, e.g. symbols, images, scenarios, desires, or actions. Cognitive processes are reliant upon knowledge from Semantic Memory. Tulving writes: *'Semantic knowledge provides the individual with the necessary material for thought, that is, for cognitive operations on the aspects of the world beyond the reach of immediate perception'* [Tul93a, p. 30]. In the architecture, for planning tasks, the Semantic Memory interacts with the Working Memory. There is also a strong connection between Semantic Memory and Episodic Memory. Semantic knowledge can be created by abstraction from Episodic Memory. Another source of semantic knowledge is explicit learning of facts and rules.

Episodic Memory

Episodic Memory is based on individual autobiographic events. Therefore, other than semantic memories, episodic memories do not apply generally. They are remembered from a first-person view. Moreover, Tulving writes: *'The act of remembering a personally experienced event, that is, consciously recollecting it, is characterized by a distinctive, unique awareness of reexperiencing here and now something that happened before, at another time and in another place'* [Tul93b, p. 68]. The requirement to literally re-experience events makes it necessary that episodic memories include emotional ratings. In both cases, with the Episodic and the Semantic Memory, memory units must be explicitly (declaratively) accessible. This requires a symbolic, localist representation. In the architecture, the prototypical Episodic Memory units are sequences of events called episodes (see Section 7.2). Events are characterized by feature elements where features can be of different dimension: e.g. drives, emotions, actions, context information.

The Superego is viewed as a special part of the Episodic Memory. It contains rules to moderate impulsive behavior (e.g. to prevent drive behavior from getting excessive), and rules for socially acceptable behavior. Social rules can be useful for autonomous agents that have to fulfill a task together as a group.

Working Memory

An active, explicit kind of short-term memory, situated in the Decision module, is Working Memory. In the case of humans, working memory is described by Solms and Turnbull as *'the ability to consciously 'hold things in mind"* [ST02, p. 83]. In the architecture, while Image Memory supports the perception process, Working Memory is associated with higher-level cognitive operations (see Section 6.6.5). It actively provides the most goal-specific information and streamlines the information flow to the cognitive processes. Thus, the main task of the Working Memory is to provide all the relevant information which is currently needed for decision making. Mainly, it is populated by active desires and their associated action plans. Desires are related to experiences that have provided a reward in the past. Another contents of the `Working Memory` are memories similar to the current perception, preferably those which have got a positive rating as outcome. Salient emotional signals, and events that have been classified as novel or exceptional also enter the `Working Memory`. Thereby, the origin for the classification can come from low-level processes (like basic emotions), or from processes of the `Working Memory` itself.

An important characteristic of `Working Memory` is that it automatically provides a task-related *focus of attention*. It can send top-down attentional information to the perceptual processes, directing the `Perception` module to actively search for specific sensory data. This data could be, for example, an object, that is needed to trigger a transition in an active action plan. Another possibility would be an event that is expected to happen next. Derived from the context, that is the scenario the system is in, such expectations are constantly projected ahead. If, in a given context, the system's expectations are not met, this is emotionally rated. Detected data which is highly unexpected is handled with increased priority.

6.4 Perception

The `External and Internal Perception` module includes filtering and symbolization processes of – internal and external – sensory data. External, like internal, stimuli are continuously streaming into the system. The enormous amount of information has to be reduced, filtering out relevant pieces of information. At the end, there shall be, on the one hand, the perception of external objects and situations, and, on the other hand, the perception of internal bodily states.

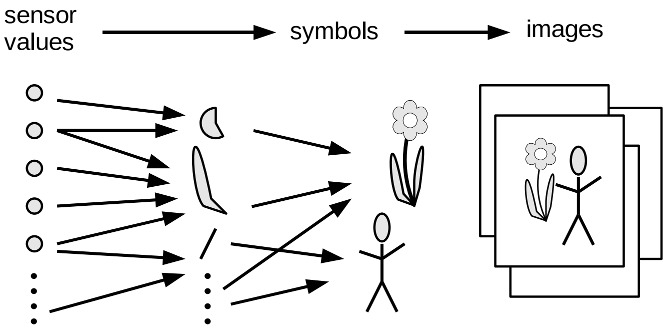


Figure 6.3: Symbolization process of sensory data. *Sensor values* are, in a hierarchic process, condensed to *symbols* which can in turn be grouped in *images*.

In the `External Perception` part of the perception module, external stimuli run through a symbolization process that consists of several levels which are hierarchically organized [PLD05, Pra06, Bur07, DB06]. Thereby, 'symbols' (i.e. chunks of information), on the lower levels represent fragments of perceptual information (perceptual primitives), such as edges, brightness, etc. Symbols on the higher levels contain more condensed information and represent more complex features, like a head, a torso, or a whole cat. On the lower levels, different kinds of sensory values (optic, acoustic, etc.) are processed separately. The higher the level, the more the symbols are derived from the association of various channels. The perception process results in assemblies of symbols that are referred to as 'images' (Figure 6.3) with an image defined as follows:

> **Image**: a collection of (internal and external) features making up an object, or a situation (or only parts of both)

The above definition entails another definition, that of a situation:

> **Situation**: a composition of all available information that characterizes the momentary state of an autonomous system

Apart from being hierarchical, the perception processes obey several further principles.

First, to reduce the computational load, the whole perception process ideally only calculates changes. This is related to the problem that, theoretically, the time gap between two situations is infinitesimally small. In [Tul83, p. 37], E. Tulving states that the prototypical unit of experience is an *event*, describing it as 'something that occurs in a particular situation'. He also distinguishes between *simple* and *complex events*, the former referring to an almost instantaneous change in a situation, the latter lasting longer in time. Given this, two further definitions are made (see also Section 7.2.3:

> **Event**: a happening when something significantly changes in a situation

> **Episode**: a sequence of events

> **Scenario**: a template of a sequence of events

Second, another important principle of the architecture is that, throughout all levels, to recognize objects, situations, and episodes (Figure 6.4), the interpretation process makes heavily use of already memorized symbol, image, and episode templates (the latter being referred to as 'scenarios'), stored in the various types of memory of the system. In Section 5.1.2, the neurobiological roots of such templates are discussed and it is emphasized that the perceptual process is more and more governed by abstract knowledge derived from learning experiences. As a consequence, human perception is not unbiased and neutral but shaped by previous memories. We recognize what we already know or what deviates just a little bit from that. In other words, in the bottom-up direction, by determining the match of sensory inputs with existing higher-level templates (e.g. symbols, images, or scenarios), objects or events are recognized. On the other hand, recognitions act as expectations that lead to internal predictions, dictate interpretations, or direct behavior. In this way, they exert a top-down influence on the perception process.

Third, perception is additionally biased by the fact that input and evaluation of data are organized as an *active screening process* searching for features that are required by the present state of body and mind. This constitutes another top-down influence from the higher-level processes of the brain on the perception process ('focus of attention').

Finally, according to the principles of embodiment and interaction, perception can be actively supported and shaped by actions. When confronted with a new and confusing object, the system could take it up, scrutinize it from different directions, etc. (see also e.g. Section 6.5.2).

Parallel to external stimuli, internal stimuli are perceived by the `Internal Perception` part of the perception module. This module watches over the 'bodily' needs of the autonomous system which are represented by internal variables. Each of these variables manages an essential resource of the autonomous system that has to be kept within a certain range. Examples for such variables are the energy level, or a resources-related health state.

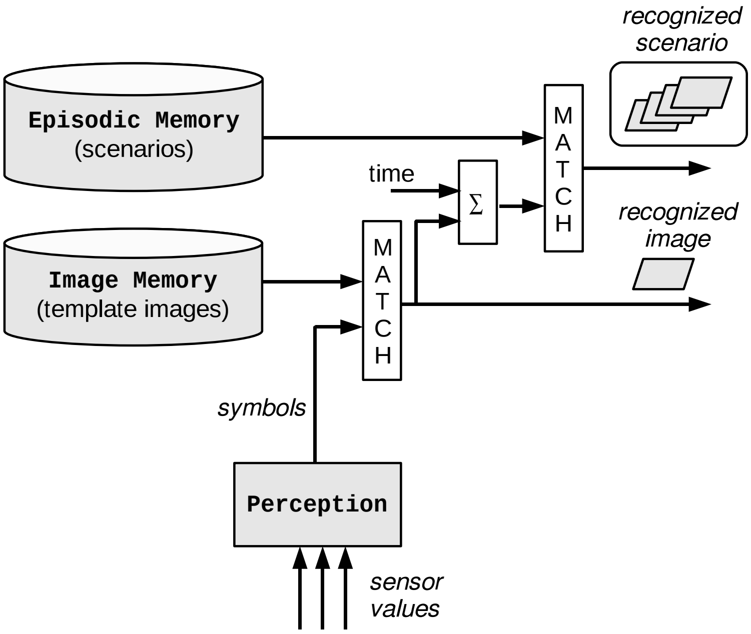


Figure 6.4: *Image and scenario* recognition via the usage of *templates*. Note that *scenarios*, which are ordered sequences of *images*, also (per se) have the character of templates. They represent 'extended' standard situations, that is, such that go beyond a snapshot of the current moment in time.

6.5 Low-level Decision Making – `Pre-Decision/Id` Unit

Freud's central finding was that most processes that determine our everyday feelings, wishes, and thoughts occur unconsciously. However, unconsciously motivated activities of the brain are not random. They serve to ensure the individual's well-being (see Section 3.1.2). In this respect, drives and affects play an important role. Freud conceptualized drives as boundary phenomenon, signaling the mind the needs of the body. Affects are ascribed a similar intermediate role. They possess physiological aspects as well as psychic aspects (this is, because they can be subjectively *felt*). Concerning their origin, Freud suspected affects as (inborn or learned) adaptations to prototypical environmental situations our ancestors had to deal with [Sol97].

In the structural model, unconscious, instinctual forces are enclosed in the ID which is governed by the *'pleasure principle'* (the seeking of pleasure). Today, J. Panksepp, a researcher of brain organization and chemistry of affective behavior, has identified at least four basic instinctual/emotional circuits shared by all mammals (see Section 3.2.4). He refers to them as SEEKING system (associated with a reward system called LUST system), FEAR system, ANGER/RAGE system, and PANIC/LOSS system (responsible for social bonding). Additionally, a PLAY system is also investigated. The SEEKING/LUST system resembles Freud's concept of the *libido drive*. It is considered as pleasure-seeking system providing for the self-preservation of the body in the widest sense, thereby motivating most of our goal-directed interactions with the world in the first place. Pleasure creates an urge for repetition of the pleasurable behavior while frustration leads to avoidance. This is an important ingredient for learning.

In the cognitive architecture, the `Pre-Decision` unit can be mapped to the ID (see also the footnote in Section 6.2). It consists of the `Drives` and the `Basic Emotions` module.

6.5.1 Drives Module

The Drives module gets input from the Internal Perception module. This input is related to 'bodily' needs, that is, technically speaking, resources of the system. In case that one of the internal resources is about to leave its range of 'well-being', this information is signified by the Internal Perception module to the Drives unit which, in turn, raises the intensity of a corresponding drive, for example hunger in the case of low energy. Provided that a certain threshold is passed, an *action tendency* to correct the impending imbalance is invoked, for the above example, this tendency would be to search for food. The output of the Drives module also goes to the Basic Emotions module where it is combined with the current external perception.

The way drives are conceptualized is influenced by Panksepp's SEEKING system which pushes mammals to search for the satisfaction of their needs in case of appetitive states (hunger, thirst, reproduction, etc.).[4] The range of 'well-being' is predefined for each resource. Drives need to be initiated a sufficient time span *before* a critical state is reached such that there is enough time to take counter measures. To activate the system, there must be several need-detector mechanisms (one or more for every resource), constantly sampling the internal world for signals indicating the appearance of a critical state. The intensity of the activation of the SEEKING system depends on the magnitude of the deviation from the normal value and is referred to as *tension*. An activation can potentially last for a long period in time, its discharge leads to pleasure. See also Section 7.2.4 for the datastructure of a drive, and Section 6.5.2 for the relation between tension and pleasure.

The SEEKING system stands in reciprocal connection with the LUST system, meaning that the satisfaction of a need discharges the tension raised by a need, and activates a reward signal referred to as *pleasure*. As a consequence, the explorative behavior due to the activation of the SEEKING system is stopped. The combination of the SEEKING system with the LUST system is very important for learning and will be discussed below (Section 6.5.2).

The SEEKING system is a general purpose system (in psychoanalytic parlance, one would say it is 'objectless'). Of course, the kind of need to be fulfilled largely determines the objects and events principally suitable to fulfill a specific need. Still, with a little interaction with the memory systems (especially Image and Episodic Memory) the system can always adapt its seeking behavior (or at least try to do so) to the environment (which may change). This is also a question of whether the system models an adult, or a child, because the knowledge how to satisfy basic bodily (internal) needs is largely learned during infancy and childhood. A technical analog to those pre-adult stages would be a training phase (compare with Sections 5.1.2 and 6.5.2).

The output of the Drives module are action tendencies. Not only single actions, but also whole sequences of actions (*routines*) can be initiated, unless the action impulse is not inhibited by the Decision module. Moreover, the Drives module can also exert influence on the Basic Emotions module, varying the strenghts of one or more of the basic emotions which are formed there. Finally, activated drives can also enter the Pre-Decision unit. There, they can, for example, become the origin of a desire.

[4] Panksepp refers to the SEEKING system not as drive but also as *basic emotional system*. This is just a question of labeling. In [ST02, p. 117], the authors discuss this issue, relating the language of Freud with that of modern neuroscientists, for example by comparing Freud's conception of drive with Panksepp's SEEKING system. Where Freud spoke of *'libidinal drives'* to denote the mental function activated by bodily needs of all kinds, modern neurobiologists use the term 'appetites'.

6.5.2 `Basic Emotions` Module

Solms and Turnbull write:

> *'Built on the foundations of core consciousness, both conceptually and anatomically speaking, is a set of connections that encode self-object relationships of universal significance. These are connections that link certain feeling states with certain classes of perception, which in turn, when activated, trigger 'pre-prepared' motor programs.'* [ST02, p. 277]

The task of the `Basic Emotions` module is to filter out stereotype situations, and to provide them with a first, rough evaluation. Stereotype situations are the ones that can be relatively easily recognized. As they are (by definition) often occurring, they usually possess some characteristic stimuli the system 'knows' about (if only implicitly, that is unconsciously). After having been recognized, an impulse for a (greatly) pre-defined behavioral response is almost immediately evoked, without much further processing.

Each basic emotion is related to a specific kind of behavioral tendency. Depending on the intensity of the basic emotion – and the strength of a potentially existing inhibitory signal from the Decision unit (Figure 6.5) – the evoked behavioral tendency is either directly sent to the `Action` module for execution, or, it is transmitted – together with the basic emotion values – to the `Decision` unit where it influences all further mental processing. Note that the basic-emotion command systems can be influenced by learning, especially in humans.

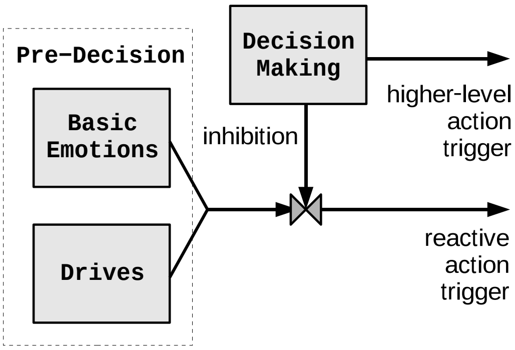


Figure 6.5: Higher mental functions can inhibit lower ones, corresponding to the fact that the output of the basic emotion systems can be inhibited by conscious, voluntary behavior. Nevertheless, there is no need to think before being able to act in a dangerous situation. Once a stimulus is associated with an unpleasant experience, the corresponding basic emotion system has to be immediately activated whenever that stimulus is encountered again. In the architecture, the association has to be therefore stored in `Image Memory`.

On the whole, basic emotions enable agents to switch between various modes of behavior based on the perception of simple, but still characteristic external or internal stimuli. This helps the autonomous system to *limit the set of potential actions* among which to choose, and to focus its attention by *narrowing the set of possible perceptions*. The system starts to actively look for special features of the environment while suppressing others.

The question which kinds of basic emotions there are is addressed for example in [McD69, Pan98]. See Sections 3.2.2 and 3.2.4 for an overview. Panksepp's basic emotion command systems are adapted to the needs of mammals in a natural environment. If not aiming at emulating human psychological or cognitive performance, they can be modified, depending on the demands of the technical application. Still, exactly the same basic emotions, when technically interpreted, can be very valuable for the design of any intelligent autonomous system.

SEEKING – The SEEKING system (which has already been described in Section 6.5.1 because of its correspondence with Freud's conception of drives) is involved in the management of essential resources. The resources may be the system's own resources, but possibly also the resources of someone else the system has to take care of, for example those of handicapped or elderly people in the case of cognitive assistants. It also contributes to learning as will be explained below.

FEAR – The task of the FEAR system is to identify potentially dangerous situations that require immediate reactions, thus, it plays the role of an alarm system. Again, the danger can be for the system itself or maybe someone else. Depending on the application and context, the associated reaction may be fleeing, hiding, or just being cautious.

RAGE – The ANGER/RAGE system is evoked when an expected positive stimulus or situation is not received. It leads to behavior where the autonomous system 'impulsively' tries to defend its resources, or to remove obstacles preventing it from reaching a goal. It can do so by selecting a potentially appropriate reaction from a (small) set of possible ones. Additionally, it can 'energize' (parts of) the system such that it momentarily increases the resources dedicated to the solution of the problem.

PANIC/LOSS – The PANIC/LOSS system is important for the establishment of social interactions in general, and especially for those between youngsters and their adult caregivers. This may be needed for automation or robotic systems that have to coordinate their behavior as an ensemble. It also plays a role for learning based on imitation which will be discussed in more detail in relation with the formation of the `Superego` (see Section 6.6.2).

In their book, Solms and Turnbull write: *'[..] all basic-emotion command systems [..] are, to variable degrees in different species, but to a very great degree in humans, open to influence by learning mechanisms. In other words, although these systems are innate, they are by no means 'hard-wired' in the sense of being unmodifiable. On the contrary, they appear to be specifically designed in a way that requires 'blanks' to be filled in by life experience (and especially early experience)'* [ST02, p. 133]. Essential for early learning processes is the mediation of adult caregivers, supported by the PANIC/LOSS system.

PLAY – A dedicated PLAY system – and to some degree also the SEEKING system with its exploratory character – substantially contribute to various forms of learning that go beyond simple reinforcement mechanisms. As outlined in [SC05], the key feature may be constant experimentation with external and internal actions during which re-usable chunks of information can be learned about possible combinations between external situations, potential behaviors, and internal states. This has to be supported by a memory that puts all these elements together, and a reward system that labels combinations as 'good' or 'bad', depending on whether they bring about pleasure or unpleasure (frustration) – the LUST system.

LUST – Historically, the system has been called *pleasure system.* It rewards behavior that has lead to the attainment of the object of a biological need or of a more general desire. It

> can also 'fire' in an anticipatory way when satisfaction has not been reached but maybe its likelihood has just got a boost. The LUST system is very important for learning. It is the emotional excitement of pleasure/unpleasure that lifts an unconscious process into consciousness and thus the focus of attention.

Essential for the pleasure system to promote learning is the appropriate modeling of the activation and decrease of the pleasure signal to support the finding of the relevant features which to associate by learning ('credit assignment' problem). In this respect, it has already been quoted (in Section 5.2.2) that Freud conceptualized the lowering of a tension to be felt as pleasure (and its raising as unpleasure). Moreover, not the absolute height of a tension is felt but rather changes within. See A. Buller's discussion of this topic in [Bul06] and the description of Buller's work in Section 4.4.

Tensions can come from unsatisfied needs (bodily as well as higher ones like desires or goals), and also from emotional arousals. For example, the activation of the basic emotion systems FEAR and RAGE is associated with varieties of *unpleasure.* Complex emotions (see Section 6.6.2) are usually related to potential future states of pleasure or frustration. To hope for something good to happen would establish a positive tension, and to fear that something bad will happen a negative one.

Learning about a novel object, for example, could consist of a sequence of exploratory actions, like maybe, first, approaching the object carefully, then, putting it to the mouth, and finally, trying to use it as a tool. From the outcomes of these trials (whether they charge or discharge any of the system's tensions), several chunks of information about the novel object are created and stored, for example categorizing it as non-fear-inducing, bad-tasting object that can be used to crack nuts. This is an example of a kind of active, bodily mediated learning driven by the SEEKING system.

6.6 High-level Decision Making – `Decision/Ego` Unit

Unless the `Pre-Decision` unit has not already 'fired' – that is, initiated the execution of an action – the more exhaustive evaluation and decision making processes of the `Decision` unit are run through. At this stage, higher-level cognitive processes like reasoning or planning come into play. Still, there are also affective components (motivations and emotions) to attribute values to cognitive representations. In the proposed cognitive architecture, these affective components are subsumed within the `Desires` and `Complex Emotions` unit. On the whole, the `Decision` unit can be referred to as EGO as it includes all the the functions attributed to the EGO, e.g. reality check, pleasure postponement by inhibition of drive impulses and acting-as-if, defense mechanisms (see Section 5.2.3).

The central task of the `Decision` unit is to associate desires, (basic and complex) emotions, and thoughts – which, in technical terms, are the elements of the `Acting-As-If` module (Section 6.6.4) and the `Working Memory` (Section 6.6.5) such that the release of an 'intelligent' action command follows. The whole construction corresponds to the agreed upon principle of psychoanalysis that meaning is achieved out of a combination of cognitive and affective components. In his psychodynamic theory, Freud described mental life as a kind of continuous battle between conflicting psychological forces such as wishes, fears, and intentions. The resolution of the conflicts, in one way or another, leads to compromises among competing motives. A great deal of the involved mental processes occur unconsciously, and the high-level conscious part of the mind has only limited access to the lower-level unconscious parts. In this respect, emotions act as a link between different levels by informing the higher cognitive apparatus about how world events relate to intrinsic needs.

Becoming aware of environmental changes or a changing status of available internal resources is a necessary prerequisite for the control of the impulsive action tendencies of the ID. The arousal of these impulses cannot be prevented by the EGO, but their execution can be inhibited. Also, for example, an original, 'primitive' action tendency can be substituted by a more complex behavioral pattern.

6.6.1 `Desires` Module

According to what has been said in Section 5.3, a desire[5] can be defined as follows:

> **Desire**: the urge or wish to re-experience a once pleasurable situation

Thus, the data structure of a desire – how it may look in detail – has to combine a need, an object of desire, and a representation of the self and its expectations concerning the interactions with the object and the fulfillment of the desire (see Sections 5.2 and 7.2). Desires aim to initiate behavior that expectedly lead to the fulfillment of the desires. Similar as in the case of drives, the fulfillment of a desire discharges one (or more) tensions. These tensions are related to unsatisfied (internal) needs of an agent, but potentially also to not yet achieved tasks the agent is due to carry out.

A desire can originate from the following sources:

- A memory (i.e., a mental image occurring maybe in the course of a perception, association, or planning process)
- A drive (i.e., a homeostatic, resource-related need)
- An emotion that builds-up a tension

The first case is the most basic one. It usually occurs 'spontaneously' (though, it may also happen in the course of a deliberative planning process). In general, the spontaneous recall of memories because of *some aspect of similarity with what is currently perceived or otherwise processed* (e.g. by the `Acting-As-If` module) is referred to as *'reminding'*, and it is the most basic process of all the processes of the `Decision` unit.

In a strict sense, the second and the third case are just derivations of the first case. The second case, with a drive being the source of a desire, shall be illustrated by an example. Given an activated drive, let's say hunger, a memory may pop up, having as content a specific kind of food, let's say chocolate. The memory may have popped up because, only recently, the system or agent has watched another agent eating chocolate. Thus, the popped up memory is just

[5] In the work, I interchangeably speak of 'desires' and 'wishes' because it does not make a difference at the current stage of elaboration of the architecture. In future, however, one could make the distinction of using the term 'wish' only for those desires which are processed on a linguistic level ('Wortvorstellungen').

the most recently perceived instance (that is, concrete episode) of a scenario related to eating (which in this case accidentally is a 'watching-somebody-eating scenario'). To summarize, as a result of spontaneous retrievals of stored images based on similarity, an unspecific drive of the `Pre-Decision` unit is turned into a specific desire of the `Decision` unit.

The third case – having an emotion as source of a desire – is also best illustrated by an example. Consider an autonomous agent that experiences a negative emotion, let's say shame, at a relatively high level for some ongoing time (or repeatedly). In this situation, the autonomous agent may elicit the desire to get rid of this unpleasurable (negative) emotion (or to prevent its re-occurrence). There are also cases where positive emotions can evoke a desire. Consider for example a system that experiences gratitude or empathy towards another system or agent. Any time it sees the other system, it may activate the desire to help the other system or agent.

In summary, desires are of similar nature as drives. Analog to them, they are the motivational source of actions, but now, more sophisticated and more enduring actions (because of an increased memory involvement). Compared to drives, desires are not necessarily directly related to the satisfaction of bodily needs but in most cases deal with the satisfaction of 'higher motives' – although it is not excluded that desires may originate in drives in a relatively straight-forward way.

Generally, a desire is triggered by a – perceived or remembered – object of desire. The object may also be a 'subject', that is, a person or autonomous system or agent. The objects of desire are not innate, but learned. For instance, when a tension is discharged, or an emotion of high intensity elicited (e.g. by an external event), the making of a new memory entry can be triggered (see Section 7.3). This is a learning process that links external object or situation, discharged or otherwise emotionally evaluated need, and previously executed action. Higher order learning processes can also be triggered, for instance such that transfer recurring episodic memories as abstract schemes ('scenarios') into the semantic memory.

Expectations about how to achieve the satisfaction of a desire ('wishfulfillment') take the form of *action plans* expanding the desire into goals and sub-goals. In Figure 6.6 the processing of a desire is shown, given that desires – as this is the case for the time being (see Section 7.2.4) – are implemented as graphs. Using graphs may evoke the impression of a finite state machine. However, the mind (and also the mental apparatus) is *not* considered to be a such, nor is the architecture aspired to result in such. Note that every moment, on many levels of the architecture, there are many processes running asynchronously and in parallel. Also, as has been mentioned in the above paragraph about learning and memory reorganization, desires (and, thus, also their representations) are not constant but *change with time* due to newly made experiences. The changes strongly depend on emotional evaluations (which are not conceptualized as state machines).

To expand a desire, there can be a contribution from the `Acting-As-If` module (representing deliberate thinking about wishfulfillment) and/or from the `Episodic Memory`. In the case of the latter, in particular experiences set on early in life, under the guidance of a caregiver, become important templates. The mechanism is related to imitation learning. Influencing factors for imitation are the similarity of situations, the (recognition of the) success of the observed behavior, and, most importantly, the status of the observed individual whose behavior is taken as a template. By imitation, out of a smaller set of innate motivational value systems, various desires are created. In parallel, also under the auspices of influential and respected others, (moral) standards and ideals acting as a counterpart to wishful drives and desires are formed. In the architecture, these

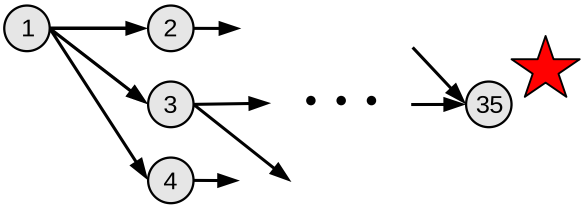


Figure 6.6: Graphical sketch showing the processing of a desire, from its initialization, to its fulfillment (indicated by a star). From a given intermediate state there may be more than one possible transitions to a successor state. Transitions can be based on (passively perceived) events, or on own actions, and they can also depend on emotions. Proceeding from one state to another on the way to wishfulfillment corresponds to the achievement of sub-goals. With the ability to perform actions, plans about how to reach the fulfillment of a desire can be actively constructed.

normative elements are collected in the `Superego`. The there contained codes of behavior have a great effect on which action tendency is finally selected for execution.

An activated desire directs the focus of attention to perceptions that are related to the fulfillment of the desire. There is also a focus on desire-relevant actions. Both filtering processes can be supported with information from the semantic memory if a cooperation with the planning (`Acting-As-If`) module of the cognitive system takes place. Other attributes of a desire include its intensity or tension (contributing to its priority), and the time it takes until a desire is given up. The latter is associated with an assessment of the degree of *success/failure* of a desire, which, in turn, is accompanied by a whole spectrum of emotional evaluations, ranging from hope to resignation. This will be discussed in more detail in the following section.

6.6.2 `Complex Emotions` Module

The `Complex Emotions` module is responsible for all emotional mechanisms that go beyond the ones defined in the `Basic Emotions` module (see Section 6.5.2). Humans can greatly shorten, prolong, or otherwise modify their more hardwired emotional tendencies (see Sections 3.2.2, 3.2.3, and 3.2.4). With an increased complexity of situations the individual finds itself in, a broader emotional spectrum is necessary to handle them. An evolutionary elaboration has brought about a progressive differentiation in emotional control mechanisms from an inaccurate, global evaluation to a more accurate, local one, manifesting itself in emotions that possess more specific appraisals of situations, and more specific behavioral responses (see Sections 3.2.1, 3.2.2, and 3.2.3). Generally, it is assumed that such more elaborated emotions – like shame, envy, mercy, or hate – arise from interactions of the more basic systems with higher brain functions [Pan98, 42]. Thus, complex emotions are directly influenced by basic emotions, and a full emotional response includes rapid and unconscious processes, as well as slow, deliberative responses – up to conscious, verbal reflections on an emotionally challenging situation [Pan98, 34].

In the architecture, the mechanisms of the `Complex Emotions` module deal with *derivations* (including mixtures) of the basic emotions referred to as *'complex emotions'*, but also with the 'pure' basic emotions themselves. This means that, for example, the (basic) emotion anger not only can be elicited or modified at the level of the `Pre-Decision` unit, but also by appraisals at the level of the `Decision` unit. In this case, however, anger is probably associated with a tendency to switch to a new method of problem solving rather than with an impulse to perform a simple attack behavior.

Input to the `Complex Emotions` module is coming from the `Episodic Memory`, the `Superego`, the `Pre-Decision` module, and the `Desires` module.

Functionally, emotions are control signals. There are two major lines along which complex emotions contribute to a more sophisticated control of behavior:

a) Some complex emotions extend present awareness in time. This class of emotions produces evaluations of expectancies that judge (future) possibilities as good or bad. Correspondingly, these emotions are derived from various *grades and mixtures of pleasure and unpleasure.*

b) Other complex emotions establish social hierarchies, contributing to rules for social coordination. For this class of emotions (labeled *'social emotions'*), the expressive aspects of the emotions using the face or the body are of particular importance.

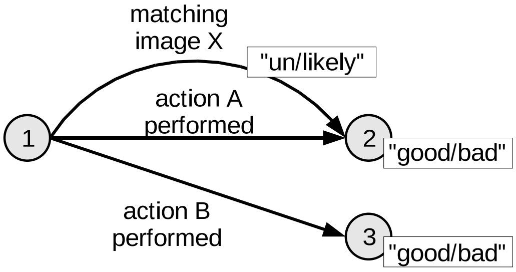


Figure 6.7: Transitions in the desire graph. The transition of one state to another can either happen (passively) by recognizing the next valid image ('matching image X') or (actively) by performing a certain action ('action A' or 'action B'). Transitions are evaluated by emotions concerning their likelihood (on a spectrum from *likely* to *unlikely*), and concerning their success (on a spectrum from *good* to *bad*; however, more specific emotions can also be used, e.g. a transition could lead to a *shameful* state).

The first class of complex emotions consists of affective reactions that *evaluate success or failure of a desire*, either prospectively or retrospectively. Thus, these emotions occur when processing a desire (Figure 6.7). Prospective emotions evaluate possibilities that are still to happen. The spectrum ranges from confidence and hope on the positive end, through stages of doubt and fear, to hopelessness and desperation on the negative end (Figure 6.8). Retrospective emotions evaluate the finalization of a desire. The spectrum includes joy, satisfaction, disappointment, sadness, and distress.[6] All the mentioned emotions can be viewed as emotional states consisting of combinations of pleasure and unpleasure (mostly frustration) in various degrees of intensity. All of them are based on an assessment of the probability that the desire with which they are connected will be satisfied. This assessment is based, most importantly, on previous experiences related to the desire in question (mainly the ratio of pleasurable experiences to the number of unpleasurable or the total number of experiences), and on the time how long the desire is already evoked but unsatisfied.

The second class of complex emotions evaluates actions that are of social relevance. With these emotions the psychoanalytic principle that the individual has always to be seen in relation to someone else is acknowledged (see Section 5.2). Functionally, these emotions foster cooperative behavior. Influencing factors are whether actions are self-initiated or other-initiated, whether they

[6] In the case of a 'negative desire', that is, the wish that something *bad* does *not happen*, the joy of 'wishfulfillment' would be referred to as 'relief'.

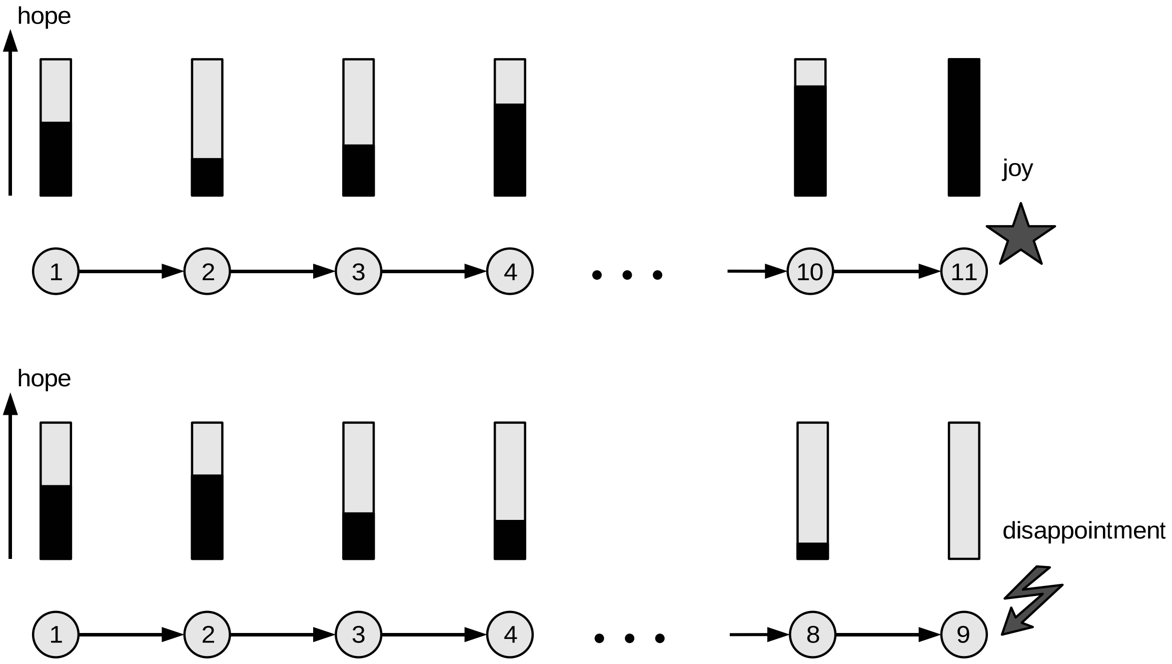


Figure 6.8: The dynamics of hope in the course of processing a desire.

are blame-worthy or praise-worthy. The emotions of this class themselves can again be positive or negative. This leads to emotions such as shame, and pride when judging own actions; or reproach, and respect and admiration when judging the actions of others. All the above emotions are related to social norms. These make up the contents of the `Superego`. The norms, together with the corresponding emotions, are 'learned' during infancy, for example, to feel pleasure when being judged positively by others.

In fact this learning makes up a contrast to basic emotions: Complex emotions are much more dependent on *acquired emotional dispositions* than basic emotions which are to a great extent predefined. Apart from the processing of a desire, the emotionally colored memories stored in the `Superego` and the `Episodic Memory` (which the former is a part of) are the main source for the evocation of a complex emotion. (Episodic) memories are not neutrally stored and remembered. Memories similar to the currently perceived image have associated emotions which are elicited when remembered, meaning that *remembering is* almost immediately followed by *re-experiencing* the associated emotion, including its physiological effects. Hence, emotions can be triggered by mental images just like by perceived images. With a given object or type of situation, possibly more than one kind of emotion may be associated, for example disgust and anger which would mix to something like contempt. The intensity of a complex emotion is influenced by the current perception, by contributions from evoked memories, and by contributions from activated desires.

6.6.3 Effects of Emotions

Emotions have several effects on the decision making process of an autonomous system (some of them having already been mentioned when discussing the `Basic Emotions` module). In particular, emotions contribute to the following tasks:

- Informing the self about important changes in the body and the environment

- Establishing an alarm system
- Establishing social coherence
- Supporting the pursuit of goals or their abortion in time in case they turn out to be destructive
- Setting up new goals or desires in order to 'fix' previous actions that have turned out to be not particularly successful
- Supporting learning processes to classify objects or other individuals, which, in turn, influences future interactions with them, for example increases or decreases the probability of the occurrence of such interactions
- Supporting focus on relevant perceptions as well as relevant actions
- Emphasizing crucial events such that they can be memorized without having to be experienced repeatedly (which could be already lethal for the system)

The stored experiences in the `Episodic Memory` include associated emotional values. Memories with high emotional values are remembered more easily and more often (compare with Section 7.2.3). In general, the processes at the level of the `Decision` unit are characterized by an increased interaction with the memory systems (in particular the `Episodic` and the `Semantic Memory`) compared to the processes at the `Pre-Decision` level. The memory interactions are also of improved quality, meaning that memories can now be accessed explicitly. This is particularly crucial for planning and inference capabilities, but also for the competence of performing reflections on affective states. In fact, it is widely assumed that direct accessibility together with explicit manipulability of memories is *the* essential feature of deliberative, conscious processing.

6.6.4 `Acting-As-If` Module

The `Acting-As-If` module produces 'thoughts' where the name of the module indicates that thinking is viewed (and correspondingly also defined) as a kind of *acting, only that there is no execution of the invoked motor commands.* Thus, thinking depends crucially on a mechanism to *inhibit* evoked action tendencies. In psychoanalysis, as has been discussed in Section 5.2.3, inhibition is considered to be a core function of the EGO. In the architecture, the `Decision Making Control` module of the `Decision` unit is responsible for inhibition.

The `Acting-As-If` module mainly serves two purposes:

Planning – The first one is thinking ahead or planning. This capability opens up a new space between a need or desire and its fulfillment. It depends on internal models about how things are supposed to go. Scripts for situations the system knows about are given by episodic memories that have been typified to scenario templates. Additionally, there is also the possibility that behavioral routines (ideally given in some symbolically encoded form) act as scenario templates. The difference between both kinds of templates is not fundamental. It just lies in the fact that routines in any case deal with *own actions*, whereas scenarios extracted from episodic memories can also deal with *observed actions of others*, or *events* that do not depend on actions at all (neither own nor other-initiated).

Both kinds of templates are used to expand desires or tasks into goals and sub-goals, and to evaluate the expected consequences of the planned actions. The evaluation process is done at every stage of the expansion. Note that the system can only *actively* pursue the currently aspired desire/(sub-)goal if it has own actions at its disposal that are suitable means to reach the desire/(sub-)goal. (Otherwise, it can only passively wait for 'favorable events' – including the actions of others – to happen.) Planning can be done in forward or backward direction. When planning forward, scripts are proceeded in their normal 'chronological' order (from the current state to the next (sub-)goal, and so on), otherwise the processing is done in reverse order (from the final desire/goal to the previous (sub-)goal, and so on). Of course in the forward case, given a certain situation in time, the system has to know which of the currently possible actions brings it nearer to the aspired future goal. Naturally, this will be more often the case for standard goals and such which do not need so much steps until fulfillment. In particular, when having to deal with long-term goals, it is favorable to combine a forward expansion with a backward expansion in order to get better and more efficient planning results.

In any case, expansion and evaluation, are supported by knowledge from the `Semantic Memory`, and by inference capabilities. Thereby, semantic facts and inference rules in particular contribute to the (explicit) judgement of the feasibility and probability of potential actions or events. These aspects, however, are also aspects that are judged (among others) by complex emotions (albeit only implicitly). In summary, expansion and evaluation are a mixture of logically derived contributions from the `Acting-As-If` module (the seat of planning and reasoning which is strongly interacting with the `Semantic Memory`), and of affective contributions coming from the `Episodic Memory` (which stores sequences of previous experiences in an emotionally rated way).

Anticipation – The second purpose of the `Acting-As-If` module, apart from advancing the fulfillment of desires, is to *constantly* experiment with potential future actions, and their consecutive costs and benefits in an *anticipatory* way. Thereby, expectations of what probably will happen next are built up. What is gained by such expectations is that the system can speed up its cognitive reactions by narrowing its focus of attention (its 'search space'), for perceptions as well as actions. Anticipation is not completely different to planning, rather an extension. It is discussed separately for two reasons. Once, it is not particularly carried out in the service of a currently active desire, but instead directly based on the scenario recognition process, and second, it is not solely focused on own actions. Actually, the algorithm constantly projecting ahead of what is supposed to happen next especially has to include thoughts about potential future actions *of others*. This capacity could be shortly referred to as *action understanding*. Hereby, in contrast to planning, the system does not know from the start the desires or goals at which action sequences of others are aimed at. Thus, the system needs an algorithm that enables it to guess the goals (*intentions*) of others from their observed actions. Such an algorithm requires – quite similar to planning – knowledge about the ends of action sequences because these ends can be interpreted as goals. Consequently, action understanding can be implemented in a similar way to planning: by initializing the scripts that are observed, within the system itself. Of course, the more inputs are just simulated (as it is gradually the case when reasoning about the intentions and beliefs of others and not just about those of oneself), the more insecure the predicted results will get.

Note that additionally to goals also the emotional states of others are inferred. This can be done because emotions are part of the scripts. Moreover, emotions are explicitly communicated via their expressive components.

To summarize, the power to *inhibit* the execution of evoked action tendencies leads to the possibility to simulate them offline – *acting-as-if.* This leads to a distinction between actual and potential actions. Knowledge about means/ends associations of action sequences is required. When making plans to fulfill desires and goals, the system can be referred to as *intentional.* Again, by running predictive simulations, the intentions of others can also be guessed in advance. A comparison between desired, anticipated, and actual ends of action sequences can be made. Such comparisons enable the system to constantly adapt its behavior to the world. When there is, additionally to the actual/potential distinction, also a distinction between own- and other-initiated actions, the system can build up a more than transient internal model of its capabilities, that is a model of it-*self* as an acting person.

Humans have the capability to simulate sequences of counterfactual inputs. The quality of plans and predictions derived by an intelligent system depends crucially on the length and the granularity of the sequences of acts/thoughts the system can project ahead. Both aspects, length and granularity, are boosted enormously when using a symbolic language with syntactic rules, a capacity that is possessed by humans. When routines and experiences are lifted from a sub-symbolic to a symbolic level, their elements can be more pronouncedly articulated and consequently more flexibly linked with other elements.

6.6.5 `Working Memory`

Working Memory is the place where the momentarily most salient external perceptions, desires, emotions, and thoughts come together (Figure 6.9). Thus, affective contributions are combined with cognitive contributions to create ‘a feeling about something’. The ultimate goal of this coupling is the release of an action command.

The contents of the `Working Memory` changes on a moment to moment basis. The question now is which chunks of information populate the `Working Memory`:

- First, the `Working Memory` contains the current external perception. This is already a construction where sensory inputs are complemented by stored knowledge from similar memories which are automatically evoked during recognizing a scenario.
- Second, these memories may activate associated emotions which can also gain entrance to the `Working Memory` in case they are above some given threshold.
- Third, the `Working Memory` contains the currently most salient desires and tasks (the ones which produce the highest tensions). They are the starting point to *actively* trigger the formation of a plan how to satisfy the specific desire or task – unless they do not already produce a very strong action tendency. The building of (longer) plans is done by the `Acting-As-If` module (see Section 6.6.4) which is responsible for a big part of the processes of the `Working Memory`.

Finally, as has been stated above, the `Working Memory` should release an action command. Most likely, the various elements of the architecture will have generated more than one action tendency. The following sources of action tendencies have been described:

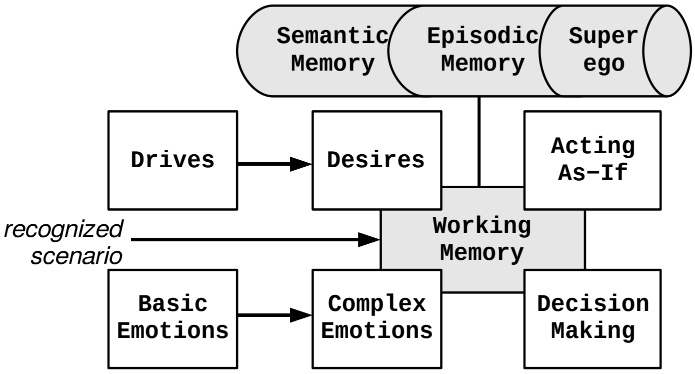


Figure 6.9: The `Working Memory` is a blackboard used by several modules. Correspondingly, it can be populated by perceptions, memories, desires, complex emotions, action plans, `Superego` rules.

- The `Drives`, and `Basic Emotions` module of the `Pre-Decision/Id` unit, producing affective action tendencies based on a quick and rough evaluation of input stimuli
- The `Desires`, `Complex Emotions`, and `Acting-As-If` module of the `Decision/Ego` unit, producing refined affective action tendencies based on appraisals that take into account emotionally rated previous experiences, and cognitively derived action tendencies based on logical inferences
- The `Superego`, containing behavioral rules that represent social ideals or standards; such rules can block action tendencies which are evoked by the other sources

Usually, the action tendencies produced by the `Id`, the `Superego`, and the `Ego` will be in conflict with each other [Fre23, Fre89] such that 'a decision has to be made' amongst them. This is a dynamic battle between the different action tendencies. Generally, actions are motivated by the goals that can be reached with them (be it the fulfillment of a need or the accomplishment of a task). A current goal may require the execution of several steps of actions. The currently pursued goal – and thus sequence of actions – depends on the height of the tension being potentially discharged by the actions (to be more precisely, not on the absolute height but on the relative *change* that is possible), and on the number of action steps it takes until this happens. The applied selection mechanism (how it may look in detail) should strongly favor actions that are associated with the highest tensions, and actions which do not need much steps to discharge a tension.

Moreover, a parameter is introduced that varies the relative strength of the `Id`, `Superego`, and `Ego` unit on the action selection process (Figure 6.10). Whether an impulsive action from the `Id` – which can occur any time during the execution of a longer action sequence – can be inhibited by processes of the `Ego` could differ between different instances of the cognitive architecture, leaving room for optimizations. The same applies to the influence of the `Superego` which delivers antagonistic action tendencies for some of the desired or planned for actions.

The switch from an old goal to a new goal – and thus behavior – is also influenced by emotions. Especially the `Complex Emotions` module not only evokes direct action tendencies, but also outputs that strengthen or weaken some of the other action tendencies (compare with Section 6.6.2). Generally, positive emotions support the prolongation or repetition of ongoing interactions, negative emotions lead to the abortion or change of current system/environment interactions. Actions with

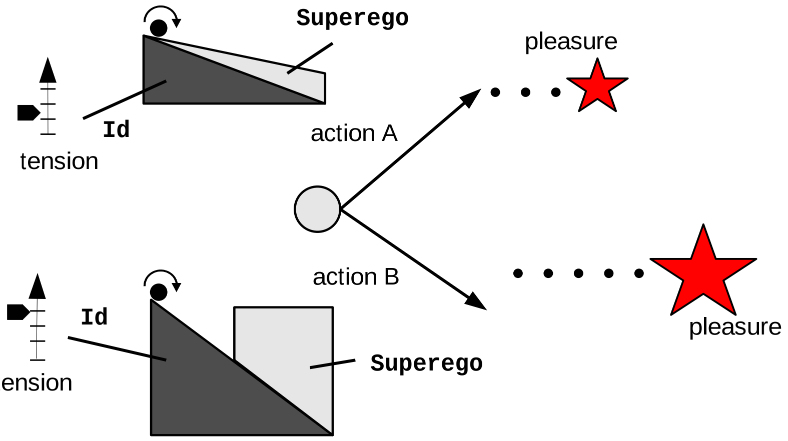


Figure 6.10: Graphical illustration of the fact that the relative strengths of `Id` and `Superego` (metaphorically depicted in the form of ramps and blocks) can vary depending on the action tendency 'to be judged' by the `Superego`. Which action tendency in the case of several competing ones is finally selected for execution not only depends on the tension (and the potential future amount of pleasure) associated with an action tendency, but also on the strength of an eventually opposing `Superego` rule.

positive associated emotions are looked for, actions with negative associated emotions avoided. A feared for negative outcome of an event, or a situation where none of the tensions can be released, may activate a 'defense mechanism'.

6.7 Actions and Behavior

Actions have influences on external circumstances and on internal states. They are built out of action primitives. These are system- and application-dependent. The `Action Sequencer` module handles sequences of actions (Figure 6.11). On the one hand, it loads them from the `Procedural Memory` where they are stored. On the other hand, when occurring repeatedly, it discovers new action patterns (*routines*) and transfers them to the `Procedural Memory` from which, in future, they can be activated as a whole, thus forming new kinds of routine behavior.

Activated routines are carried out by the `Execution` module. It handles – in strong cooperation with the `Procedural Memory` – the physical aspects necessary to execute the actions belonging to the system's action repertoire. Thus, the specific look and content of the `Execution` module – and how it cooperates with the `Procedural Memory` – depend greatly on the 'hardware' of the autonomous system or robot, that is, on how the system is generally embodied in its environment, and on how this general embodiment is encoded in appropriate information structures controlling the actuators of the system to produce real physical actions in space and time. In principle, there is a similar problem like on the perception side, only in reverse order: From a high-level symbolic and abstract representation of behavioral commands it has to be determined how to control a (potentially great) number of actuators. It may be speculated that the nearer one gets to the actuator level, the more preferable the usage of a distributed representation may be. Such a usage might correspond well with the largely implicit character of procedural memory. However, an investigation of the question of the sensorimotor coordination of movements and how to implement such, f.i. in a given robot, is beyond the scope of this work.

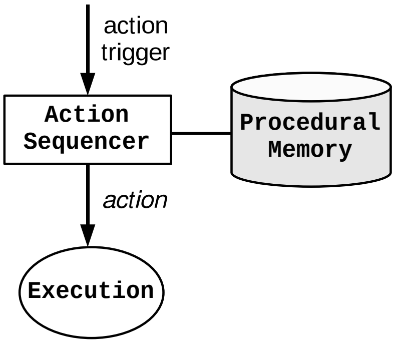


Figure 6.11: The action sequencer feeds and reads the procedural memory to constitute and to make use of routine actions.

This work deals with an abstract conceptualization of the behavioral levels and an analysis of how abstract behavioral representations interact with the other modules and levels of the architecture. As a starting point, it is important to note that actions (unless completely arbitrary and useless ones) are always *means* to achieve *goals*. Thus, the used data structures must provide a link between actions and the goals that can be achieved with them. Such links ('means/ends associations') can be chained. For example, crossing the street can be a means to get to the other side of the street, which someone may want to do in order to enter the railway station over there, in order to buy a ticket such that he can take a train, in order to get from A to B, to visit his parents.

Knowing about the sequential structure of routines enables a system to anticipate ends of action sequences. This is an important capability for two reasons (compare with Section 6.6.4.

The first one is planning (see Section 6.6.4). However, to be useful for planning purposes, it is not effective for routines to be just implicitly given (as this would be the case when routines were solely represented in distributed, sub-symbolic structures). It has advantages it there is also a form that allows routines to be explicitly addressable, currently most easily implemented with some kind of symbolic representation. The crucial point is that it has to be possible to separately address – and in the following also manipulate – parts of a routine such that these parts can be newly combined and flexibly linked with other elements of the architecture. As has already been described, such processes are carried out by the `Acting-As-If` module, the main module for planning, in cooperation with the `Working Memory`. The latter acts as a common workspace where different elements can meet, be it in a cooperative or conflicting way.

Note that in planning mode, starting from a desire (that is, a goal-like state), the system has to search for actions that are suitable means to reach a currently aspired desire (goal).

The second reason for the necessity of knowing about means/ends associations is that they are a prerequisite to anticipate the behavior of others. In this mode, agents can guess from perceiving the actions of others about the goals they are aspiring to. This is very important for the generation of appropriate context-dependent behavior.

Knowing is not enough, we must apply. Willing is not enough, we must do.

Johann Wolfgang von Goethe

7 Implementation

The proposed cognitive architecture needs to be implemented in order to evaluate its power. However, it is impossible to achieve a 'final' implementation of the proposed cognitive architecture in one step right from scratch. Thereby, 'final' refers to an implementation that can really produce behavior qualitatively comparable to that of humans in terms of understanding *and* feeling. There are mainly two reasons for this.

First, the architecture is very comprehensive, supposed to include, at least in principle, all the affective and cognitive capacities involved in human decision making and acting. However, every functionality, every module, and all the relations within and between the modules require a profound analysis on their own.

Second, although the neuro-psychoanalytic approach combines physiological and neurological data with psychoanalytic insights and concepts – and thus takes into account both objective *and* subjective knowledge about the functioning of the human mind – there are still many aspects that remain undetermined. Thus, the designer of the technical system has to make choices between several possible alternatives, for instance when it comes to model the data structures or algorithms to be used.

To tackle the above problems, the chosen strategy within the ARS project is to successively produce more and more refined and improved implementations of the proposed cognitive architecture in an *iterative process* during the coming years. A starting implementation which represents a first step towards the final version has been worked out and programmed. More specifically, artificial creatures called 'Bubbles' have been invented and put in a simplified environment. This is because according to the principle of embodiment the cognitive architecture cannot be implemented without the specification of the environment where the autonomous systems equipped with the architecture have to prove themselves (their 'ecological niche' or field of application).

In the following, first the question of distributed versus symbolic representations is again shortly revisited. Second, the datastructures of the proposed cognitive architecture are discussed. Although the solutions used in the current implementation are presented, the section is kept quite general, highlighting the aspects that have to be respected by *any* future implementation, regardless of how it may look in detail. Thereafter, the *Bubble Family Game* (BFG) is introduced. Finally, referring to the environmental setting of the BFG and to the data structures and modules as realized in the prototypical implementation, all the processes running in the cognitive architecture, from input to output, are described. This serves to *qualitatively* illustrate the incorporated functional capacities of the cognitive architecture.

Note that in this foundational work of neuro-psychoanalytic design no quantitative results are presented to demonstrate and analyze the power of the proposed cognitive architecture (although some of the cited references about the ARS project already include such results). To summarize, details of the starting implementation are included for the following two purposes: to generally discuss implementational issues, and to deliver a picture of the architecture 'in action' (in contrast to the structural description of the architecture given in Chapter 6).

Note further that the architecture has to be implemented using familiar elements of IT – like numbers, datastructures, algorithms, modules, and interfaces. It may turn out that with currently available platforms a sufficiently performant implementation will not be possible. This is however not assumed a principal impossibility, rather one that – should it arise – is supposed to be overcome in the not so distant future with a new generation of hardware and/or software components and paradigms.

7.1 From Sensor Data to Symbolic Representations

The decision about which kinds of representations to use is a very fundamental one.

In biology, as discussed in Section 2.3, organisms show a mixture of physical and informational processes. Just like the outer world in which they are embedded, organisms are made of physical matter. In both cases, the dynamics of the matter is given by physical laws. In Section 2.3.1, it has been stressed that every informational process needs a physical carrier and every physical process has an informational aspect. However, because of the complex organization of organisms, the physical processes happening within organisms gain an informational dimension that goes beyond that of not so organized matter. In fact, the more organizational levels there are, the more 'dense' the processed information can get. This means that there may be representational levels where the processed information deals less and less with low-level matter-specific details but more and more gains the power to encode higher-level, more cognitive phenomena. The higher the information processing level, the lesser the influence of the physical carrier can be assumed.

When technically creating an intelligent autonomous system, the balance lies more on the informational side than on the physical because in this case physical interactions are in an increased manner and already starting at a very early processing stage replaced by informational algorithms that try to functionally capture all the effects of the real physical interactions. In the case of autonomous embedded and embodied systems, real, physical interactions mainly remain relevant between sensors or actuators on the one hand, and the environment on the other hand. There, physical forces and laws directly apply. Consequently, these processes possess an intrinsic, natural meaning and can thus provide the foundation for all the other, more abstract meanings created and processed within the architecture (in the sense of an anchor).

Within the cognitive architecture, it is assumed that, at the lower levels, directly after the sensor/environment respectively sensor/actuator interfaces (mainly within the `Perception` and `Execution` module), techniques like neural networks or statistical methods should be used. From the resulting distributed representations, symbolic representations should be extracted and used throughout most of the rest of the cognitive architecture. The main reason for this choice of representations is that distributed representations are very adaptive but badly accessible and not so modularly combinable like discrete ones. On the other hand, when pinning down 'information-carrying' roles on specific abstract, discrete elements right from the start in a completely prespecified way, it is not possible to catch emergent meanings of an ongoing dynamics (for a discussion of distributed, connectionist versus localist, symbolic representations see Sections 2.1 and 2.4).

There are several possibilities how to derive symbolic representations out of distributed ones in an adaptive way, that is without having the symbols pre-specified by a human programmer. One possible example is presented in [Bru07]. In this work, also belonging to the ARS project, the described algorithms to get from real-world sensor data to semantic interpretations are based on statistical models. More specifically, based on a hierarchic hidden markov model framework, symbols are created out of distributed and diverse sensor data. The envisaged field of application is building automation. The goal is a system that can learn in an unsupervised way to symbolically classify previously unknown situations and scenarios.

Another approach of how to learn meaning without explicit instruction is presented in [Lar03]. There, symbols are automatically extracted from sensory experiences based on the regularities that are experienced in the sensory data. By grouping what is similar and what is not similar, descriptions are built (compare with Section 2.4.2). The resulting clusters can be activated by activating a member of the cluster. Dependent on the frequency of coincidental activations, associations between clusters are formed. In sum, low-level perceptual information is transformed in an adaptive way into a higher-level, symbolic knowledge representation scheme with intrinsic meaning. This is possible because an extraction of the meaning embedded in the context of the information space.

An important data structure of the ARS-PA model, as described in the following section, are *images*. They can be viewed as an intermediate data structure between a dynamically coupled, analog representation, and a decoupled, symbolic representation. (Their role is similar to that of *metaphors* in the case of language.)

7.2 Data Structures

In the following, general remarks concerning the data structures to be used in the architecture are made. Additionally, it is referred to their current realization within the *Bubble Family Game* simulation test bed.

7.2.1 Feature Elements: Smallest Data Units of 'Psychological Dimension'

Above as in Section 6.4, it has been described how, in a hierarchical symbolization process, sensor readings are more and more condensed until a level is reached where whole objects or situations are represented by symbols. These symbols, in contrast to the representational constructs of the symbolization process, already possess a 'psychological dimension', that is, they represent entities which can be the elements of psychological theories. Before further classifying the symbols of 'psychological dimension', it has to be specified what to understand by a situation (because these symbols shall represent situations). In the context of our neuro-psychoanalytically-inspired architecture for autonomous systems or agents a situation is defined as follows:

> **Situation**: a composition of all information that characterizes the momentary state of an autonomous system or agent (as perceived by itself)

The elements this composition is made of (and thus the symbols of 'psychological dimension') are referred to as *feature elements*. They can belong to different categories. This is because the information characterizing a situation can come from different sources: a) the internal state as currently perceived, b) the state of the environment as currently perceived, and c) the actions the system or agent currently exerts on the environment or has a tendency for. In the present implementation of the architecture, there are the following different **feature element categories**:

- *Template images* – representing the perception of the environment
- *Drives* and *desires* – representing internally perceived bodily signals, and internally arisen states of motivation or urges to do something
- *Basic* and *complex emotions* – representing evaluations
- *Actions* – representing currently executed actions, or just tendencies or dispositions for specific actions

7.2.2 Images

In the previous section, it has been tried to technically define the colloquial term 'situation'. However, situations as such are *not* an element of any of the architecture's memory systems. Further conceptual refinements are made to get to the used data structures.

A central data structure of the architecture are *images*. They have already been defined in Section 6.4. The definition there can be extended to include template images (TIs):

> **(Template) Image**: a (predefined) collection of (internal and external) features making up an object, or a situation (or only parts of both)

Several remarks can be made. First, like situations images are composed of feature elements belonging to the different categories as described above. An example of an image could be: a person sitting in front of me, shouting out loudly, and looking angry at me. Thus, the term 'image' is very general. It refers to visual information as well as to other kinds of sensory modalities. Second, an image can also be a collection of more abstract pieces of information which are already the result of evaluation, condensation, or other kinds of processes occurring in the architecture. In this case, the content represented by images can become progressively more complex. Correspondingly, images not only appear in the perceptual realm as *perceived images*, but also as *mental images* when originating in higher processes of the architecture (like remembering or planning). Third, template images are predefined collections or snapshots of what can be perceived or remembered. They are included permanently in the `Image Memory`. New combinations of symbols (representing previously unknown objects or situations) can be constructed. Under which conditions their initially transient status turns into a permanent one is addressed in the following section.

7.2.3 Extending Time: Events, Episodes, Scenarios

Situations are continuously streaming into the system. Theoretically, the time gap between two situations can be infinitesimally small. The concept of perceiving/recognizing situations by comparing them with known template images (which are necessarily detail-omitting abstractions with the character of samples) brings already a discretization of the continuous flow of situations. Also, what is newly stored (in the `Episodic Memory`) as experience and later on remembered are only *extractions* of the continuous input stream.

According to Tulving, the prototypical units organizing episodic memory are events. In Section 6.4, an event has been defined as follows:

> **Event**: a happening that arises at a particular time when something in a situation changes significantly

The data structure of events is that of images. Thus, like images, events consist of the feature elements listed above.Additionally, events might get an *intensity* (called salience value in [Gru07]) and an *emotional tone* evaluating the importance or impact of an event as a whole. See Figure 7.1 for a graphical illustration.

Events happen within situations, that is, they change states where little or nothing has happened for a while. They always have a beginning and an end, although beginning and end may be very close together (compare [Tul83, p. 83]). Tulving decomposes situations in a *setting* and a *focal element*. The setting corresponds to the static aspects of a situation that change only slowly, whereas the focal element is given by what (suddenly) happens within such a static setting. Events can be merely caused by the environment, or they can be caused, or at least influenced, by the system's or agent's own actions, or by the actions of another system or agent (to recognize the distinction between these possibilities is not always trivial for a system). To store new experiences in the `Episodic Memory`, the input stream has to be monitored and significant changes in some of the features of a situation have to be detected.

Analog to template images, there are stored *sequences of events*, acting as templates to recognize 'extended situations', that is such, that go beyond a simple point in time. They are called *scenarios* and defined as follows:

> **Scenario**: a template of a sequence of events (and states in between)

In the current implementation, scenarios are implemented as directed graphs consisting of states and transitions, the latter being triggered by template images representing perceived (or remembered) events or actions. A graph is activated by detecting its initializing image within the stream of input images. By traversing the graph, scenarios are perceived and recognized. Usually, every moment, one or more of such scenario recognition processes (which can run in parallel) are in an 'activated' condition, that is, in progress. One event can belong to several scenarios. Recognizing the current scenario provides the system with a 'basic awareness' of the present situation and context it finds itself in. Thereof, most further processes of the architecture are built, like for example the encoding or retrieval of specific *episodes* which are defined below.

The detection of an event can cause the *encoding of an experience*. Usually, a stored experience not only consists of a single event, but of a sequence of events and states, referred to as *episode*:

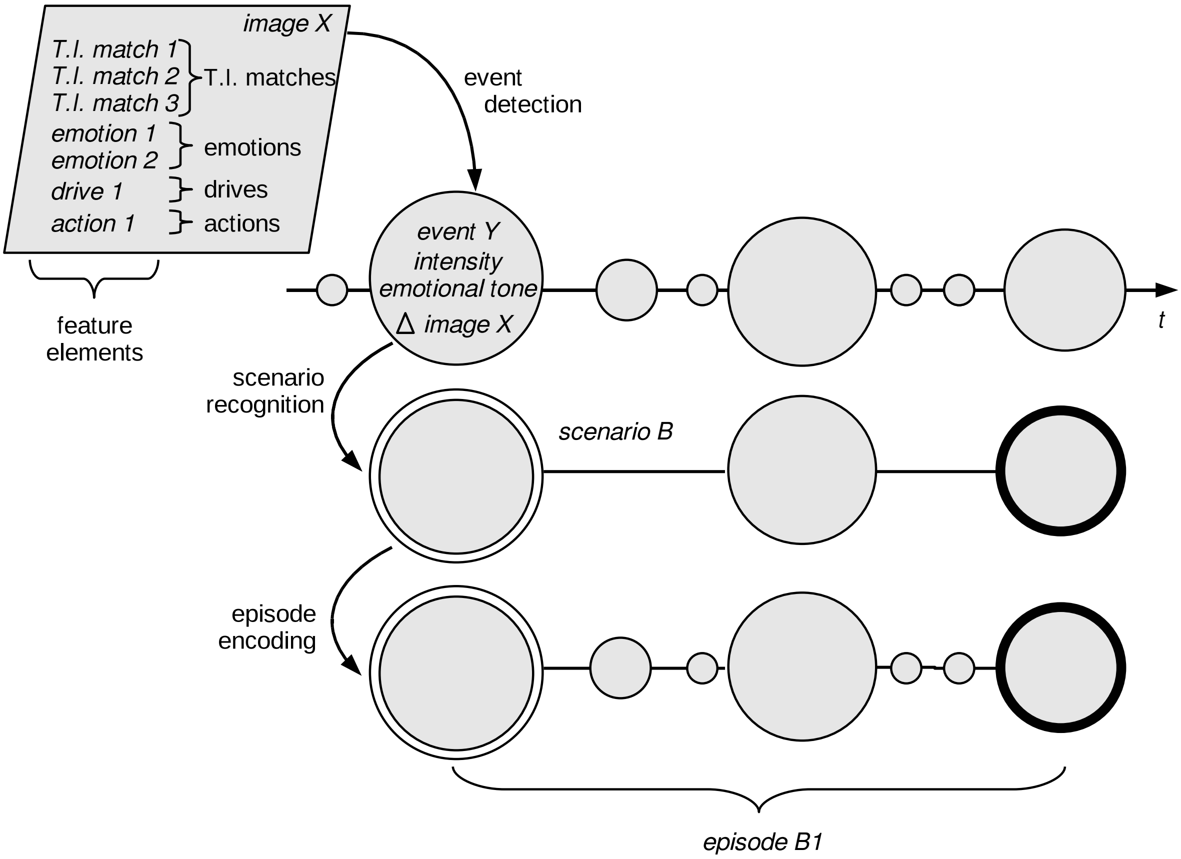


Figure 7.1: Illustration of the relations between feature elements, images, events, scenarios, and episodes. Individual values are in *italics*, a double circle indicates a start state, and a thick circle an end state. The detection of *event Y* is based on a significant change in *image X*, and it starts the scenario recognition process of *scenario B* because *event Y* is the start state of this scenario. If the intensity of *event Y* is high enough, an episode encoding process is also started in parallel. In this process, also events that happen in between the events required by *scenario B* are registered. *Episode B1* is an instance of *Scenario B*.

> **Episode**: a sequence of events (and states in between), including sequences of own and/or other-initiated actions

Episodes are realizations or instances of scenario templates. Note, that scenarios do not fully determine episodes. They only deliver the minimum structure of events making up a scenario, whereas there may be many details that are not fixed and thus can vary significantly. This means that one and the same scenario will usually have many different episodes as realization. Note further, that scenarios represent semantic knowledge. They are standard scripts (schemata) of previous experiences which have lost their individuality in an extraction process, now denoting general information.

Episodes have a context, a content, and an impact.

Context – The context of an episode is given by the scenario which it is an instance of. Episodes are grouped by this context. Activated episodes belonging to a certain scenario are sorted according to the time it takes until they are completed.

Content – The content of an episode consists of the sequence of events and states that make up an episode. One and the same event can belong to different episodes, and episodes can be nested and overlap in time.

Impact – The impact of an episode is derived from the emotional values of the events belonging to the episode, in particular the final emotional value at the end of an episode has a strong influence.

Episodes are managed by the *Episodic Memory* which handles their encoding, storage, and retrieval (see [Gru07] for details of the design of an `Episodic Memory` within the ARS-PA architecture). According to Tulving: *'Recollection of an event, or a certain aspect of it, occurs if and only if properties of the trace of the event are sufficiently similar to the properties of the retrieval information'* [Tul83, p. 223]. Retrieval can occur *spontaneously* or *deliberatively.*[1] Thus the *Episodic Memory* must provide the following functions:

- Detect when an agent has experienced something 'worth' remembering
- Initiate encoding
- Manage stored episodes, including their potential *decay*
- Return experiences deliberatively searched for by a *cue*
- Provide unbidden (spontaneous) retrieval

Note that experiences can be sequences of events (that is episodes), or just single events. Sources for retrieval cue formation are the processes that are currently running within the cognitive architecture, especially the ones within the `Decision` unit. In the present implementation, events, subsets of features, and additionally scenario templates can be used as retrieval cues, thereby realizing retrieval queries based on behavioral, emotional, goal-oriented, spacial, and temporal information of previous experiences (see [Gru07, p. 67]). Especially the current emotion acts as a strong cue for spontaneous retrieval. As far as temporal queries are concerned, it is noteworthy, that it usually cannot be queried for a specific, absolutely indexed moment in the past because what is stored is just the *chronological ordering of events*, not the absolute time of their occurrence.[2] When remembering an event that is part of an episode, it can be further cued for preceding or succeeding events. With this mechanism, a whole temporal sequence can be reconstructed. This is used for example for making action plans.

The return data structure of recalled events potentially includes elements from all the possible feature categories (T.I.-matches, drives, desires, emotions, and actions). Repeatedly recalled events (and herewith also the episodes they are part of) are 'strengthened' such that they will be even more often remembered in future.

[1] Note that neither the term *spontaneous retrieval* nor *deliberative retrieval* is to be misinterpreted as requiring a kind of homunculus.

[2] The usage of absolute time-stamps is not considered psychologically plausible. When not trying to 'imitate' a human but 'just' building an intelligent machine, time-stamps might be introduced.

7.2.4 Drives and Desires

To technically model a drive or a desire, the following aspects must be provided for:

- Activation – When and how is a drive activated?
- Intensity – How is its initial intensity determined? How does its intensity change with time?
- Aboutness – Which kind of need (motivation) does the drive promote?
- Duration – How long does a drive last?
- Abortion – When and how does a drive end?

According to psychoanalysis, *drives* are enclosed in the ID and defined as border phenomenon between the body and the psyche (see Section 5.2.2).

> **Drive**: *'[..] a measure of the demand made upon the mind for work in consequence of its connection with the body'* [Fre15c, p. 214]

Drives are not only associated with bodily need states, they also produce the arousal and energy (tension) that raises an individual's interest in the surrounding world, because there, the objects to fulfill the needs can be found. The neurobiological basis of drives has been investigated intensively by Panksepp who speaks of a SEEKING system (see Sections 3.2.4 and 6.5.1).

In the current implementation, *drives* are modelled with the following elements:

- *Resource-related variables*
- *Tensions*
- *Associated action tendencies*
- *Relations to emotions and desires*

With these elements, the above mentioned aspects are addressed. Each of the needs of the autonomous system or agent gets one or more *associated variables.* These variables are abstractions which represent the state of internal resources the system or agent must keep within certain ranges in order to survive. Several need-detector mechanisms constantly monitor the internal variables. In case of a pending imbalance, they activate a corresponding drive, which in turn, initiates appropriate counter actions, for example some kind of search behavior.

The tension of a drive depends on its deviation from a set-point. Different drives can compete with each other. The aboutness of a drive is currently largely predefined (compare the discussion in Section 6.5.1). As output, the `Drives` module evokes action tendencies. It can also influence emotional strengths, and it can lead to the activation of desires.

The execution of an action tendency evoked by the `Drives` module can be inhibited by the `Decision` unit. As has been said, drives often initiate sequences of actions (routines). Usually, a change in behavior is evoked by external factors, like for example, perceived external stimuli. For example, when the searched for object is found, seeking behavior is stopped and consummatory behavior is initiated, leading to the satisfaction of the need. This discharges the need's tension. The associated variable gets back towards its set-point. Pleasure is experienced.

Like the satisfaction of a drive, the fulfillment of a desire brings about pleasure. According to psychoanalysis, the seeking of pleasure is the most fundamental motivation for actions. In Section 6.6.1, the following definition of a desire has been given:

> **Desire**: the urge or wish to repeat a previous experience of pleasure

Thus, desires have motivational (because of the urge), emotional (because of the pleasure), and cognitive components. The latter come into play when reflections about the fulfillment of the desire are created.

Like with drives, the aspects activation, aboutness, tension, duration, and abortion of a desire have to be determined. These aspects have to be dealt with by determining relations between the following components making up a desire according to psychoanalysis (see Section 5.2.1):

- Internal *needs*
- External *objects of desire* (things or persons)
- *'Action plans'*, that is, representations of
  - the system itself
  - in the course of satisfaction of the desire

Much more as in the case of drives, the objects of desire are not predefined from the beginning, but learned. This may happen rather early, during a 'training phase' of the system, or only rather recently.

In the current implementation, desires are modelled as *directed graphs*, similar to scenarios. This is no wonder, since a desire aims at the *revival of a once experienced pleasurable situation* (or a sequence thereof). The graphs can be viewed as essentially being *action plans*, all the aspects characterizing a desire being contained either in the states or in the transitions. (In the case of goal-expansion, several such action plans can also be recursively linked to each other.

The processing of a desire starts with its activation because of an (actual or only virtual) encounter with the need and/or object of desire. Generally, triggering can happen because of

- what is currently happening, and thus perceived, or
- what is currently remembered, imagined, fantasized, etc. in the course of some further processing in the cognitive architecture.

Thereby, a desire can have its origin in an *external circumstance*, like dropping by an object of desire, or in the system's *internal state*, like having an activated drive that evokes a memory of a situation where this drive is satisfied. Also, an emotional state, like being ashamed, may evoke a desire. See also the discussion of this topic in Section 6.6.1. In any case, a memory of a once pleasurable experience popping up in the course of some processing may trigger the initialization of a desire (when leading to a strong enough pleasurable, tension-discharging end state). Several (conflicting) desires may be active in parallel, at least 'in the background'. This means that during an evaluation step of the `Decision` unit, only the desire with the highest tension can access `Working Memory`, however, there may be others with 'non-zero' tension and the height of which may change depending on the ongoing processes.

Usually it takes a sequence of states to reach the fulfillment of a desire. The sequence of steps leading to the satisfaction of a desire can either be completely known to the system. In this case, the action plan to reach wishfulfillment is a given blue-print, a once successful sequence of images stored in the `Episodic Memory` that can again be followed. The other case is that some parts of the action plan leading to wishfulfillment have to be worked out. This can happen by the `Acting-As-If` module on explicit request from the `Working Memory` when the latter is processing the desire that is currently 'residing' there.

When working out *action plans*, desires are expanded into sub-goals ('goal-expansion'), leading to a recursive concatenation of action plans. For example, the desire to see a specific movie may require to get to the cinema, which, in turn, may be either achieved by using a car, a bicycle, or public transportation. Before entering the screening room, a ticket has to be bought, which first may require to get some money, etc. To enable the formation of (new) action plans, actions and events have to be *tagged with the goals that can be reached with them*, enabling the concatenation and nesting of goals and means to reach them.

In general, the states of a *desire graph* are represented by images (as in the case of *scenarios*), however, the edges representing *transitions* can either be triggered by

- (Template) images – representing passively perceived events towards wishfulfillment
- Actions – representing actively initiated steps towards wishfulfillment

The transitions can also be made dependent on a *context specified by by a given scenario.* In each stage of the processing of a given desire, the following parameters can be newly determined:

- *Tension* – indicating the current intensity of the desire
- *Likelihood of success* – indicating the current chance of reaching the satisfaction of the desire
- *Complex emotion trigger* (hope, resignation, ..) – influencing the further pursuit of the desire
- *Decay, timeout, and abortion conditions*

The tension of a desire may vary (increase or decrease) while it is processed, due to, for example, a changed emotional state, or a newly recognized scenario indicating a changed external situation. The likelihood of success is influenced by the number of steps the system is still away from wishfulfillment[3], however not necessarily in a linear way. The basic operation to determine the likelihood of success is of course to put the number of successful cases in relation to the number of overall cases. Thereby, 'success' is first of all given by emotional ratings stored in the `Episodic`

[3] In the work, I do not distinguish between the notions 'desire' and 'wish', see however the comment in Section 6.6.1 on this issue.

Memory. Eventually (if there is for example enough time), this emotion-mediated basic rating can be combined with 'more rational' evaluations. In any case, the algorithms to be used have to derive the likelihood value either from information stored in the Episodic or in the Semantic Memory. The likelihood of success as well as the complex emotions caused by a desire have influence on the parameters of the desire in the subsequent stages of desire processing. They can for example intensify a desire, or accelerate its abortion. They can also determine which path to pursue in the (likely) case that the action plan has several alternative branches.

7.2.5 Basic and Complex Emotions

Though being closely linked to internal (bodily and brain) states, emotions are typically triggered by environmental events [Pan05, p. 32]. This can directly result in a direct (instinctual) reaction, but also for instance in a modulation of the perception process. Thus, emotions not only inform about the inner state, they also do the task of linking the inner state with the outer world. This linkage is the reason why emotions are intrinsically *evaluative* (see Section 3.2).

Neurologically, the basis for the linking are two 'maps' of the body within the brain, one representing the inner state of the body (being the source of emotional states), and one representing the sensorimotor apparatus of the body, that is, its outer anatomy which makes it move around and act in the world. In the brain, there is an area where the two maps come together, thereby *'giv[ing] the emotion-generating part of the brain direct access to one of its action-generating mechanisms.'* [ST02, p. 111]

Having the task of bringing together perceptions and actions in an evaluative way, emotions are to be thought of as a *bundle of processes*. In particular, emotions also influence cognitive activities in the higher brain areas, and in turn are influenced by them. See Figure 3.3 for an overview of the involved processes.

In the architecture, the data structure of a *basic emotion* includes the following elements:

- *Type* – pleasure/unpleasure or more specific
- *Intensity* or *tension* – influenced by the desirability, likelihood, etc. of an event
- *Cause* – mainly to determine whom to evaluate (self or other)
- *Associated action tendencies*
- *Decay* and *abortion specifications*
- *Connections* to other elements of the architecture

A *complex emotion* additionally contains:

- *Context* – given by episodes which are rated by the emotion
- *Connection* to basic emotions and drives

Note that the aspects *type* and *tension* of a basic or complex emotion directly model its emotional quality, whereas the other introduced aspects serve to model possible relations between emotionally afflicted objects and the self.

By and large, the various types of emotions can be either classified as 'good' or 'bad'. Solms and Turnbull write that *'degrees of pleasure and unpleasure calibrate the basic qualitative range within which the 'sense' of emotion is experienced.'* [ST02, p. 108] Thus, an overall state of pleasure and of unpleasure can also be calculated. Apart from this, the various emotions are handled distinctively.

Emotions can result out of unsatisfied drives or desires (e.g. getting angry when being hungry for too long). They can also come from the (spontaneous) remembering of images or episodes. Usually, an evoked emotion can sustain for some time after the triggering event has passed (leading to *moods*). Thereby, a previously induced but lasting emotional state can influence emotions that are triggered later.

7.3 Protoypical Implementation

The proposed architecture describes a psychologically inspired cognitive model from a high-level perspective. It gives little or no advise, how to implement it in a given machine (computer, robot, etc.). Functionalities and data structures are arranged within several functional modules because this is what one gets in a relatively straightforward way when translating the psychoanalytic model into a technical model. It is not yet clear, however, if this arrangement is the he most easiest to implement or the one with the best performance on a given platform. The future of the project will very likely consist of several attempts of bringing the architecture to reality in order to prove its usefulness. The worked out implementations might suggest alternative ways of describing the model, leaving even the strict modular way beside.

Thus, the development of the cognitive architecture is considered an iterative task, carried out via implementations, with the current implementation being the first step. The prototype of the ARS-PA architecture, called *Bubble Family Game* (BFG), has limited but extensible features and is continuously improved. The BFG is not supposed to fulfill a particular engineering task (like serving as an alarm system in an office building or as a service robot in a private home). The fictional environment where artificial creatures have to survive provided by the BFG is intended to serve as an object of study. As has already been mentioned, the proposed cognitive architecture cannot be implemented and analyzed without first specifying an environment where the system has to perform its activities. This is because the specific environment has a critical influence on sensory images, actions, tasks, desired behaviors, etc. Below, the environmental setting of the BFG will be described together with the cognitive system the Bubbles are equipped with, the latter being given by the first, prototypical implementation of the proposed cognitive architecture. The presented descriptions serve to qualitatively demonstrate the potential of the architecture when 'in action'.

As, at the moment, it is not yet the goal to equip a real robot or another machine with an 'ARS-PA program', each of the following components of the BFG is simulated:

- Body of the artificial autonomous agents
- Environment

- Mind/mental apparatus of the artificial autonomous agents

The body of the artificial agents is a rather simple one. As there are no worth-to-talk-about physical dimensions and structures considered, the creatures are called 'Bubbles'. Despite its simpleness, the body of the creatures (as explained in Section 3.1.2) is extremely important for the occurring drives and emotions, for the grounding of the symbols, and for the development of context-dependent appropriate behavior in general.

The Bubbles are equipped with the following bodily features:

- Eyes and a proximity sensor
- Feet (or some other means of movement)
- Ability to consume energy sources ('eat')
- On-board energy storage

The process of symbolizing the sensor values into meaningful chunks of knowledge is considered as already done. The 'symbols' the simulated Bubbles get are like <`energy source in front`> or <`other Bubble on the left`>. Bubbles also have a homeostatic level, which is again relatively primitive. A kind of battery must be kept within a certain level. Living, walking etc. drains the energy level. To refill it, there are energy sources that can be used.

Similar to giving the artificial agents a body, it is also necessary to put them into a setting that stimulates their senses and that allows for interactions with the surroundings, with other simulated Bubbles, or even with other autonomous agents like humans. Figure 7.2) shows the environment of the prototypical implementation, a

- 2-dimensional, finite playground, with
- a number of different energy sources and objects, and
- a number of other Bubbles.

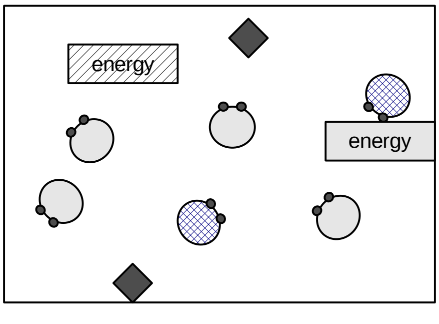


Figure 7.2: The Bubble simulation environment. Artificial autonomous creatures ('Bubbles'), energy sources, and obstacles are placed into a 2D-world.

The setting might look trivial, resembling existing artificial life games, but it is necessary to offer stimulation of a certain analyzable and feasible complexity to the Bubbles. Putting several Bubbles in the BFG environment already allows the definition of a great variety of rather challenging tasks for the Bubbles.

The third and actually most interesting part of the simulations is the cognitive system of the Bubbles – their mind/mental apparatus. In the first version, the following parts of the ARS-PA architecture are implemented:

- *Drives*: hunger, seek, play
- *Basic emotions*: pleasure, anger, fear
- *Complex emotions*:
  - Social emotions: shame, pride, reproach, admiration
  - Emotions derived from pleasure/unpleasure: hope, joy, disappointment
- *Memory types*: `Image Memory`, `Episodic Memory` (managing *scenarios* and *episodes*, and contributing to the formation of *desires*), `Procedural Memory`

Most drives are directly related to internal properties of the body (e.g. hunger is raised when the energy level drops below a certain threshold), others are raised when a (momentarily predefined) mixture of internal and external conditions applies. See [BLPV07] for details of the implementation. By using these drives, one would get a reactive, autonomous machine, potentially self-sufficient and somehow surviving.

The other building blocks of the first prototype, however, enhance the Bubbles features significantly by implementing functions from sensing up to recognizing episodes. This is by far not complete but it allows already for anticipating potential outcomes. The topmost function of the Bubble – episode handling – enables it to perform for instance the following:

1. A scenario is recognized.
2. The scenario is encoded and associated to a number of already experienced episodes, matching in varying degrees.
3. One of the episode leads to a highly desirable state, therefore it is activated and further pursued.
4. Other episodes leading to danger or other not so pleasurable states, evoke fear and are therefore avoided.

A context-specific episode handling is in particular necessary in conflicting situations where it is not so clear what is the best thing to do and what will happen next, and where the time frame for decisions is very tight.

To make the 'social life' of the Bubbles more interesting, the playground setting has been enhanced by some rules, for example:

- There are energy sources that can only be 'cracked' when two or more Bubbles work together.
- Bubbles can form gangs, trying to dominate the playground.

- Sometimes Bubbles desire company to play with one another or to promenade together.

The purpose of the rules is to lead to situations which require cooperative (and also competitive) behavior. There is a particular interest to configure Bubbles in such a way that they can solve a problem or task as a team. In this respect, the first step is to design Bubbles such that *social acceptance* and *task achievement* are *pleasurable* states in general and hence desired for. Social emotions (which are a subclass of complex emotions) contribute to the establishment of cooperative behavior. Figure 7.3 graphically depicts some of the social emotions implemented in the prototype, and also some of the associated variables. These variables are used to model social acceptance and its effects.

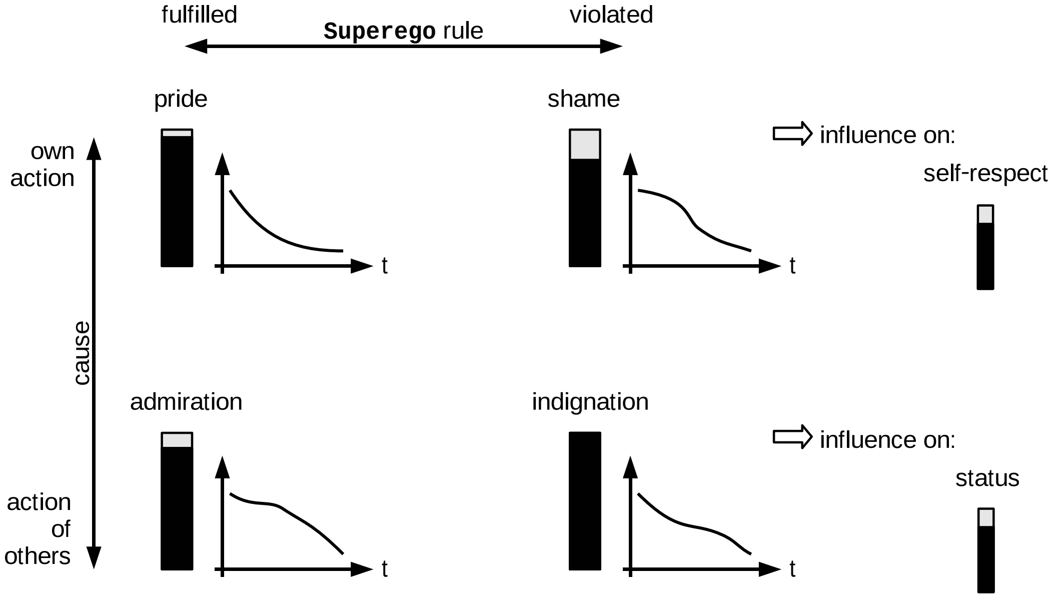


Figure 7.3: Some social emotions plus associated variables used in the BFG. The graphic illustrates that evoked emotions can be of different intensity and that they do not vanish instantaneously.

Moreover, for solving cooperative tasks, Bubbles have to remember relevant pieces of previous experiences, for instance:

- I am hungry but the kind of energy source next to me is too big to be 'hunted down' alone.
- The Bubble over there is one that I have helped last time.
- With high probability, it will help me now – so let's ask.

Remembering and recognizing such complex situations consisting of a number of temporally and spatially correlated events – episode handling – is, however, already a pretty sophisticated function, and based on a number of lower-level functions:

a) Condensing, in a hierarchic process, current sensor data into feature elements forming images

b) Decomposing the continuous stream of perceived images into sequences of events

c) Associating characteristic images with basic emotions (B.E.) and drives

d) Sequencing events

e) Recognizing scenarios and encoding (storing) episodes

f) Retrieving specific episodes out of the experiences stored in memory

g) Initializing desire sequences based on certain episodes that are reminded

Each of the processes is described in more detail below. For a graphical depiction of the processes, their relations, and the involved memory types and data structures, see Figure 7.4.

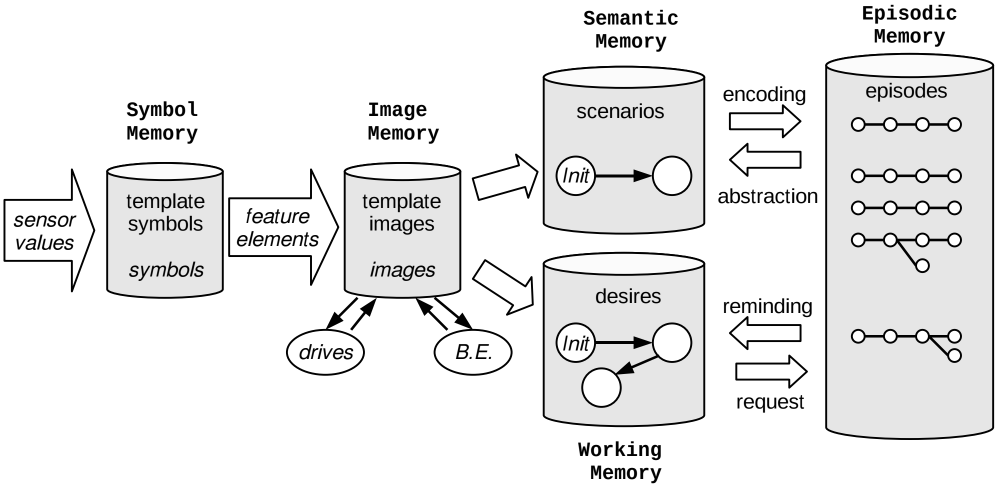


Figure 7.4: The diagram shows types of memories used in the ARS-PA prototype. It also depicts various data structures as they are used in the different types of memory, and the processes by which they are related to each other. Data values which are currently processed are in *italics*. These data values are not yet stored permanently.

a) **Image Match** – *Images* consist of sets of *feature elements* which consist of *symbols*. The memory responsible for image processing is called `Image Memory`. Initially, it only contains template images. They are stored permanently. Images generated out of the current sensory perception are stored temporarily in a buffer.

b) **Event Detection** – Currently perceived images (or collections of known symbols) are gradually matched (in a given order) against the template images to filter out 'important' (salient) images. This happens for example if one of the feature elements corresponds to a predefined element (a 'stimulus'), or is in an extreme range, or has suffered a significant change. The identified images get the status of *events* and receive further processing. With time, new images are added to the original set of template images, for example in case the images are connected with a high emotional arousal.

c) **Drive and/or Basic Emotion Evocation** – Template images can contain a *drive* and/or *basic emotion* element. The input stream of images is exactly first matched against these template images that are associated with a drive or basic emotion (see Figure 7.5). The eventual initialization of a drive or basic emotion happens because of the identification of characteristic stimuli also contained in the template images. With this mechanism, the system can react quickly to potentially dangerous or beneficial standard situations. Newly perceived images can also become tagged with a drive or basic emotion value, mainly because of similarity with an already tagged image, or because of a temporal coincidence (in this respect it is of relevance that emotions do not vanish instantaneously). Being tagged qualifies images as 'important', leads to their further processing, and promotes their permanent storage. With time, via emotional tagging the system learns to quickly decide for a growing number of situations whether they are potentially useful for the achievement of a basic need or not.

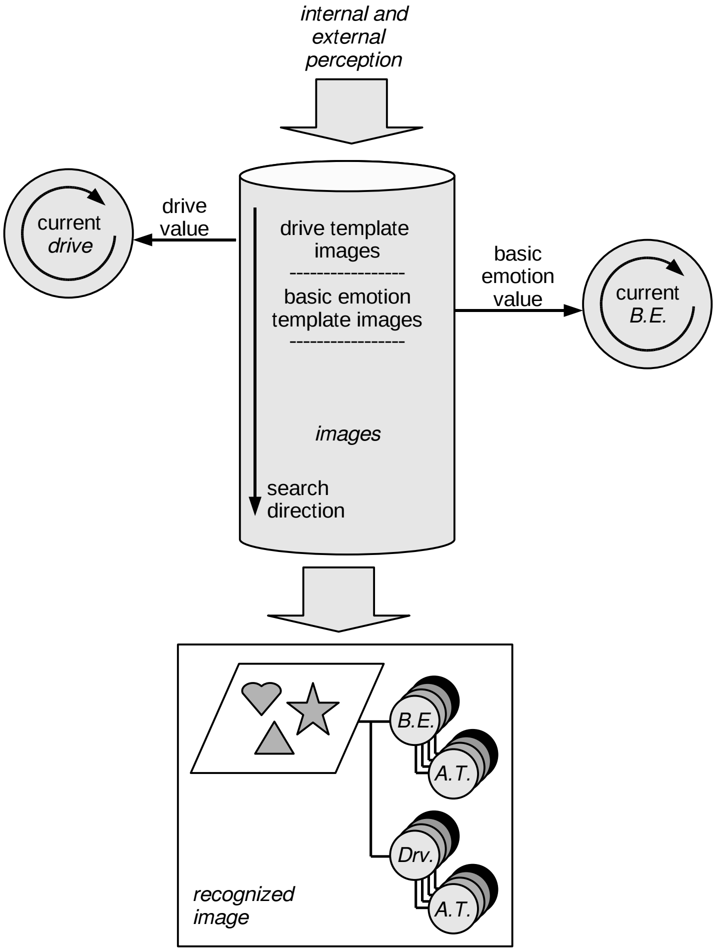


Figure 7.5: The procedure of recognizing an image. Perceived collections of symbols are first matched against drives and basic emotion template images. This serves to detect significant stimuli that indicate, significant situations that need to be dealt with quickly. In case there are matches, the current drive and basic emotion values are updated by the drive and basic emotion values associated with the matching template images. Matching drives and basic emotions can also induce associated action tendencies. The whole procedure results in an emotionally evaluated recognized image.

d) **Event Sequencing and Scenarios** – Events that form a temporal sequence and stand in some mutual context are represented – for the time being in the first prototype – as graphs. Thereby, the transitions to come from one state to the next are triggered by events (something has to happen to bring the Bubble from state A to state B). There is a set of predefined event sequences referred to as *scenarios.* Note that scenarios are templates. With these, the system designer gives the Bubble some kind of jump start by denoting important, recurrently occurring incidents and how to handle them. If certain behavioral patterns are known to be effective for survival from the very beginning, it is wise to equip the Bubbles with this knowledge from the start and to spare them the effort of learning. Scenarios represent cause-and-effect knowledge the system has.

e) **Scenario Recognition and Episode Encoding** As in the case of images, via a matching process, scenarios are recognized, and 'important' ones are identified. These are encoded in the `Episodic Memory`. Thereby, encoding is literally implemented as recording process. As a result of this construction, each stored (that is recorded) *episode* is a specific, refined instance of a given scenario. It includes the scenario-defining elements and additionally other elements that depend on the particular experience that has been recorded. Episodes take into account that no situation is ever experienced exactly twice. With the installation of a *generalization* process that detects, within the stored episodes, repeatedly occurring patterns, simplified, abstract patterns are added as new scenarios to the set of already stored scenarios. Thus, over time, the number of stored (that is, known) scenarios grows.

f) **Episode Retrieval** – Similar to the `Image Memory`, the `Episodic Memory` can be looked up in an associative manner. With the recognition of a scenario, all the stored episodes that are instances of this scenario are spontaneously retrieved ('reminded'). Thereby episodes that lead to emotionally tagged endpoints play an important role as described below.

g) **Desire Activation** – Analog to scenarios and episodes, desires are also currently implemented as graphs. Similarly to the case of scenarios, there are predefined desires that the Bubbles start with, for instance the desire to be socially accepted. A Bubble that recognizes the present ongoings as known desire sequence, or is reminded of an episode that has a strongly emotionally tagged endpoint, can 'activate' this desire with a certain tension. Note that there are also 'negative desires' where the Bubble wishes that something bad does not happen (or rather fears that it will happen). For the transitions of a desire graph, there are two possibilities: They can be triggered by events that the Bubble just passively perceives ('awaits') as they are happening, or actions performed by the Bubble itself (the Bubble takes action to actively proceed along the graph (see Figure 6.7). On the way to the fulfillment of a desire, the tension describing its intensity varies. The transitions in a desire are evaluated with a 'chance' of success which has influence on the pursuit of the desire, telling if satisfaction is near or if it is still a long way ahead. If the desire is finally satisfied, the tension is discharged. While a desire graph are processed, complex emotions emotions (like hope, confidence, or resignation) can be evoked.

All types of graphs have timeouts, which 'deactivate' the particular graphs if they remain in one state for too long. A timed-out desire graph typically leads to frustration.

The experienced episodes – which are way more detailed than the pre-given scenarios or desires – are stored in the `Episodic Memory`. Individual scenarios or desires might have a number of previously experienced episodes 'associated' with them which means that these episodes are more detailed examples of certain pre-defined templates [Gru07].

Due to their associative architecture, matching *Init* events or scenarios are found instantaneously. Known situations 'come to the Bubble's mind' (that is, into its `Working Memory`) and are the foundation for action planning. Typically, actions are taken, to satisfy drives or desires. The action planning procedures are based on an architecture described in [RLD+07]. In particular, it is analyzed how to represent *actions* in a graph so that they do not come into obvious physical conflicts (like simultaneous eating, drinking, and talking).

The ARS-PA prototype Bubble's `Episodic Memory` gradually 'forgets' entries that are not needed for a certain time. In turn, entries are more likely remembered, if they have had a big (emotional) impact on the Bubble (e.g. changed its energy level, were a threat, etc.). Each Bubble has its own, personal *Episodic Memory* that stores its experiences and previous adventures as episodes.

Apart from potentially triggering and activating a certain desire, any event can also increase or decrease the *tension* of an already active desire. It is not unusual that several (similar) scenarios and desires are simultaneously active, since they are initialized when the image recognition indicates a strong similarity of a stored *Init* template image with the currently perceived input.

The `Episodic Memory` is implemented as an associative memory, which means that an entry X can be searched by its meaning and not by its address. The search is flexible and fuzzy, so searching for a 'cup falling from the table' scenario will, for example, retrieve those memorized entries where it fell down and broke, where it did not break, and where it was caught by someone. It currently has (beside the graph time-outs, which are introduced for the sake of system stability) no explicit sense of time, but it can tell the timely order of events.

The entire simulation loop of a Bubble from perception to action has currently one concurrent block of functionality that consists of the following parts:

1. **Perception**: *Symbolized sensor data* is composed to *images.*

2. **Image Matching**: The sensed image is compared with stored ones (*template images*) and an ordered list of matches is formed. If the system has been waiting for a certain image (having set its focus of attention on something specific), once this image is detected, the its matching level is further increased. All *matching images* (e.g. those with recognition level above some threshold) can *eventually* issue a *reactive action tendency*, sorted by matching level, and importance. Matching images (if containing predefined characteristic stimuli) can directly lead to the activation of *drives* and *basic emotions*, even to contradicting ones[4].

3. **Scenario Processing**: If a special *Init* image is recognized, the corresponding *Init* state is evoked (i.e. the *scenario is activated*). If an edge of an active *scenario* graph is 'fulfilled', there is a change of state. Recognized scenarios can activate a desire. Sequences with high intensity are recorded as *episodes.*

4. **Desire Processing**: Similar to scenarios, but now 'open' edges can create *desire action tendencies.* They are sorted, depending on the current intensity of the associated *tension*, and the *likelihood of success.* Complex emotions are updated. Sequences with high intensity are stored.

5. **Merging of Action Tendencies**: The sorted lists of all existing *reactive* and *desire action tendencies* are *merged.*

6. **Superego Processing**: Activated *action tendencies* may be in eventually in conflict with exiting `Superego` rules. Some are simply not allowed, others gradually inhibited. New action tendencies can also be generated due to the `Superego` rules. At present being very rudimentary, later for instance it will be equipped with a mechanism to deduct social rules out of the reactions of other Bubbles to own actions.

[4] Think of the two stored images <<`food near, body weak`>, <`hunger, pleasure`>> and <<`enemy near, body weak`>, <`hunger, fear`>> and the three perceived symbols <`enemy near`>, <`food near`>, and <`body weak`>.

7. **Conflict Detection and Resolution**: Finalization of the ordering of action tendencies from all sources whereby different sources can be differently weighted.

8. **Execution**: At the moment, the BFG can only *execute* one *action* per simulation step and Bubble.[5]

Although the Bubble society may macroscopically appear as ‘agents’ on the first sight (which they actually are), the intrinsically conflicting agent system of interest is inside each individual agent (Bubble). So the real (or better the more important) multi-agent-system is inside the Bubbles. The various modules and even the contents of the modules are typically in conflict with each other which leads to a complex behavior.

[5] The action ‘move from A to B’ is executed via a ‘god mode’, where the Bubble just moves there without thinking about every single step in between. It wants to go somewhere, decides to do so and it just *happens magically.* The processes leading to decision formation and selection are the main focus of this work, not the robotics.

What is not started today is never finished tomorrow.
Johann Wolfgang von Goethe

8 Discussion and Outlook

After disscussing and assessing characteristic qualities of the proposed architecture, some potential applications are listed, followed by conclusions and proposals for further research.

8.1 Achievements of the Proposed Architecture

In [SCZ05], it is argued why cognitive architectures are essential for creating intelligent autonomous systems and why it is not enough to just mathematically specify the details of processes assumed to contribute to specific cognitive capabilities. The mind cannot be understood purely on its behavioral outputs. However, for modeling internal structures and processes that eventually result in this or that behavioral output, theoretical assumptions in advance are indispensable. These assumptions come from theories of cognitions. In so far as a cognitive architecture incorporates a theory of cognition, it is more than just a loose collection of mathematical equations and/or computational algorithms. The point is not to imply limitations to equations and/or algorithms, but to stress the critical dependence of the performance of the intelligent autonomous system as a *whole* on the way equations/algorithms *cohere* in the functional architecture.[1] In particular, details can only be modeled and worked out based on assumptions provided by the incorporated theory. These details include the specification of structures, their modular division, the relation between modules, forms of knowledge representation, types of memories, types of perception, learning, reasoning, etc. The underlying theoretical assumptions are anything else than irrelevant. They represent a commitment concerning the structure and the dynamics of the resulting cognitive capacities, and they lead to implications based on which comparisons between different architectures can be made and the quality of them judged.

The presented cognitive architecture is based on the neuro-psychoanalytic view of the human mind. In the following, cognitive and behavioral phenomena that can be implemented in a technical system based on this approach will be discussed.

8.1.1 Levels of Cognitive Modeling

The idea that cognitive modeling – for computational purposes and also in general – should result in the development of a hierarchic, multi-level model is quite common. It is also assumed within

[1] The situation is similar to what is depicted in Figure 2.6.

the ARS project, however, with an important restriction: As has already been stated in the first paragraph of Chapter 6, the higher the level, the more the proposed system turns into one, where more or less ‘equally powerful psychic forces‘ battle with each other on the same level (see Sections 6.6.5 and 4.4). Leaving the psychic level aside, the reason why cognitive architectures should result in a multi-level organization is related to a simple fact: the physical processes in which singular neurons engage do not possess mental qualities. Instead, mental experience is a functional phenomenon appearing only on top of a complex structural organization as a result of several, partly parallel and partly sequential, processes (see Chapter 5). However, there are several possibilities how to determine the different ‘levels’ of analysis and modeling that have to be provided for. An overview of the issue, can be found in [SCZ05]. There, for example, the distinction proposed by Newell and Simon is listed: 1) physical level, 2) symbol level, and 3) knowledge level [NS76], or that from Marr: 1) computations, 2) algorithms, and 3) implementations [Mar82]. Within the ARS project, as elaborated in [Bur07, pp. 14–17], the question of levels is not addressed from a theoretical computational perspective, but from a neuro-psychological standpoint based on the work of A. Luria [Lur73]. In [SCZ05], also a non-computational angle is taken when requiring a comprehensive cognitive architecture to provide for the following levels:

- Physiological level
- Componential level (intra-agent level)
- Psychological level
- Social level (inter-agent level)

All of these levels are addressed by the ARS-PA architecture.

The *physiological* level is captured by adhering to the principle of embodiment. Of course, a technical system has no physiological body like an organism, but it still has a) a material body, b) the functioning of which depends crucially on potentially limited resources, and c) sensors and actuators which are in direct physical interaction with the environment. Even if, in the beginning, both, the autonomous system equipped with the ARS-PA architecture as well as the environment where it is embedded, are just simulated, the principle of embodiment is still respected by the ARS-PA architecture because the information processes taking place in the architecture (and making up its understanding of what is going on) are really rooted in the interactions of the system with its environment. This includes a hierarchic condensation process of sensor data into symbols, thereby grounding symbols in a natural way (see Section 6.4), the incorporation of drives and (basic) emotions to deal with limited resources, thereby functionally modeling homeostasis (see Sections 6.5.1, 6.5.2, 6.6.2), actions and procedures that are both symbolically represented as well as in a sensor-near form where they can directly control the actuators of the system see Section 6.7).

The *componential level* is given by the modular architecture of the ARS-PA model. There are interfaces defined between the individual components such that they can synchronize, cooperate, or inhibit each another. The various modules have different particular tasks, their real functional outcome, however, is only visible when they work on a common task (see Section 7.3; also [Gru07, pp. 86–98]).

While the previous level is mostly given by the structure of the system, the *psychological level* is a result of the processes running on this structure. The mechanisms of basic emotions and complex emotions are enhancing the system with psychological qualities. In contrast to traditional (maybe deterministic rule-based) decision making systems, such systems can evaluate situations in a very complex, for instance asymmetric, way. Danger, even if very unlikely, can be overestimated, while potential chances can be perceived not that intense. The effect can even adjust itself adaptively during run-time. The result of complex internal interactions with many factors of influence weighted in an asymmetric and unusual way can be behavior that superficially looks completely chaotic and illogical. System-specific, individual behavior can also arise out of the `Episodic Memory` that gives the autonomous systems a 'personality' of their own.

Complex emotions and the `Superego` enable the individual to contribute to a social group, which leads to a new, *inter-agent level* of complexity. The current testing setting of the Bubble Family Game will be replaced – once when designing a real application – by a technical setting like a factory or public space where the autonomous system interacts with existing processes, other autonomous systems, and humans.

8.1.2 Comprehensiveness and Coherence

Cognitive architectures have to model such diverse capacities as perception and situation evaluation, recognition and categorization, decision making and choice, remembering and learning, prediction making and monitoring, problem solving and planning, reasoning and belief maintenance, action execution, social interaction, and communication. Most of the existing architectures address just a few of the above capacities, and often they are explicitly tailored to just one of them. Roughly, architectures can be divided into low-level ones using reactive control, and high-level ones using deliberative problem solving. In the last year, there have been suggestions of architectures combining both approaches (e.g. CLARION [SST05]; H-CogAff [SCS04]). However, most of them either do not get general enough, or they stay too vague, or they are too complicated for implementation. CLARION for example combines implicit and explicit representations and mechanisms, but it does not include emotional mechanisms, and thus it is not as comprehensive as the proposed architecture. H-CogAff suffers from the last two mentioned shortcomings: It has a very complex and rich structure, however, almost none of the included mechanisms are really specified out. This is related to the fact that it is not built on an underlying psychological theory.

The presented approach, based on neuro-psychoanalysis [nps07], sketches how low-level and high-level cognitive phenomena can be unified along a number of fronts. Thereby, emotions acting on each level of the architecture play a central role. What is most novel of the approach, is the introduction of psychoanalytic concepts and insights. They provide a coherent frame which guides the arrangement of all the diverse processes that make up human intelligence and personality. They also allow to exploit subjective, introspective knowledge of the human mind for a technical design.

8.1.3 Combined Bottom-Up/Top-Down Approach

The inspiration and scientific foundation of the suggested cognitive architecture comes from neuro-psychoanalyis, a relatively new interdisciplinary effort to bridge the gap between neurological findings and psychoanalytic concepts describing the human psyche [nps07]. Leading researchers of both camps are contributing to this dialog, mutually trying to combine and reconcile their perspectives.

Section 4.1 of this work describes approaches to computationally model cognitive capabilities with the aid of emotions based on neurology, neurobiology, and/or ethology. These approaches are presented under the label 'low-level approaches'. In contrast, Section 4.2 shows approaches which are purely based on psychological theories of emotions. These theories leave physiological and neurological aspects completely out.

The proposed ARS-PA model, based on neuro-psychoanalysis, tackles the brain/mind problem from two sides, since neither neurophysiology, nor approaches dealing only with the human mind/psyche alone can lead to systems with human-like affective and intellectual capabilities. Neuro-psychoanalysis benefits from the fact that high-level findings can give hints to understand low-level mechanisms and vice versa.[2] This analytical advantage becomes a synthetic advantage, when high-level structures guide the design of the lower levels and vice versa. Moreover, as the neuro-psychoanalytic perspective is also an inherently evolutionary perspective it allows for instance to have a look at the cognitive capacities of more ancient and thus much simpler organisms which can particularly contribute to the design of the lower levels.

8.1.4 Combined Distributed/Symbolic Representation

Bottom-up approaches often use distributed representations (like e.g. the models described in Section 4.1.2), whereas top-down approaches mostly use symbolic representations (like e.g. the models described in Section 4.2). Using distributed representations for the sensor-side and symbolic representations for the cognitive side is certainly a good solution from a practical point of view, since the two methods fit well to the required tasks on the respective levels. Whether such a representational division is more than pure convenience but corresponds in some functional sense to the working of the human mind, can only be speculated today, but would be in principle compatible with the neuro-psychoanalytic view as it would stress the difference between the physical level of neuronal processing and the psychic level of subjective experiences. At this point, it shall be once more stressed that throughout the whole work the term 'symbol' is not understood in a classical, mathematical sense as completely arbitrary marker without internal structure (compare with Section 2.4.2). Rather, it is assumed that distributed, implicit representations 'converge' to symbolic constructs that can enter computations. Even if symbolic representations may 'just' arise out of a complex organization of non-symbolic components, when once established, symbolic representations are considered to be different (and actually more powerful) than distributed representations (important symbolic properties being f.i., according to [Roc01, pp. 14-15], decoupling from ongoing dynamics, ruleful composition, systematic semantic interpretation).

The power of open-ended derivations like in a formal system is often considered not to be achievable with purely dynamic, non-local, and non-symbolic types of memory (see the arguments given in e.g. [Roc95, Min91, Cla05]). Explicit symbols – acting as tokens for categories like 'food', 'fruit', 'tool', but also non-static categories like 'a running person' or 'spinning wheels', etc. – allow the inference of 'a mango is eatable' based on the knowledge that 'a mango is a fruit', any moment in time, even if there is currently no sensory perception of a mango or whatsoever and even without ever having seen or eaten a mango before. A distributed, implicit representation of food or fruit cannot be related in such an unconstrained way – that is, so independent from actual physical correlations – to other representations as this is possible with the help of a symbolic language (compare with Section 2.4). Therefore, it is sensible to condense on the way to the higher levels,

[2] Note however that low-level mechanisms can never explain or completely determine high-level mental capabilities.

as it is done in the architecture (see Sections 6.4 and 7.1), the initially distributed sensor readouts to an explicit symbolic representation. A central element in this condensation process are images.

8.1.5 Images and Scenarios

Images are an important data structure of the proposed architecture. They are acting as an *intermediate structure between plain numbers* coming from sensor readouts and *symbolic representations* used by the cognitive modules. Information stored in numbers is always opaque [Min91]. The creation of images is an early step in the hierarchical process of attaching meaning to sensory information (compare with the discussion of icons as simplest form of referential relationships in Section 2.4.2). Based on similarities and temporal correlations, sensor values are grouped into sets and remembered as images. Symbols are created in a hierarchical process, starting with sensor-near symbols that represent for instance basic geometric shapes or sounds. Later, these symbols are combined to other, more meaningful symbols, representing for instance a dog or a house (compare with the approach presented on [num05]). Images do not only consist of visual, auditory or tactile information. They also get a physiological and emotional component attached, as well as an action tendency (see the definition of feature elements in Section 7.2.1).

The next step is to combine images into temporal sequences of events (and/or actions). Note that this is not a contradiction to the fact that images (and also symbols) can already represent motions. So this is now a combination of several static and/or moving representational chunks on a higher scale, leading to *scenarios* (and *routines*) as core elements of the proposed architecture. Both of them have the character of templates, and the system is equipped with a set of them from start on. During run-time, ongoing interactions lead to the storage of ever finer-grained episodes, based on the set of originally available scenarios, but unique to the individual system. Via generalization and categorization, the finer-grained episodes can again be unloaded from irrelevant details and turned into new templates for future experiences (compare with 7.4). The special role of repeating sequences for the development (formation) of the human memory systems, and for the interpretation of what is currently going on is stressed both in modern psychology [Nel96, MW06] and psychoanalysis [Den99, Dor93].

8.1.6 Integration of Affective and Cognitive Components

Throughout the architecture affective components (drives, emotions, desires, etc.) are utilized to perform evaluations. It is described how they are linked to other cognitive processes (such as acting-as-if) and how they can modulate behavior. Drives are introduced to deal with the problem of keeping track of internal limited resources. Still, drives, and even more desires which partly arise out of drives and partly out of perceived or remembered events, are neither itself an entirely intern phenomenon, nor the passive product of external forces (see Section 7.2.4). Basic emotions are used to evaluate significant standard situations, often such that require quick reactions. Complex emotions are used to promote social behavior and to judge the progress of desires and goals. Drives and desires are not only used to passively promote already raised needs and goals, but also to actively explore objects and to support the generation of expectations and plans (see Section 7.2.5).

The raising and discharge of tensions (connected with pleasure and unpleasure) is suggested to model the dynamics of drives and emotions. To study the parametrization of this dynamics in more depth in order to best capture object/environment contingencies is an important topic for further research (see also [Bul06]).

8.1.7 Episodic Memory

The architecture directly supports the formation of an emotionally rated episodic memory which is central for many other processes running within the architecture and operating on its contents. By interacting with the environment, existent knowledge is used to recognize and interpret what is going on in the environment, and newly made experiences, in turn, contribute to the formation and storage of new knowledge. The implemented processes can be for example compared with the role of the episodic memory as described in [MW06]. There it is stressed that humans first develop their procedural memory, and later – in combination with the learning of a language – their semantic memory, and even later their individual episodic memory. Thereby, the contents of the procedural memory are behavioral schemes which work as templates for the other types of memory. Early templates consist of concrete sensory impressions. Later on, due to generalization and categorisation, templates become more abstract. With time, more and more experiences of interactions with the environment are stored. Already stored templates rest in memory until they are activated by a current stimulus. With the learning of a symbolic language templates can also be activated without an explicit external stimulus (i.e. imagined or fantasized).

8.2 Potential Future Applications

The future outcome of this fundamental research work is expected to be a technical system that can evaluate complex situations, focus on the 'important' parts of an information overload, and come to appropriate decisions by remembering and interpreting previous experiences. Such a system can be used within a number of technical processes, some examples of which are considered below.

Note also that in Section 6.5.2, it has been briefly described how the basic emotions of mammals can be given a technical interpretation and functionality.

Industrial and Process Control

Factories and production plants typically try to optimize their costs, their throughput, their raw material efficiency, and their emissions. Such systems consist of a number of dependent control loops, logistics, alarming systems, and resource management systems. Keywords are *enterprise resource planning* (ERP) and *manufacturing execution system* (MES). Such systems are structured in a traditional way, using one or several databases, controllers, and control networks to run the processes. It is the task of the designer to keep system complexity low. As a solution, often entities that should work as a concerted ensemble are artificially separated to keep things simple. A system that is able to cope with high grades of complexity might might substantially improve current solutions.

Intelligent Buildings

Modern buildings are equipped with building automation systems where tens of thousands of sensors are used to run the air conditioning system more efficiently, to safe money with heating, or to optimize the internal logistics. Previously independent processes (window shades, heating, ventilation, presence sensors, lighting, etc.) are now available in one network. Exploiting all potential synergies is, however, a task too complex for traditional systems. A system with more ‘self-awareness’ about its current, resource-related state can be expected to find more optimized operation modes.

Self-sufficient Computer Systems

In [NOR03], it is described how computer systems can use aspects of affects and cognitions to improve their availability and reliability. A given example is that of a failure of a disk drive in a RAID (redundant array of inexpensive drives) system which makes the system ‘anxious’ and alert to further potential failures. One broken drive does not stop the entire system (because of available redundancy), but a further defect might indeed cause a total breakdown. The evoked anxiety can lead to the lowering of safety margins. Additionally, the system might autonomously decide to start backups or mirroring services which it would not have started otherwise. One of the question that arises with such an application is whether the cognitive ‘alarm system’ should be run in parallel to the ‘standard system’, or whether these two instances should be incorporated into one system. The latter case would be the more ‘natural’ for the ARS system, as it is built on the premise of permanently keeping track on its resources in order to ‘guarantee its bodily well-being’.

(Semi-)autonomous Navigation

Today’s airplanes and trains already demonstrate how (semi-)autonomous navigation can improve the safety, efficiency, and comfort of transport systems. The next logical step will be automatic control of truck columns and individual cars. These two applications are way more complex, since more parties are involved and more degrees of freedom have to be dealt with. To make a decision, huge amounts of information from various sources have to be weighted against each other in a very short time span. Think of a crowded city highway with hundreds of cars changing lanes, joining, or leaving the highway. Each of them individually must pursue its route and plan its next actions, but at the same time also ‘cooperate’ with others and project ahead their next actions.

Safety-critical Systems

Functional safety (short safety) tries to prevent humans from being injured or killed [61502]. Typically, such safety-related systems are very simple in their structure. Usually, large and complex amounts of data are not processed, however, there are several applications where exactly this is required. A good example is public safety in the light of terrorism. Checking if someone drops a suitcase in a train or a bus still needs human operators sitting in front of surveillance screens. An automated system would save a lot of human resources, and could also supervise more than one place in parallel. The demands on such a system are, however, high. It must be able to interpret visual and other sensory information, merge information from various sources, evaluate suspicious behavior, thereby eventually taking the current political situation into account, etc. Another example where it can become rather difficult to decide situations are conflicting safety rules (e.g. automatic sealing of rooms in the case of a bio-hazard while people are within the rooms, automatic argon-based fire extinguishing in a data center while personnel is present). As it is not possible to define all the situations that may occur in advance, autonomous systems that can cope with dynamic environments are required.

Autonomous Robots

Autonomous robots can be useful for several purposes, among them activities in environments which are too hostile for humans. As goal-oriented agents, they might be sent to the sea floor, into outer space, or into some military mission. In cases where remote control is not easily possible, the robot should make its own decisions, based on its own knowledge, local information, implemented evaluation mechanisms, predefined moral concepts, etc. A broadband communication link to human operators might be not always possible because of huge distances, or because of distorted and unreliable communication channels. Thus, the ability to cope with complex situations must be 'at site', that is, on board of the robot.

Generally, it might become necessary for certain applications to encapsulate the authorization of an autonomously deciding system like a robot into a simple, pre-given rule-base. For instance, if the system is protecting something valuable on behalf of an insurance company, the company will certainly ask for some guaranteed performance. However, the system might be too complex to guarantee specific features, in particular of the running system. Statistics might help. Another possibility is the definition of minimum and maximum values that may not be exceeded in any case. For such a solution, the autonomous system can optimize within certain limits, while minimum and maximum values are hard-wired.

8.3 Conclusions and Outlook

In this work a new comprehensive cognitive architecture for autonomous systems has been proposed, addressing not only both low-level and high-level cognitive capacities, but also describing a way how to combine them into one *unified* model. The architecture is designed based on a functional model of the human brain/mind (more specific Â´mental apparatus', see Section 5.2.1) as given by neuro-psychoanalysis (see Section 5.3). The latter is a relatively recent scientific effort of bringing together neurological findings of the organization of the human brain with psychoanalytic concepts and models hypothesizing the working of the human mind (psyche) on a functional level [nps07].

One of the reasons for choosing neuro-psychoanalysis as basis is that it strongly informs how to pursue a combined bottom-up/top-down approach. This is important because in the case of building an intelligent machine (as in nature), the problem is how to link the sensor data level with a level of 'semantic understanding'. The reason why neuroscientific knowledge is in particular combined with *psychoanalytic* concepts (and not, for instance, with those of other psychological branches) is twofold. First, psychoanalysis offers *'the most coherent picture'* of the functioning of the human mind/psyche [Kan99, p. 505], and second, psychoanalysis really takes seriously the introspective observations of individuals concerning their mental experiences (mostly by provoking verbal reports of subjective states). It is this *aspect of subjectivity* that makes mental states unique among all other states of nature, and, one of the core assumptions of this work is to consider this unique feature of human intelligence as important ingredient that should not be neglected for a technical design even if, at first sight, it might seem to be in contradiction with an objective approach. Of course, when granting (to some extent) subjective knowledge entrance into technical modeling, one has to take care not to leave scientific ground. In this respect, the combination of psychoanalytic concepts with sound scientific facts produced by contemporary neuroscientific research proves itself again as being a balanced choice.

The concepts and design principles derived from the neuro-psychoanalytic view are of course *not necessarily* different from those derived by other neurological, psychological, or cognitive theories. This aspect of scientific compliance is welcome and actually sought for in this work, some parts being explicitly dedicated to establish it (in particular Sections 2.1 and 3.2) . On the other hand, their is the claim throughout the work that the proposed approach presents a new, promising way towards creating an artificial intelligent autonomous system, a goal that, so far, has not been reached by the existing 'four' approaches, neither by classical symbolic AI (see Section 2.1.1), nor by connectionist approaches (see Section 2.1.2), nor by the embodied agents approach (see Section 2.1.3), nor by existing computational models using emotions (see Chapter 4). Apart from the arguments already brought in the previous paragraph when explaining the usefulness and the uniqueness of the choice of neuro-psychoanalysis as a basis, the new approach – potentially leading to a 'fifth' generation of artificial autonomous systems – informs the design of the proposed cognitive architecture in some important aspects, and this in a threefold way concerning its constitutive elements, its organization, and its dynamics.

A key feature of the proposed architecture is the introduction of affective elements. It is also described how the affective elements are linked, within the architecture, with more traditional elements used by artificial intelligence, like for instance planning algorithms. The affective elements introduced in the architecture are drives, (basic and complex) emotions, and desires. The functional value of drives is given by the linkage of drives with bodily resources and bodily well-being and, by the fact that drives bring about *activity.* Particularly the value of a seeking drive leading to active search behavior, and its extension leading to a playful exploration of cause/effect contingencies of the surrounding environment without acute necessity have been stressed. The introduction of *tensions* and the description of how their plummeting gives rise to pleasure according to psychoanalysis is a further important aspect of the proposed cognitive architecture, essentially contributing to the dynamic of the ongoing processes (see Section 7.2.4).

Pleasure and unpleasure make up the most fundamental emotional scale [ST02, p. 108], rating outcomes of actions depending on their success or failure concerning the fulfillment of a need. All other, more specific emotions within the architecture also act as evaluations. Basic emotions quickly but rather automatically deal with standard situations (which often can potentially destroy the system). Complex emotions deal with more complicated relationships between the system and its surroundings, including the relationships of the system to other systems it has to cooperate with in a team, or those to human beings it has to interact with. Complex emotions arise not only because of the perception of a characteristic stimulus. They largely depend on emotionally rated images and scenarios contained in the episodic memory of the architecture. This emotionally afflicted, individual memory enhances the capability of the system to *anticipate future situations on the basis of past experiences* and represents another important feature of the proposed architecture especially contributing to its context-sensitivity. Desires (and also drives) are motivations to actions, but – similar as in the case of complex emotions – they arise out of the 'popping-up' of experiences stored in the episodic memory. This happens in the following way: Previous experiences related to the current situation which are positively emotionally rated are 'desired' to be repeated, negatively rated ones are 'desired' to be avoided.

Images and scenarios are introduced as fundamental data structure for the various types of memories. The system starts with a set of predefined images and scenarios, and also with (emotionally supported) mechanisms to create and store new images and scenarios, thereby extending the original sets. This is a kind of learning. About learning, it has been stressed to distinguish predefined knowledge coming from the programmer of the system, knowledge acquired in an adaptive training phase, and knowledge acquired during run-time. In a biological setting this corresponds to knowledge provided by genes, learning during childhood, and learning as an adult. From a psychoanalytic point of view, learning during early infancy is strongly shaping the future emotional setting as well as the cognitive processes of an individual. In this respect, the relationship of the infant with its caregiver is determining. In [Bul05], the author shows how such an infant/caregiver relationship can lead, within a psycho-dynamic approach based on the plummeting of tensions, to the emergence of a mutually understandable proto-language. The mechanism can also be implemented in the proposed cognitive architecture, all the necessary elements being available. It is my claim that the affective elements of the architecture can also easily support the accommodation of many other kinds of learning mechanisms, a topic that has been partly addressed in the work (see the end of Section 6.5.2 about operant conditioning and 6.6.2 about imitation learning). In general, finding the right mixture between predefined knowledge, system adaptation guided by a supervisor during a training phase, and run-time learning is certainly a topic that has to be explored in more depth in the future, including how it can be fruitfully supported by contemporary psychoanalytic insights.

A simple prototypical implementation of the architecture has been described. In future, this first prototypical implementation will be iteratively improved, using better realizations of the used data structures and algorithms. The design of the architecture principally allows the incorporation of already existing datastructures and algorithms. Of course the datastructures and algorithms have to be fitted to the given modular organization, inspired by the id-ego-superego model of Freud, and also to all the other incorporated psychoanalytic principles, f.i. the dynamics of drives and desires. However, this should be a solvable problem for several highly elaborated, special purpose algorithms (like retrieval algorithms to search for similar memory entries, categorization algorithms to group repeatedly occurring series of events and episodes according to their invariant features, clustering algorithms to create symbols out of sensor data, image recognition algorithms etc.). The given psychoanalytic model and the incorporated principles mainly provide a general framework indicating which elements are necessary and how they are interwoven with each other.

Together with the current implementation, a simulation environment for testing and evaluating the proposed cognitive architecture has been designed (the ‘Bubble Family Game’). In parallel to the improvement of the used algorithms, this testing environment has to be enhanced (or new testing environments have to be developed) such that real psychological problems and questions can be better reflected with the simulation. This is important for the evaluation of the approach, allowing a sound analysis of the question which kinds of psychological phenomena it can account for. Additionally, using a simulation environment that can mirror real psychological experiments will also be a first step towards the application of the technical system as a tool in psychological research. In this respect, the higher-level modules of the architecture have to be elaborated in more detail, potential topics of research being the change of the focus of attention, defense mechanisms, a deeper understanding of the dynamics of emotions (their elicitation, endurance, decay, and mutual interaction), the role of the superego as an ideal-providing entity, the introduction of censor mechanisms to suppress certain desires, etc.). It certainly cannot be hoped for to be just one step away from building a technical system that can achieve a human-level psychological performance. This is anyway not the primary goal of the proposed work. Still, any progress in this direction can certainly improve human-computer interaction and thus contribute to construct technical systems that can better understand and predict human behavior, both aspects are certainly of relevance for some of the envisaged applications, like surveillance systems for airports or sports stadiums, or service robots aimed to support the life of elderly or handicapped persons at their homes.

The character of the present work is an introductory one, drawing a unified picture of the new approach, considering its bottom-up as well as top-down aspects. Maybe the focus has been put a little bit more on the bottom-up direction (how to get from sensor values to meaningful symbolic representations and finally appropriate actions), after all the starting point of this work is the construction of autonomous systems that can make sense of ever increasing amounts of sensor data. In the ARS project, the attempt is to build machines that are supposed 'to really understand' what is happening, where they are, if the situation is bad or good, what were the different possibilities they could do now, and which would be the hypothesized consequences in each case. It has been argued that a key for this is the described combination of affective concepts, such as emotions and desires, with syntax-oriented symbol manipulating inference capacities ('acting-as-if'). The crucial point is to equip machines with *evaluative* mechanisms such that they can autonomously and adaptively acquire information and turn it into *meaningful* pieces of knowledge. The so derived knowledge is grounded in the interactions of the system with the ongoings in the world. The resulting representations, organized in affectively charged images, scenarios, routines, and further to be specified, more complex datastructures, such as acts, relationship matrices, ideals, etc. [Den99], reflect a context-sensitive picture the system has constructed about its environment but also about itself: its past and future action possibilities and how the environment most probably will answer when the system selects one action for execution compared to another.

References

[61502] IEC 61508. Functional safety for electrical/electronic/programmable electronic safety-related systems, 2002.

[Alb96] J. S. Albus. The Engineering of Mind. In *Proceedings of 4th Int. Conf. on Simulation of Adaptive Behavior: From Animals to Animats*, 1996.

[Bad97] A. Baddeley. *Human Memory: Theory and Practice.* Psychology Press, 1997.

[Bat79] G. Bateson. *Mind and Nature: A Necessary Unity.* New York: Bentam Books, 1979.

[BLPV07] W. Burgstaller, R. Lang, P. Pörscht, and R. Velik. Technical Model for Basic and Complex Emotions. In *INDIN07*, 2007.

[Bra84] V. Braitenberg. *Vehicles: Experiments in Synthetic Psychology.* MIT Press, Cambridge, MA, 1984.

[Bro86] R. A. Brooks. A Robust Layered Control System for a Mobile Robot. *IEEE Journal of Robotics and Automation*, RA-2 (1):14–23, 1986.

[Bro91a] R. A. Brooks. Intelligence without reason. In *Proceedings of 12th Int. Joint Conference on Artificial Intelligence*, 569–595, 1991.

[Bro91b] R. A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991.

[Bru07] D. Bruckner. *Probabilistic Models in Building Automation: Recognizing Secnarios with Statistical Methods.* PhD thesis, Institute of Computer Technology, Vienna University of Technology, 2007.

[Bul02] A. Buller. Volitron: On a Psychodynamic Robot and Its Four Realities. In *Proc. of the Second Int. Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, ed. C. G. Prince et al., 17–20. Edinburgh, Scotland, 2002.

[Bul05] A. Buller. Building Brains for Robots: A Psychodynamic Approach. In *Proc. of Pattern Recognition and Machine Intelligence PReMI 2005*, ed. S. K. Pal et al., 70–79. Springer-Verlag Berlin Heidelberg, 2005.

[Bul06] A. Buller. Machine Psychodynamics: Toward Emergent Thought. Technical Report TR-NIS-0005, Advanced Telecommunications Research Institute International (ATR) Network Informatics Laboratories, Kyoto, Japan, 2006.

[Bur07] W. Burgstaller. *Interpretation of Situations in Buildings.* PhD thesis, Institute of Computer Technology, Vienna University of Technology, 2007.

[Car95] P. Cariani. Towards an evolutionary semiotics: the role of symbols in organisms and adaptive devices. In *Proc. of the International Seminar on Evolutionary Systems (ISES) Vienna*, eds. S. Salthe and G. Van de Vijver, 1995.

[Cha02] D. Chandler. *Semiotics: The Basics.* London: Routledge, 2002.

[Cla97] A. Clark. *Being There. Putting Brain, Body, and World Together Again.* MIT Press, 1997.

[Cla04] D. S. Clarke. *Sign Levels: Language and Its Evolutionary Antecedents.* Springer, 2004.

[Cla05] A. Clark. Beyond the Flesh: Some Lessons from a Mole Cricket. *Artificial Life*, 11, (1-2):233–244, 2005.

[Cn97] L. Cañamero. Modeling motivations and emotions as a basis for intelligent behavior. In *Proc. of the First International Conference on Autonomous Agents*, ed. W. L. Johnson, 148–155. New York: The ACM Press, 1997.

[Dam99] A. Damasio. *The Feeling of what Happens: Body and Emotion in the Making of Consciousness.* Harcourt Brace & Company, New York, 1999.

[Dam03] A. Damasio. *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain.* Harvest Books, 2003.

[DB06] G. Russ D. Bruckner, B. Sallans. Probabilistic Construction of Semantic Symbols in Building Automation. In *Proceedings of 2006 IEEE INDIN'06*, 2006.

[Dea98] T. W. Deacon. *The Symbolic Species: The Co-Evolution of Language and the Brain.* W. W. Norton & Company, 1998.

[Den99] F-W. Deneke. *Psychische Struktur und Gehirn.* Schattauer GmbH, Stuttgart, Germany, 1999.

[Des75] R. Descartes. In *The Philosophical Works of Descartes*, eds. E Haldane and G. Ross, volume 1 and 2. Cambridge Univ. Press, Cambridge, U.K., 1975.

[DFKU07] D. Dietrich, G. Fodor, W. Kastner, and M. Ulieru. Considering a technical realization of a neuro-psychoanalytical model of the mind. In *Proceedings of ENF 2007 – 1st Int. Engineering & Neuro-Psychoanalysis Forum*, 13–20, July 23, Vienna, 2007.

[DKM$^+$04] D. Dietrich, W. Kastner, T. Maly, C. Roesener, G. Russ, and H. Schweinzer. Situation Modeling. In *Proceedings of the 5th IEEE International Workshop on Factory Communication Systems WFCS04*, 93–102, 2004.

[DL$^+$06] T. Deutsch, R. Lang, , G. Pratl, E. Brainin, and S. Teicher. Applying Psychoanalytic and Neuroscientific Models to Automation. In *Proc. of the 2nd IET Int. Conf. on Intelligent Envrionments*, volume 1, 111–118, 2006.

[Dor93] M. Dornes. *Der kompetente Säugling: Die präverbale Entwicklung des Menschen.* Fischer, Frankfurt am Main, 1993.

[DS00] D. Dietrich and T. Sauter. Evolution potentials for fieldbus systems. In *Proc. of the WFCS 2000, IEEE Int. Workshop on Factory Communication Systems, Instituto Superior de Engenharia do Porto, Portugal*, 343–350, 2000.

[Ede87] G. M. Edelman. *Neural Darwinism: The Theory of Neuronal Group Selection.* Basic Books, Inc., New York, 1987.

[Ede89] G. M. Edelman. *The Remembered Present: A Biological Theory of Consciousness.* Basic Books, New York, 1989.

[Ede03] G. M. Edelman. Naturalizing consciousness: A theoretical framework. *PNAS*, 100:5520–5524, 2003.

[EF94] W. Ebeling and R. Feistel. *Chaos und Kosmos: Prinzipien der Evolution.* Spektrum Heidelberg, Berlin, Oxford, 1994.

[Ell92] C. Elliott. *The Affective Reasoner: A process model of emotions in a multi-agent system.* PhD thesis, Northwestern University, Evanston, Illinois, 1992.

[Fou90] ed. E. Foulkes. *Selected Papers of S.H. Foulkes: Psychoanalysis and Group Analysis.* Karnac, New York, 1990.

[FP88] J. Fodor and Z. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28:3–71, 1988.

[Fre91] S. Freud. *On Aphasia.* London: Imago, 1891.

[Fre00] S. Freud. Die Traumdeutung. In *Gesammelte Werke*, volume II und III. 1900.

[Fre15a] S. Freud. Das Unbewußte. In *Gesammelte Werke*, volume X, 263–304. 1915.

[Fre15b] S. Freud. Instincts and their vicissitudes. In *S.E.*, volume 14: 111. 1915.

[Fre15c] S. Freud. Triebe und Triebschicksale. In *Gesammelte Werke*, volume X, 209–232. 1915.

[Fre15d] S. Freud. The unconcsious. In *S.E.*, volume 14: 161. 1915.

[Fre16] S. Freud. Introductory lectures on psycho-analysis. In *S.E.*, volume 16: 339. 1916.

[Fre20a] S. Freud. Beyond the pleasure principle. In *S.E.*, volume 18: 7. 1920.

[Fre20b] S. Freud. *A general introduction to psycho-analysis.* Horace Liveright, 1920.

[Fre23] S. Freud. The ego and the id. In *S.E.*, volume 19: 3. 1923.

[Fre26] S. Freud. Hemmung, Symptom und Angst. In *Gesammelte Werke*, volume XIV, 111–206. 1926.

[Fre33] S. Freud. New introductory lectures on psycho-analysis. In *S.E.*, volume 22: 3. 1933.

[Fre36] A. Freud. The ego and the mechanisms of defence. In *The Writings of Anna Freud*, volume 2. 1936.

[Fre89] S. Freud. *An outline of psycho-analysis.* W. W. Norton & Co., New York, 1940/1989.

[Fre02] S. Freud. *Abriss der Psychoanalyse.* Fischer Taschenbuchverlag, 1940/2002.

[Fri04] N. Frijda. The psychologists point of view. In *Handbook of Emotions*, chapter 5, 59–74. The Guilford Press, second edition, 2004.

[GH01] S. C. Gadanho and J. Hallam. Emotion-triggered Learning in Autonomous Robot Control. *Cybernetics and Systems*, 32(5):531–559, 2001.

[GM04] J. Gratch and S. Marsella. A Domain-independent Framework for Modeling Emotion. *Journal of Cognitive Systems Research*, 5(4):269–306, 2004.

[Gru07] A. Gruber. Neuro-Psychoanalytically Inspired Episodic Memory for Autonomous Agents. Master's thesis, Institute of Computer Technology, Technival University Vienna, 2007.

[Har90] S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

[Har02] S. Harnad. Minds, Machines, and Searle 2: What's Right and Wrong about the Chinese Room argument. In *Views into the Chinese room: New essays on Searle and artificial intelligence*, eds. J. Preston and M. Bishop. Oxford, Clarendon, 2002.

[Hay98] S. Haykin. *Neural Networks: A Comprehensive Foundation (2nd Edition).* Prentice Hall, 1998.

[Heb49] D. Hebb. *Organization and Behaviour.* New York, Wiley, 1949.

[Jua99] A. Juarrero. *Dynamics in Action: Intentional Behavior as a Complex System.* The MIT Press, 1999.

[J.v82] J.v.Uexküll. The Theory of Meaning. *Semiotica*, 42(1):25–87, 1940/1982.

[Kan99] E. Kandel. Biology and the future of psychoanalysis: A new intellectual framework for psychiatry revisited. *Am. J. Psychiatry*, 156:505–524, 1999.

[Kel91] J. P. Kelly. The neural basis of perception and movement. In *Priciples of Neural Science*, eds. E. H. Kandel, J. H. Schwartz, and T. M. Jessell, 283–295. Prentice-Hall, London, 1991.

[KK92] S. Kosslyn and O. Koenig. *Wet Mind.* New York: Free Press, 1992.

[Kur90] Ray Kurzweil. *The Age of Intelligent Machines.* MIT Press, 1990.

[Lar03] S. D. Larson. Intrinsic Representation: Bootstrapping Symbols From Experience. 2003.

[LB07] B. Lorenz and E. Barnard. A brief overview of artificial intelligence focusing on computational models of emotions. In *Proceedings of ENF 2007 – 1st Int. Engineering & Neuro-Psychoanalysis Forum*, 1–12, July 23, Vienna, 2007.

[LBP02] M. Leuzinger-Bohleber and R. Pfeifer. Embodied Cognitive Science und Psychoanalyse: Ein interdisziplinärer Dialog zum Gedächtnis. In *Psychoanalyse im Dialog der Wissenschaften: Europäische Perspektiven*, ed. P. Giampieri-Deutsch, 242–270. Stuttgart, Kohlhammer, 2002.

[LDS01] D. Loy, D. Dietrich, and H. Schweinzer. *Open Control Networks.* Kluwer Academic Publishers, Boston/Dordrecht/London, 2001.

[Led96] J. E. Ledoux. *The Emotional Brain.* Simon and Schuster, New York, 1996.

[Led03] J. Ledoux. *A Synaptic Self: How Our Brains Become Who We Are.* Penguin Book, 2003.

[Lur73] A. R. Luria. *The Working Brain: An Introduction to Neuropsychology.* Harmondsworth, Penguin Books, 1973.

[Mar82] D. Marr. *Vision.* W. H. Freeman, New York, 1982.

[MB93] D. McFarland and T. Boesser. *Intelligent Behavior in Animals and Robots.* MIT Press, Cambridge, MA, 1993.

[McD28] W. McDougall. Emotion and feeling distinguished. In *Feelings and Emotions. The Wettenberg Symposium*, ed. M. L. Reymert, 200–205. Worcester, MA: Clark University Press, 1928.

[McD60] W. McDougall. *An introduction to social psychology.* London, Methuen, 1908/1960.

[McD69] W. McDougall. *An outline of psychology.* London, Methuen, 1928/1969.

[MF95] D. Moffat and N. Frijda. Where there's a Will there's an Agent. *Intelligent Agents: ECAI-94 Wokshop on Agent Theories, Architectures, and Languages*, 245–260, 1995.

[Min91] M. Minsky. Logical versus Analogical or Symbolic versus Connectionist or Neat versus Scruffy. *AI Magazine*, 12(2):34–51, 1991.

[Min06] M. Minsky. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind.* Simon and Schuster, New York, 2006.

[Mow60] O. H. Mowrer. *Learning Theory and Behavior.* John Wiley, New York, 1960.

[MW06] H. J. Markowitsch and H. Welzer. *Das autobiographische Gedächtnis.* Klett-Cotta, 2006.

[Nel96] K. Nelson. *Language in cognitive development.* Cambridge University Press, Cambridge, 1996.

[New90] A. Newell. *Unified Theories of Cognition.* Harvard University Press, Cambridge, MA, 1990.

[NL04] A. Nuxoll and J. E. Laird. A Cognitive Model of Episodic Memory Integrated With a General Cognitive Architecture. In *Proc. of the 6th Int. Conf. on Cognitive Modeling*, 220–225. Mahwah, NJ: Lawrence Earlbaum, 2004.

[NOR03] D. A. Norman, A. Ortony, and D. M. Russell. Affect and machine design: Lessons for the development of autonomous machines. *IBM SYSTEMS JOURNAL*, 42 (1):38–44, 2003.

[NS76] A. Newell and H. Simon. Computer science as empirical inquiry: symbols and search. *Communication of ACM*, 19:113–126, 1976.

[OCC88] A. Ortony, G. L. Clore, and A. Collins. *The Cognitive Structure of Emotions.* Cambridge University Press, Cambridge, MA, 1988.

[O'K90] J. O'Keefe. A computational theory of the cognitive map. *Progress in Brain Research*, 83:301–312, 1990.

[Ort03] A. Ortony. On making believable emotional agents believable. In *Emotions in humans and artifacts*, eds. R. Trappl and P. Petta. MIT Press, Cambridge, MA, 2003.

[Pal04] P. Palensky. Requirements for the Next Generation of Building Networks. In *Proceedings of the International Conference on Cybernetics and Information Technologies, Systems and Applications, (ISAS CITSA 2004), Orlando, Florida*, 225–230, 2004.

[Pan98] J. Panksepp. *Affective Neuroscience: The Foundations of Human and Animal Emotions.* Oxford University Press, New York, 1998.

[Pan03] J. Panksepp. At the interface of the affective, behavioral, and cognitive neurosciences: Decoding the emotional feelings of the brain. *Brain and Cognition*, 52:4–14, 2003.

[Pan05] J. Panksepp. Affective consciousness: Core emotional feelings in animals and humans. *Consciousness and Cognition*, 14:30–80, 2005.

[Pei06] C. S. Peirce. Pragmatics in Retrospect: a last formulation. In *The Philosophical Writings of Peirce*, ed. J. Buchler. New York, 1906.

[PLC07] P. Palensky, B. Lorenz, and A. Clarici. Cognitive and Affective Automation: Machines Using the Psychoanalytic Model of the Human Mind. In *Proceedings of ENF 2007 – 1st Int. Engineering & Neuro-Psychoanalysis Forum*, 49–73, July 23, Vienna, 2007.

[PLD05] G. Pratl, B. Lorenz, and D. Dietrich. The Artificial Recognition System (ARS): New Concepts for Building Automation. In *Proceedings of 6th IFAC Int. Conf. on Fieldbus Systems and their Applications*, 48–55, Puebla, Mexico, Nov 14-15, 2005.

[Pra06] G. Pratl. *Symbolization and Processing of Ambient Sensor Data.* PhD thesis, Institute of Computer Technology, Vienna University of Technology, 2006.

[PS99] R. Pfeifer and C. Schreier. *Understanding Intelligence.* MIT Press, 1999.

[Pyl84] Z.W. Pylyshyn. *Computation as Cognition.* MIT/Bradford, Cambridge, MA, 1984.

[QDNR99] T. Quick, K. Dautenhahn, C. Nehaniv, and G. Roberts. The Essence of Embodiment: A Framework for Understanding and Exploiting Structural Coupling Between System and Environment. In *Proceedings CASYS9́9, Third Int. Conf. on Computing Anticipatory Systems*, HEC, Liège, Belgium, 1999.

[Rei96] W. S. Reilly. *Believable Social and Emotional Agents.* PhD thesis, School of Computer Science, Carnegie Mellon University, 1996.

[Rie81] R. Riedl. *Biologie der Erkenntnis: Die stammesgeschichtlichen Grundlagen der Vernunft.* Paul Parey, Berlin, Hamburg, 1981.

[RKLC80] A. Reber, S. Kassin, S. Lewis, and G. Cantor. On the relationship between implicit and explicit modes in the learning of a complex rule structure. *Journal of Experimental Psychology: Human Learning and Meory*, 6:492–502, 1980.

[RLD$^+$07] C. Roesener, R. Lang, T. Deutsch, G. Gruber, and B. Palensky. Action planning model for autonomous mobile robots. In *INDIN07*, 2007.

[RLFV06] C. Roesener, B. Lorenz, G. Fodor, and K. Vock. Emotional Behavior Arbitration for Automation and Robotic Systems. In *Proc. of the 4th IEEE Int. Conference on Industrial Informatics*, 2006.

[RM86] D. E. Rumelhart and J. L. McClelland. *Parallel Distributed Processing: Exploration in the Microstructure of Cognition.* MIT/Bradford, Cambridge, MA, 1986.

[RN04] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach (2nd Edition).* Prentice Hall, 2004.

[Roc95] L. M. Rocha. Selected Self-Organization and the Semiotics of Evolutionary Systems. In *Evolutionary Systems: Biological and Epistemological Perspectives on Selection and Self-Organization*, eds. S. Salthe, G. Van de Vijver, and M. Delpos, 341–358. Kluwer Academic Publishers, 1995.

[Roc01] L. M. Rocha. Evolution with material symbol systems. *Biosystems*, 60 (1-3):95–121, 2001.

[Roe07] C. Roesener. *Adaptive Behavior Arbitration for Mobile Service Robots in Building Automation.* PhD thesis, 2007.

[Rot03] G. Roth. *Fühlen, Denken, Handeln: Wie das Gehirn unser Verhalten steuert.* Suhrkamp, Frankfurt am Main, 2003.

[Rot04] G. Roth. Das Verhältnis von bewusster und unbewusster Verhaltenssteuerung. *Psychotherapie Forum*, 12:59–70, 2004.

[Rus03] G. Russ. *Situation Dependent Behavior in Building Automation.* PhD thesis, Vienna University of Technology, 2003.

[SC81] A. Sloman and M. Croucher. Why robots will have emotions. In *Proceedings of the 7th Int. Joint Conference on Artificial Intelligence*, 197–202, 1981.

[SC05] A. Sloman and J. Chappell. The Altricial-Precocial Spectrum for Robots. In *Proceedings IJCAI'05, Edinburgh*, 1187–1192, 2005.

[Sch70] S. Schachter. Some extraordinary facts about obese humans and rats. *American Psychologist*, 26:129–144, 1970.

[Sch97] D. L. Schacter. *Searching for Memory: The Brain, the Mind, and the Past.* Basic Books, 1997.

[SCS04] A. Sloman, R. Chrisley, and M. Scheutz. The architectural basis of affective states and processes. In *Who Needs Emotions? The Brain Meets the Machine*, eds. M. Arbib and J.-M. Fellous. Oxford University Press, Oxford, New York, 2004.

[SCZ05] R. Sun, L. A. Coward, and M. J. Zenzen. On Levels of Cognitive Modeling. *Philosophical Psychology*, 18(5):613–637, 2005.

[Sea80] J. Searle. Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3:417–424, 1980.

[Sim69] Herbert A. Simon. *The Sciences of the Artificial.* MIT Press, Cambridge, MA, first edition, 1969.

[Slo97] A. Sloman. What sort of control system is able to have a personality? In *Creating Personalities for Synthetic Actors*, eds. R. Trappl and P. Petta, 166–208. Lecture notes in AI, Springer Berlin Heidelberg New York, 1997.

[Slo01] A. Sloman. Beyond shallow models of emotion? *International Quarterly of Cognitive Science*, 2(1):177–198, 2001.

[Smo88] P. Smolensky. On the Proper Treatment of Connectionism. *Journal of Behavioral and Brain Sciences*, 11:1–74, 1988.

[Sol97] M. Solms. What is Consciousness? *Journal of the American Psychoanalytic Association*, 45:681–778, 1997.

[Sol04] M. Solms. Freud Returns. *Scientific American*, May:82–88, 2004.

[Sol07] M. Solms. What is the 'mind'. A neuro-psychoanalytical approach. In *Proceedings of ENF 2007 – 1st Int. Engineering & Neuro-Psychoanalysis Forum*, 21–24, July 23, Vienna, 2007.

[SS98] J. Sandler and A. M. Sandler. *Internal objects revisited.* Karnac Books, London, 1998.

[SST05] R. Sun, P. Slusarz, and C. Terry. The Interaction of the Explicit and the Implicit in Skill Learning: A Dual-Process Approach. *Psychol ogical Review*, 112(1):159–192, 2005.

[ST02] M. Solms and O. Turnbull. *The Brain and the Inner World.* Karnac/Other Press, Cathy Miller Foreign Rights Agency, London, England, 2002.

[Ste98] R. J. Sternberg. *In Search of the Human Mind.* Harcourt Brace & Co., Orlando, FL, second edition, 1998.

[Sto97] T. Stonier. *Information and the Internal Structure of the Universe: An Exploration into Information Physics.* Springer, 1997.

[Str03] K. T. Strongman. *The Psychology of Emotion.* John Wiley, fifth edition, 2003.

[Tod82] M. Toda. *Man, Robot and Society.* Martinus Nijhoff Publishing, 1982.

[TS96] E. Thelen and L. B. Smith. *Symbolic Thought in a Dynamic Cognition: A Dynamic Systems Approach to the Development of Cognition and Action.* MIT Press, 1996.

[TSSS01] E. Thelen, G. Schoner, C. Scheier, and L. B. Smith. The Dynamics of Embodiment: A Field Theory of Infant Perservative Reaching. *Behavioral and Brain Sciences*, 24:1–86, 2001.

[Tul83] E. Tulving. *Elements of Episodic Memory.* Clarendon Press, Oxford, 1983.

[Tul85] E. Tulving. How Many Memory Systems Are There? *American Psychologist*, 40:385–398, 1985.

[Tul93a] E. Tulving. Memory concepts 1993: Basic and clinical aspects. In *Human memory*, eds. P. Andersen, O. Hvalby, O. Paulsen, and B. Hokfelt, 27–45. Amsterdam: Elsevier, 1993.

[Tul93b] E. Tulving. What is Episodic Memory. *Current Directions in Psychological Science*, 2:67–70, 1993.

[Vel98] J. Velàsquez. Modeling emotion-based decision-making. In *Emotional and Intelligent: The Tangled Knot of Cognition*, 164–169. D. Cañamero, 1998.

[VTR92] F. J. Varela, E. T. Thompson, and E. Rosch. *The Embodied Mind: Cognitive Science and Human Experience.* MIT Press, 1992.

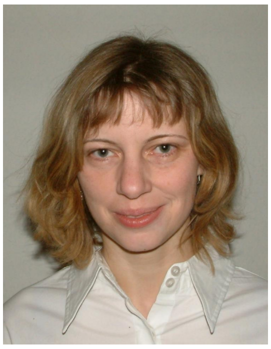
Internet References

[nps07] Neuro-psychoanalysis, International Neuropsychoanalysis Society, 2007. http://www.neuro-psa.org.uk.

[num05] Numenta Platform for Intelligent Computing, Numenta Inc., 2005. http://www.numenta.com.

Curriculum Vitae

Dipl.-Ing. Brigitte Palensky



Personal Data

Family status:	Married, former name Brigitte Lorenz
Birth:	May 17, 1970, Mistelbach, Austria
Nationality:	Austrian
Languages:	German, English, French

Education

2004 – 2007	Ph.D. studies in AI and Cognitive Science, Vienna University of Technology
1998 – 2001	Studies in Computer Science, Vienna University of Technology, focus on logics, evolutionary algorithms, self-organization
1989 – 1998	M.Sc. (Dipl.-Ing.) in Technical Physics, Vienna University of Technology, thesis: *Wavelet analysis - Theory and Practical Application,* passed with distinction
1984 – 1989	Senior high school with special focus on business and administration (*BHAK*), Mistelbach, Austria, graduation (*Matura)* with distinction

Professional and Research Experience

2004 – 2007	Research at Institute of Computer Technology, Vienna University of Technology IRON (Integrated Resource Optimization Network) project: *Increasing the energy efficiency with recent communication and IT technologies* ARS (Artificial Recognition System) project: *Translating the neuro-psychoanalytic view on the human mind into technical terms*
2001 –2003	Envidatec GmbH – Energiedienstleistungen, D-21079 Hamburg Development, test, and documentation of IT-based energy optimization services

Scientific Publications

M. Stadler, P. Palensky, B. Lorenz, M. Weihs, C. Roesener: Integral Resource Optimization Networks and their techno-economic constraints. *Int. Jour. on Distributed Energy Systems*, 11(4): 299–319, 2005.

B. Lorenz, C. Roesener, P. Palensky: Project IRON - Integral Resource Optimization Network Study. In *4th Int. Conf. on Energy Economics* (IEWT 2005), Vienna, 144–145, 2005.

G. Pratl, B. Lorenz, D. Dietrich: Artificial Recognition System (ARS): New Concepts for Building Automation. In *6th IFAC Int. Conf. on Fieldbus Systems and their Applications* (FET'05), Puebla, Mexico, 48–55, 2005.

C. Roesener, B. Lorenz, G. Fodor, K. Vock: Emotional Behavior Arbitration for Automation and Robotic Systems. In *4th Int. IEEE Conf. on Industrial Informatics* (INDIN06), Singapore, 6 pages, 2006.

C. Roesener, P. Palensky, M. Weihs, B. Lorenz, M. Stadler: Integral Resource Optimization Networks - a new solution on power markets. In *3rd Int. IEEE Conf. on Industrial Informatics* (INDIN 2005), Perth, Australia, ISBN: 0-7803-9095-4, 6 pages, 2005.

M. Weihs, H. Bruckner, B. Lorenz, P. Palensky: Integral Resource Optimization Network. In *9th Symposium on Energy Innovation*, Graz, Austria, 12 pages, 2006.

B. Lorenz, E. Barnard: A brief overview of artificial intelligence focusing on computational models of emotions. In *1st Int. IEEE Engineering & Neuro-Psychoanalysis Forum*, July 23, 2007, Vienna, Austria, 1–12, 2007.

P. Palensky, B. Lorenz, A. Clarici: Cognitive and affective automation: Machines using the psychoanalytic model of the human mind. In *1st Int. IEEE Engineering & Neuro-Psychoanalysis Forum*, July 23, 2007, Vienna, Austria, 49–73, 2007.