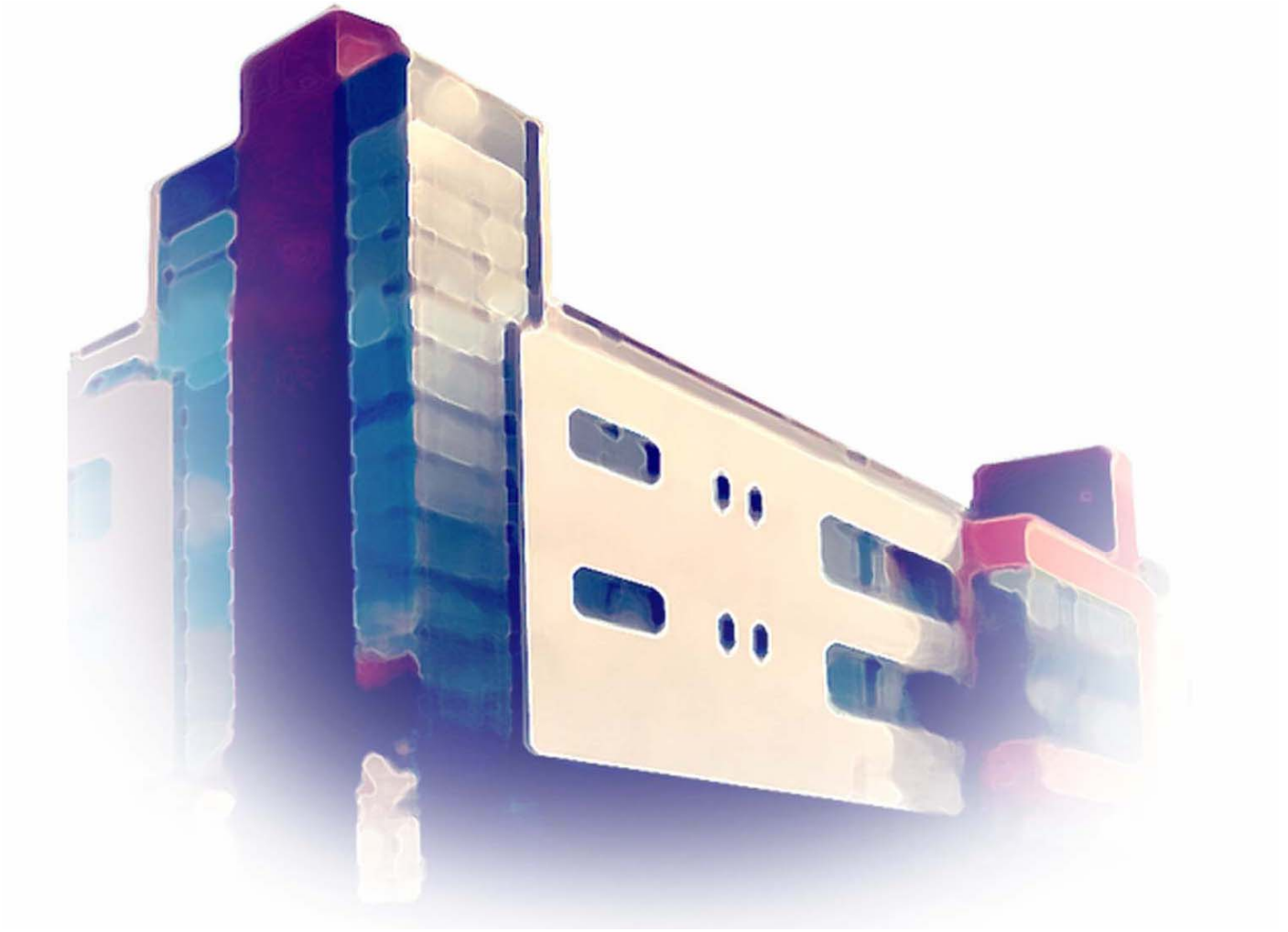
Michael Ernst Elbers

Intent on Acting – Acting on Intentions

A comparison of Belief-Desire-Intention (BDI) approaches and Ψ-theory (Psi) with respect to the agent architectures proposed, focussing on intention and utility in the regulation of action



PICS

Publications of the Institute of Cognitive Science

Volume 01-2006

ISSN: 1610-5389

Series title: PICS
Publications of the Institute of Cognitive Science

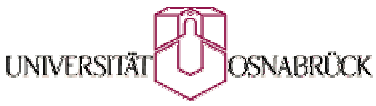
Volume: 01-2006

Place of publication: Osnabrück, Germany

Date: January 2006

Editors: Kai-Uwe Kühnberger
Peter König
Petra Ludewig

Cover design: Thorsten Hinrichs



Universität Osnabrück
Master's Programme Cognitive Science

Master's Thesis

Intent on Acting — Acting on Intentions

A comparison of Belief-Desire-Intention (BDI) approaches and Ψ-theory (Psi)
with respect to the agent architectures proposed,
focussing on intention and utility in the regulation of action

by
Michael Ernst Elbers

Submitted in partial fulfillment of the requirements for the degree of
Master of Science in Cognitive Science

First Examiner: Prof. Dr. Kai-Uwe Kühnberger
Second Examiner: Dipl. Inform. Claus Hoffmann

Abstract

This thesis provides a comparison of two approaches to explaining practical reasoning and behaviour in natural agents, particularly humans, and modeling it in artificial agents. The concept at the centre of both these approaches is the intention. Scrutinizing the nature of intentions and the role that they are attributed to in the regulation of action forms the focus of attention, with a particular emphasis on the structural and procedural elements that either approach proposes. Moreover, the notion of utility as a key concept in goal-directed (practical) reasoning is investigated in this regard. The first approach is the Belief-Desire-Intention (BDI) theory originally proposed by Bratman and its most widely studied model called the Procedural Reasoning System (PRS). It is the predominant approach to constructing rational artificial agents in the research community today. The second approach is Dörner's Ψ-theory, an agent model in itself, which combines assumptions concerning perception, memory, cognition, motivation and emotion.

Contents

List of Figures

"Die Absicht ist die Seele der Tat."
—German Proverb

1 Introduction

The intention is the soul of the deed. This popular german proverb, which may just as well be translated as *the intention is the spirit of the deed*, contains a fact well known to all of us: human behaviour is normally the result of human intentions. Usually, when we want to predict the behaviour of someone else, we try to find out what he intends. Likewise, when we analyze and judge someone's behaviour in retrospect, we tend to ask ourselves what his intentions were. In fact, the notion of intention is one we so frequently encounter in everyday life that few of us would hesitate to claim that they knew what an intention is. Perhaps even fewer people still would deny that intentions contribute considerably to how we behave, that they are a key part of our practical reasoning, i.e. reasoning which is directed towards action. However, most of us would be silenced upon further inquiry of how intentions actually come to be in the first place or how precisely they influence what we do. Whatever our exact understanding of the nature of intentions, it is a common assumption that they form a fundamental part of what makes us tick — they are the *soul* of many of our actions.

It should not come as a surprise that intentions have attracted the interest of many scientifically minded individuals over the past centuries and millenia. It should be equally unsurprising that the vast majority of theories that have been developed in this area are the work of philosophers, psychologists and sociologists. Unfortunately, from a scientific perspective these disciplines have tended to have fundamental disadvantages compared to the exact sciences, because they are *not exact*. Above all, they have not been in a position to study directly or model adequately the machinery that is assumed to produce what they all theorize about, namely the mind — until recently. Modern imaging techniques such as the EEG[1] or the fMRT[2] now allow the study of the working brain almost in real-time and the advent of powerful computing devices has opened up the possibility to actually model the brain's structure and mechanisms. Even though both still lack some precision, we are beginning to understand in much more detail how the 'machine' inside our heads works. We can watch it at work and we can simulate its behaviour computationally on the basis of the theories we have for it. This is the state at the heart of cognitive science research today, a field which draws from the humanities and the exact sciences alike.

The knowledge gain concerning the underlying mechanisms of nature has always evoked the invention of useful, useless and harmful devices that would not have been

[1] Electroencephalogram
[2] functional Magnetic Resonance Tomography

conceivable without this new knowledge. Understanding the machinery of the mind seems to be no exception to this rule and the development of practical applications, mainly in artificial intelligence, is its expression. If we are able to computationally generate forms of lifelike behaviour for the purpose of testing our theories then we can also put these forms of behaviour to some other use. This line of reasoning has sparked the creation of artificial agents, computational creatures that exhibit some form of intelligent and autonomous behaviour, and which have already been applied successfully to a number of application domains. Agent research is still only in its infancy, but the incorporation of complex psychological, philosophical or sociological theories of behaviour promises the development of very powerful devices and it has already shown a small fraction of its potential.

In this thesis we compare two approaches to agent development that are founded on two different theories. The theories are united in the fact that they see the *intention* as the core concept in the generation of rational behaviour, particularly in humans. *Belief-Desire-Intention* is a philosophical theory by Michael E. Bratman which has become the dominant theory in the agent research community. Unlike many other philosophers who defend their theory merely by thought experiment, Bratman was among the first to try and implement his theory in an artificial practical reasoning system even though the theory itself does not make any structural claims. Its most famous incarnation is the *Procedural Reasoning System* and its descendants which will serve us as the reference implementation for the comparison. Bratman's theory explains how we form new intentions from the reasons our beliefs, our desires and our already formed intentions give us. It also shows when we maintain them, when we drop them them, how they influence our planning and how they determine the level of stability and rationality of our behaviour.

The second theory is the so-called Ψ-theory by psychologist Dietrich Dörner. Dörner is a proponent of *synthetic psychology*, i.e. he believes that the development of a psychological theory should go hand in hand with the development of a computational model. Consequently, his theory is generative, i.e. it is realized directly as an artificial agent. The theory is profoundly inspired by ideas from the field of cybernetics such that it postulates an agent to have basic needs and urges which motivate its behaviour. Furthermore regulative processes, interpreted as emotions, influence the way behaviour as well as the cognitions that precede it occur. Yet, being the centre of the theory, intentions ultimately determine what the agent does; they organize the response of the agent to its environment. The theory as a whole is very complex and extends not only to motivation and emotion, but it also includes assumptions concerning knowledge representation, memory and perception. Implementations of Ψ-theory are very rare, and, whilst they have been applied in psychological experiments, the theory has not yet had any effect on the development of agents for practical applications.

1.1 Aims of the Thesis

The primary aim of this thesis is to compare the two approaches of BDI and Ψ structurally and procedurally, i.e. on the basis of the concrete agent architectures they propose in PRS-like agents and Ψ-agents respectively. This comparison shall focus on the concept of *intention* and its role in the generation and regulation of goal-directed behaviour. To this end the question of how an agent is able to consider some goals as more desirable than others must be answered which is traditionally associated with the notion of *utility*. The comparison is to show in how far the conception of intention is similar for both approaches and in how far the inclusion of motivational and emotional aspects by Ψ-theory influences this conception and accounts for the differences that exist.

We derive two further goals from our primary focus. Firstly, in preparation of the architectural comparison, the aim is to show in how far the similarities between the two architectures are owed to the fact that Ψ-theory already complies with the criteria for beliefs, desires and intentions which Bratman proposed in his theory of BDI. Secondly, as a result of the architectural comparison, arguments for and against a hybrid system consisting of parts both from the PRS and Ψ shall be laid out and the specific constraints under which such a hybrid makes sense shall be discussed.

It is important to stress that the emphasis of this thesis is on the two architectures as computational frameworks for developing agents, in particular for practical applications. This means that questions concerning the neurobiological or psychological plausibility of the underlying theories are largely skirted albeit not wholly neglected.

1.2 Outline of the Thesis

The next chapter (chapter 2) *Terms and Concepts* introduces the main ideas, like agency, mental attributes, intentionality and environment, that are necessary for an understanding of this thesis.

The following two chapters then present the two approaches: *BDI – Practical Reasoning in Rational Agents* (chapter 3) and *Ψ – Practical Reasoning in Motivated and Emotional Agents* (chapter 4). The state of the art concerning the theoretical underpinnings as well as the proposed agent architectures is described, applications and open research questions are briefly outlined.

A Constructive Comparison (chapter 5) is dedicated to the comparison of the two approaches and is subdivided mainly according to the three aims laid out for this thesis: Section 5.2 deals with the question whether or not the Ψ approach can be regarded as a form of BDI, section 5.4 provides the architectural comparison, and section 5.5 considers the possible mutual enhancements for BDI as realized in the PRS and Ψ.

The thesis closes with a *Summary and Conclusions* (chapter 6).

The beginning of wisdom is to call things by their right names.
—Chinese Proverb

2 Terms and Concepts

This chapter introduces the main concepts necessary for an understanding of this work. It will start by giving a first notion of the central term *agent*. It will then draw upon the insight gained from an apparent dilemma: On the one hand, the concept of agency is useful as a metaphor, while it is, on the other hand, a scientific necessity to define it, at least when applied in well-defined domains. This principal issue will motivate putting a focus onto the properties of agents, and first and foremost among them *autonomy*. Having developed a notion of what constitutes an agent, we will take a look at the main motivations for describing and designing artificial systems as agents, and sketch some of the areas of application where this has already been done. In the wake of the core concepts of agency more terms in need of explanation will arise which must be mentioned but cannot be dealt with in great detail. Among them are some elementary conceptions such as *agent architecture* and *agent system* as well as those necessary for characterizing the types of *environments* and the types of *problems* an agent can be faced with.

2.1 Agents

2.1.1 Basic Ideas and Notions

The term *agent* plays a central role in this thesis. But what exactly does it mean? If we start by asking for the literal meaning of the word *agent*, we are told that it stems from the Latin word *agens* which means "driving force" or "acting entity". Furthermore, we are informed that it is used to denote someone who acts as a delegate, a representative, a mediator or a facilitator for some other person or institution (Duden, 1994). This reflects our immediate understanding very well and captures two most important aspects of the common usage of the word, namely "1) one who acts, or who can act, and 2) one who acts in place of another with permission" (cf. Franklin and Graesser, 1996). Furthermore, the translation as "driving force" seems to hint at another important aspect attributed to agency: the autonomy of the acting entity, i.e. that it is the main locus of the reasons for its actions. We immediately associate issues like these with the word *agent*, because they are part of our intuition of the term[3]. In fact, it is this intuition, this metaphorical understanding of agency that has contributed most to the proliferation of the agent concept in the scientific

[3]Many people, not familiar with its usage in AI or cognitive science, maybe think of a secret agent first, like James Bond, for instance. It is the fact that James Bond works on behalf of someone else, but in his own special way, of course, that makes him an agent. The same holds for your travel agent, an artist's agent or an estate agent.

community and beyond. The metaphor has proved fruitful in the discussion on artificial intelligent systems.

Ultimately however, science demands clear cut concepts and, if possible, mathematical definitions of them. Bernedo Schneider (2004) spots a general cyclic process at work that may eventually lead to a unified definition — and the rendering redundant of the metaphor — by a perpetual sequence of definition attempts and tests. The focus of this process is, or should be, he concludes, on an agent's necessary and sufficient properties. These emerge and amalgamate at the intersection of agent development in heterogeneous fields of practical application and the theoretical undertakings in the scientific community. It seems doubtful though that a general definition is possible or desirable in the case of the concept of an *agent* (for a more elaborate discussion, see Bradshaw, 1997, pp.5-11). Bernedo Schneider (2004) also points out the oddness of wanting to define a metaphor, i.e. define an image, in the first place. In this context Russell and Norvig (1995, p.33) state that "[t]he notion of an agent is meant to be a tool for analyzing systems, not an absolute characterization that divides the world into agents and non-agents." Franklin and Graesser (1996) refer to this idea when adding that "[t]he only concepts that yield sharp edge categories are mathematical concepts, and they succeed only because they are content free. Agents 'live' in the real world (or some world), and real world concepts yield fuzzy categories." At the present stage of the scientific endeavour, agreement on one absolute definition seems distant. Irrespective of this principal issue, many attempts have been made and they manifest themselves in a multitude of different and sometimes contradictory definitions, as has been pointed out elsewhere (cf. Franklin and Graesser, 1996; Bradshaw, 1997; Wooldridge, 2001; Bernedo Schneider, 2004, and many more).

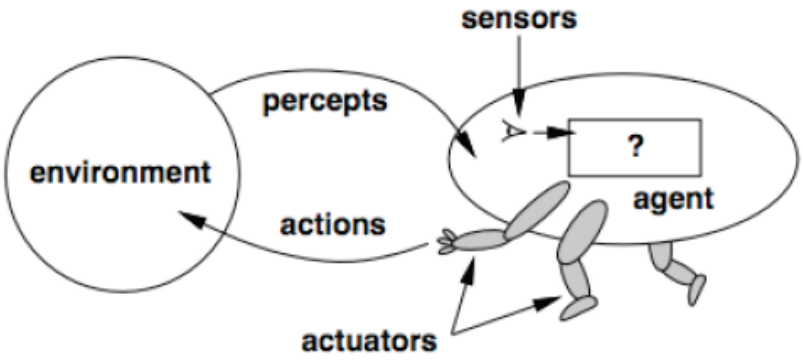


Figure 1: A basic agent. Taken from Russell and Norvig (1995).

The focus of attention of most definitions is on agent *properties*. Figure 1 shows a sketch of an agent and its most elementary properties. It can sense its environment via sensors (perception) and act upon it via effectors/actuators (action) — it is embodied,

and embedded in an environment[4]. This minimal requirement is the solid basis, but no more than that. The scientists of the field have put forward an impressively long list of agent properties describing agents as being *reactive*, *autonomous*, *goal-oriented*, *temporally continuous*, *communicative*, *learning*, *mobile*, *flexible*, and the list goes on (cf. Franklin and Graesser, 1996). Even from a naïve perspective one soon realizes that some seem to be subsumable under others, that they are part of or prerequisites for others. Both Franklin and Graesser (1996) and Wooldridge and Jennings (1995) collapse the host of properties down to four:

- *autonomy*
- *reactivity*
- *proactiveness* (or *goal-oriented*)

and — in the last point they differ —

- *social ability* (Wooldridge and Jennings, 1995)
- *temporal continuity* (Franklin and Graesser, 1996).

Wooldridge (2000) explains the terms as follows. "At its simplest, autonomy means nothing more than being able to operate independently. (...) However, autonomy is usually taken to have a slightly stronger meaning than this. (...) [W]e regard a rational agent as 'having its own agenda' ..." (pp.3-4). Reactivity is meant as "being responsive to changes in the environment" (p.4) and "[p]roactiveness means being able to exhibit goal-directed behavior" (p.4). In section 2.1.3 we will discuss these properties in more detail. Social ability is described as "the ability to interact (through negotiation and cooperation) with other self-interested agents" (Wooldridge, 2000, p.5). Finally, the property of temporal continuity claims that an agent needs to be "a continually running process", unlike a normal program which "runs once and then goes into a coma, waiting to be called again" (Franklin and Graesser, 1996). Very many definitions demand *autonomy*, *reactivity* and *pro-activeness* and they may be regarded as the lowest common denominator at the current state of the art. They form one of the weakest definitions of agency. The more attributes we include the stronger the definition becomes and the more things we can exclude as non-agents.

The demand for social ability already suggests that agents often interact with other agents in the furtherance of their goals. Well, whose goals exactly? For simplicity, let

[4]Note that the issue of having sensors and effectors, having a body of some sort, i.e. the relation of an agent to its environment which may possibly include other agents, is a very delicate one in its own right. It is part of the debate on *embodiment* and *situatedness/embeddedness*. We will have to neglect such matters in this thesis, because they are beyond its limits. Please refer to Wilson and Keil (2001) for further reading on this subject.

us assume that agents pursue the goals of their users in some way (generally humans) — or subgoals thereof — and not their very own goals (They can! We will see this later on). Even so, there is one important distinction to be made, namely between what we will call a *multi-agent system* and a *society of agents*[5]. When viewed from afar, a *multi-agent system* can also be seen as being a single agent with certain specific goals toward which the agents that constitute it (subagents) are working. Very often, they will have been explicitly designed for this purpose, i.e. to interact in such a way that the overall system may achieve its goals as best as possible. A multi-agent system will persist over some time and this is the lifetime of the single agent it forms. A *society of agents*, on the other hand, tends to be a somewhat transient construct. Agents join and leave the group autonomously. They, too, may pursue a common goal and intelligent group behaviour may emerge. Yet, once that goal is achieved or individual agents believe that it can no longer be achieved, that it is no longer feasible for them to stay in the group etc., they disperse.[6] (see Franklin and Graesser, 1996; Huhns and Stephens, 2001)

So, obviously, agent properties can serve in making classifications, that is for laying down taxonomies (cf. Franklin and Graesser, 1996). It is, however, somewhat arbitrary which dimensions one chooses to classify agents and how these are nested to produce more elaborate subclasses. Other frequently used dimensions for classification include: control structures, environments or applications. For our purposes, these subtleties may be neglected. Nevertheless, there are two attributes we need to at least mention, since they are frequently used to classify agents: *intelligence* and *rationality*. Both concepts are very hard to pin down exactly, the former more so than the latter, and it will not be tried here. Our intuitive understanding of the term *intelligence* will suffice. As for *rationality* we may content ourselves with a short quote from Russell and Norvig (1995, p.4) who state that rationality is an "*ideal* concept of intelligence (...). A system is rational if it does the right thing" (or concisely: if it "maximizes expected utility"[7]) and Wooldridge (2000, p.1) who puts it this way: "An agent is said to be *rational* if it chooses to perform actions that are in its own best interests, given the beliefs it has about the world"[8]. In fact, both concepts are so intricately related to the concept of agency that many people involved in agent development and research use the expressions, *intelligent agent*, *rational agent* or just *agent* synonymously.

---

[5]Note: this is not exactly the usual usage of the terms. In fact, societies of agents are normally seen as a subclass of multi-agent system. See the difference in the statement by Franklin and Graesser (1996): "[i]n addition to multiagent systems that can reasonably be viewed as constituting a single agent, other multiagent system are better classified as societies of agents".

[6]It is worthwhile noting the difference between what is meant by *agent* and *society of agents* in this context and what people like Minsky and Kokinov mean by *agent* and *society of mind* (cf. Minsky, 1986; Kokinov, 1994).

[7]This notion is part of the definition of a rational agent in decision theory.

[8]For further reading on the issue of defining *intelligence* see Urchs (2002, especially pp.49-53) or Wilson and Keil (2001). For discussions on the concept of *rationality* the reader may refer to Wilson and Keil (2001), Bernedo Schneider (2004), (Bratman, 1987, chapter 4) and Russell and Norvig (1995).

Before we close this section, there is one more important issue to bear in mind, especially in the following discussion: It is that most properties of agents are not absolute but gradual in that an agent does not normally have them or not, but it has each one to a certain degree. From this follows that an agent is always an agent to a certain degree. This important issue is also reflected in the fact that agency also depends on the environment the agent is set in or rather the relationship between its environment, its own sensory and motor organs and any additional capabilities it has. Consider a fish, for instance: in its accustomed environment, the water, we can regard it as an agent, while changing its environment, say to the streets of your home town, will render it unable to act appropriately, and eventually it will even die. Most of all it lacks appropriate effectors, like legs for example, but its gills aren't of much use either. Of course, the effects of being ill-equipped will not always be as dramatic as in the case of the unfortunate fish, but the general point is made.

2.1.2 Motivations and Applications

Before continuing to discuss the properties of agency, we should first take a look at what it is that motivates researchers and developers to build artificial agents and where agent solutions have already been applied. In order to fully appreciate both the great enthusiasm of the agent research community and the magnitude of its task, it is worthwhile to establish the bigger picture.

Humans have a long evolutionary tradition of devising and utilizing tools (Bernd, H. and Hippchen, T. and Jüngst, K. and Strittmatter, P., 2000). First, men made *hardware tools* like the hand-axe or the spear. Among them were also *gathering tools*, such as the telescope, the microscope, the thermometer or x-ray devices, which allowed man to overcome the limited capabilities of his sensory organs. Subsequently, *software tools* were developed making it possible to understand and explain structures and relationships that were not visible otherwise. This is the nature of atomic models and models of molecules, plan and process diagrams and the modern mapping techniques like mind mapping, concept mapping or their many relatives.

Tools, in the ordinary sense, have been one way for humans of extending their powers and the range and the domains of their application. Another has been to make use of one's peers by collaborating with them, and tapping their knowledge. Inventing the division of labour, man further increased the range of things he could achieve. From a certain perspective, humans also serve as 'tools' to other humans but, obviously, they are very different from ordinary tools and the consequences — good or bad — of 'using' them are also considerably different. We have good reasons not to call them 'tools', of course. We would rather name them our collaborators, colleagues or assistants, at least when viewed from this instrumental perspective. They are themselves complex cognitive systems which (at least some of the time) allows them to produce incredibly intelligent

and autonomous behaviour making them potentially excellent problem solvers in their own right. However, human beings can be moody, emotional and they can have their own goals and motives which may be contradictory to ours. At first glance, such attributes seem to make them less attractive as 'tools', but do they not also add much to their abilities in certain situations? Is it at all possible to have their potential advantages without their potential disadvantages? Where are the benefits of having assistants that are more like us? And can we not also use them for the purpose of learning more about ourselves, about what makes us tick? Can we not also see them as models of us, making our internal cognitive structures and their relationships 'visible' and explainable?

To sum it up, there are two fundamental concerns at the heart of agent research today — that of the study of cognition and that of building practical tools. These concerns in mind, the concept of agency has spurred the imagination of scientists in many different subject areas, most strongly in the fields of cognitive science, artificial intelligence, cognitive psychology and economics. It has gained a particularly firm foothold in cognitive science and artificial intelligence who have arguably contributed and benefited most. But where exactly do they hope to benefit from thinking in terms of an agent metaphor? The following account can only touch on this subject and is by no means complete.

The first strain of research may be broken down into several different areas, like *cognitive modeling*, modeling cognition from a psychologically motivated point of view, *artificial life*, modeling cognition in order to understand more about life, how it originated and how it develops, or *socionics*, modeling social aspects of cognition. Generally, one wishes to know more about the processes that link the input we receive via our sensors (eyes, ears, nose, mouth, tactile sensors etc.) to the output we produce via our effectors (arms, legs, mouth etc.). The term *cognition* is commonly used for most of these processes. To pick out the area most relevant to this thesis, cognitive modeling, it must be said that the most ambitious approaches, namely those that attempt the integration of all or many cognitive capabilities have largely never been used to implement agents. The common term for them is *cognitive architecture* and the collection of assumptions and hypotheses that underly them are often called *Unified Theories of Cognition* (Newell, 1990). Cognitive architectures are intended to generate diverse cognitive abilities from very few atomic structures and processes. The most widely known are *ACT-R* (Anderson and Lebiere, 1998), *Soar* (Newell, 1990), *DUAL* (Kokinov, 1994) and the architectures developed by Sloman in his *Cognition and Affect Project* (cf. Sloman, 2003).

The second strain of research is more application oriented. Developers produce artificial systems that serve some practical purpose as problem solvers. Here we find many subfields of *artificial intelligence* that often study and build virtual (software agents) and physical agents (robots) alike, and employ much the same methods for both. Among them is *distributed artificial intelligence*, a discipline focussing on *multi-agent systems* in distributed problem solving scenarios. Another strand of research lies in the study of *human-computer interaction* or *human-computer communication* paired with methods of ar-

tificial intelligence. It is an attempt to change the way we manipulate the objects we work with in order to reduce the *information overload* and the *function overload* we are increasingly faced with in modern software applications. Therefore, a paradigm shift from direct manipulation to indirect manipulation is being postulated (cp. Pauk, 1997; Lenzmann, 1998). The goal is to develop

> "intelligent background processes that can successfully clone their users' goals and carry them out. Here we want to *indirectly manage agents*, not directly manipulate objects. (...) The creation of autonomous processes that can be successfully communicated with and managed is a qualitative jump from the tool — yet one that must be made." (Kay, 1990, p.203 and p.206)

Finally, there is also much interest in agent-based systems for technical and software-developmental reasons that go beyond the use of the agent metaphor, but into the details of which we cannot go here. They include the distribution of data or control, the integration of legacy systems and open systems (cf. Wooldridge and Ciancarini, 2001).

The areas of practical application for agents are wide-ranging even if we focus our attention solely on software agents (cp. Pauk, 1997; Bradshaw, 1997); refer also to section 3.3. Such agents are used in detecting and correcting errors in the area of system- and network management. They can serve as mobile and autonomous information gatherers or email filters working inside a network. Users communicate with them via PDAs (personal digital assistants) and give them information management tasks requiring some degree of sophistication. In work groups agents can schedule meetings, appointments or other dates on behalf of their users and in interaction with other agents within the work group. Agents can also be of assistance in the handling and the distributed management of commercial processes where they improve the modeling of reality by decentralizing control. Orders are left in the care of agents which then initiate and monitor the execution of commercial processes necessary to complete the order getting into contact with other agents and further delegating tasks. Another important sector is the application of agents in electronic commerce. Intelligent agents can engage in negotiation processes for obtaining certain products, they can help to find service providers, compare prices and eventually perform transactions in the interests of their clients.

All of the applications we just mentioned seem intriguing and challenging and many more can easily be thought of. Yet, to understand how agents manage to achieve what is demanded of them, we need to dig a little deeper than we have done so far. Therefore, let us now look further below the surface at the properties and abilities that define agents.

2.1.3 Properties and Abilities

2.1.3.1 Autonomy — and its Prerequisites In defining the term *autonomy* we face much the same problem that we were faced with finding a definition for the term *agent*.

It is an abstract concept for which our intuition seems clear, yet pinning it down exactly is anything but trivial. A sure sign that it is a hard nut to crack, is that it has attracted generations of philosophers. In this section we shall again start from our intuition and then gradually intersperse the proposals of scientists. So, instead of a definition we will arrive at some sufficient consensus.

Who or what has the property of being autonomous? A table? No! A tree? Well, a little maybe. A horse? Yes, to some extent. A lot more than the tree, for certain. A human? But, of course! What about a state, i.e. a society of humans? Don't we sometimes refer to states as being autonomous? What is it that makes some of them autonomous, some not, and some more so than others?

Looking at the examples above, several things come to mind immediately. First, take the sequence "table,tree,horse", which we can substitute by "object,plant,animal". This gives us the sequence "none,a little,some more" in terms of the degree of autonomy (or just the other way around for *heteronomy*). It seems that the ability to act, i.e. to show behaviour of some sort, is crucial, which in turn is exactly what we associated with the notion of agency. In order to resist the temptation of simply replacing and explaining one abstract term with another, we should clarify a bit what we mean by behaviour[9]. Dretske (1992), a philosopher, explains the behaviour of a system as 'movement' (or an event) caused by some internal causal origin of that system. He also explains that every 'movement' usually has more than one cause. Behaviour is rather something we attribute to a system when we say that *the main cause* of the 'movement' — which is the cause we, for our current purposes, choose to take as the main one — is one that has its origin *within* that system.

A second aspect we will notice is that autonomy also appears to have something to do with being *independent* from something or someone. Pfeifer (2003) differentiates between being dependent on one's environment — food, oxygen, materials — and being dependent on other actors (we may substitute: agents, or systems). In reference to the latter, he states that the more some actor A knows about the *internal* state of an actor B, and the better it knows the rules by which it can influence the internal state of B, the better A can control B. Speaking in Dretske's terms, we might put it this way: The more A becomes the main cause of B's 'movements' and thereby relocates the causal origin to the outside of B, the less we can attribute behaviour to B, or: the more B will depend on A. Pfeifer (2003, p.144) concludes that this shows "that autonomy is not a property of an actor, but a relation between two actors"[10]. Therefore, autonomy is a *relative* term.

[9]The distinction between such terms as *behaviour* and *movement* that we make in this section shall not be made in the following parts of the thesis, unless explicitly stated otherwise. This applies also to the term *action* which is extensively used in Ψ-theory (cf. Schaub, 1993; PSI-Glossar, 2005, for the difference between behaviour (German: Verhalten) and action (German: Handlung)). We would not gain much even if we did make a clear-cut distinction and it will be sufficiently clear what we mean without it.

[10]I shall occasionally cite from german language sources where there is no english translation available. I have translated these citations into English myself.

An agent can be autonomous relative to its environment, relative to other agents in its environment or relative to its designer, for example. In this thesis we will be interested most in the autonomy of assistant agents in relation to their users.

Finally, if we think back to section 2.1.1, the context of autonomy and other necessary properties of agents — especially *reactivity* and *pro-activeness* —, we must state one further point: Autonomy seems to be inconceivable without many of these properties. In other words, autonomy subsumes such properties as *reactivity*, *pro-activeness*, but also others that were mentioned, like *flexibility*, and more loosely *communication* or *cooperation*, which in turn relate to *social ability*. Let us pin down these important terms a little more[11].

Reactivity can be attributed to an agent, if it is capable of maintaining an ongoing interaction with the environment, and responding in a timely fashion to changes that occur in it. This is taken to mean that an agent will need some "hard-wired" behavioural patterns that allow it to act when, for instance, deliberation has failed or just is not feasible.

Pro-activeness is a property that is closely associated with the ability to have and generate goals. In addition, it demands that the agent actively pursues its goals — *goal-oriented*. It must be capable of taking the initiative. This property is very closely related to the general concept of *intentionality* and the *intentional stance* which we will discuss in the following section 2.1.3.2.

Flexibility means that an agent should be able to vary and amend its actions whenever it is forced to — this is not always seen as a necessary condition. Usually, this is due to external influences: For example, the environment has changed such that a chosen goal-oriented course of action is no longer applicable or feasible. In a more general sense, flexibility includes *learning* or *adaptivity*, i.e. the ability to change behaviour based on previous experience.

Cooperation refers to the ability of agents — again, this is not a universally accepted demand — to interact with other self-interested agents. This is a major topic in *Distributed Artificial Intelligence* (DAI) whether with respect to multi-agent systems or societies of agents. In DAI one distinguishes clearly between such concepts as *cooperation*, *coordination*, *negotiation* and *interaction*. For our purposes, however, this is not as important, and therefore, by and large, we shall use them interchangeably. Lastly, cooperation requires agents to have *communication* skills, facilities and languages.

2.1.3.2 Mental Attributes — Intentionality and Utility We saw in the previous section that we associate the concept of *behaviour* with certain systems (or entities), many of

[11] please also refer to (Weiss, 2001, the glossary) and to (Wooldridge, 2000; Franklin and Graesser, 1996)

which we would call *agents*. We also noted that this association, or attribution, largely depends on our point of view, i.e. whether we believe that the main causes of the system's actions lie within it. In deciding on our point of view concerning the causal origin of behaviour, we take a stance. Then, if we want to explain the behaviour or whatever causes the 'movements' of the system, we take yet another stance. In this context, Dennett distinguishes three stances (Wilson and Keil, 2001, see the entry "Intentional Stance"):

physical stance We use this stance when we make use of the physicality of a system and its consequent subjection to the laws of physics. "When I predict that a stone released from my hand will fall to the ground, I am using the physical stance."

design stance We use this stance when we speak of a designed object and its behaviour in terms of the way it is designed, and not primarily in terms of its physical properties and how these explain its behaviour. "Suppose I categorize a novel object as an alarm clock: I can quickly reason that if I depress a few buttons just so, then some hours later the alarm clock will make a loud noise."

intentional stance We use this stance when we explain the behaviour of a system "by treating it as if it were a rational agent that governed its 'choice' of 'action' by a 'consideration' of its 'beliefs' and 'desires'." Evidently, this means that we attribute mental states of some sort to the system, whether these be reduced to beliefs and desires, or further include intentions.

While the behaviour of nearly every system can be explained by any of these stances, we choose the one most appropriate. Here, appropriate would mean, for example, that we usually do not take an intentional stance when we want to explain why our car has broken down: The design stance or even the physical stance will be more appropriate. Also, we would not normally explain the behaviour of human beings from the physical stance but rather from an intentional stance. We do not find it appropriate to say that the car 'wants' or 'desires' not to be restartable, but rather that it just cannot be restarted because it was designed to work using v belts and these happen to be broken. Similarly, it is more appropriate of me to think that my next door neighbour has not turned on the music in his room full blast because he believes that I am away than to say that certain physical forces were acting on his arm and this then caused it to be extended towards the stereo system and so on. The stances are to be taken as strategies for the prediction of behaviour. We take the stance that allows us to make such predictions as correctly and as easily as possible.(refer also to Russell and Norvig, 1995, p.820-822 for a short discussion on the intentional stance and related ideas)

Now, what does *intentionality* mean? And what does the idea of taking one of the three stances mentioned have to do with it — especially the *intentional stance*?

When one is interested in designing artificial agents, one has to decide how to model the control of behaviour, or more concisely, what structures and what processes to use

in order to produce behaviour. Analogous to taking a stance in the explanation and prediction of behaviour, one must opt for a certain level of description or level of design and implementation. Since artificial agents are very often thought of as models of human agents, and since many theories of the human mind, behaviour and problem solving originate from psychology, and since psychology tends to attribute specific mental states to humans explaining behaviour as interactions (or rather transitions) of these states, one common level of description for agent behaviour is that of *mental attributes*.

Perceiving, remembering, believing, desiring, hoping, knowing, intending, feeling, experiencing and the like are all thought of as mental attributes (Jacob, 2003). There are as yet no concise definitions or common conceptions of these terms, and how they relate to one another. Yet, the ongoing dispute about them has provided one useful insight, namely that mental states always seem to *refer to something*, they are *directed towards something*, they are *about something*. For some mental states this property of *aboutness* or *intentionality*, as it is commonly called, yields the idea of *goals* towards which these mental states are 'directed'. The idea is that behaviour is generally directed towards goals. Arguably, the mental state that is associated most with the furtherance of goals is the *intention* — however, one must not confuse the terms *intentionality*, *intension* and *intention* (again, see Jacob, 2003). The assumed relevance of intentions explains why so many theories of behaviour are in fact theories of intention regulation at their core, whether they are called theories of action regulation, practical reasoning or whatever else. This certainly holds true for the two theories discussed in this work.

Speaking of goals, there is one more issue worth mentioning explicitly that is related to *intentionality*. We need to bear in mind that in order to be able to have goals and then act according to them, it is essential to be able to somehow generate options and then choose between them, which, in turn, implies some sort of *payoff* or *utility* — irrespective of the level on which this happens, how it happens and wherein this utility is 'grounded'. In a way, an agent must be able to tell apart 'good' and 'evil', so to speak. In fact, and here we return to the subject of rationality, Russell and Norvig (1995, p.472) state that the principle of maximizing expected utility "could be seen as defining all of AI", ergo it is at the centre of agency. Therefore, this principle seems to be a fundamental hallmark of intelligence and rationality.

Obviously, for an agent to be designed in terms of mental attributes such as those that we so often take for granted in humans may be seen as promising a 'natural access' for a user to the behaviour of an artificial agent assisting him. It might support him in successfully taking an intentional stance towards the agent and more appropriately predict its behaviour. This also holds true for the opposite direction, i.e. for the agent constructing a mental model of the human user. Either way, a consequence of this may be the improvement of certain aspects of the agent's assistance. This leads us directly to our next concern.

2.1.3.3 Assistance — the User Irrespective of how we define the concept *agent*, the vast majority of practical applications will require artificial agents to be useful to someone or something other than themselves. For many, this utility aspect is an integral constituent of agency. So, the question is: What makes an agent useful? And if we remind ourselves of the considerations made at the beginning of section 2.1.2, we can ask the more succinct question: How much will agents have to be tools and how much will they have to be individuals? Do we need axes, 'mindless slaves' or 'mindful selfs'?

Even without going to great lengths, several things become apparent. Although the extreme of complete 'toolness' is rather at odds with the idea of agency, at least in the form of direct manipulation and the spelling out of every detail of what the system should do, such systems may be the most adequate, i.e. useful ones, in certain situations. Or as Laurel (1997) puts it: "Only users who want to use agents should have them; others should have other options". Going to the other end of the spectrum and imagining an artificial agent as a complete individual, like a human person, a colleague at work for instance, we will arrive at demands similar to those formulated by Laurel (1997) (and similarly expressed by (Negroponte, 1997), for instance): "A good agent will do what I want, tell me all I want to know about what it's doing, and give me back the reins when I desire." This short quote contains many basic and often echoed demands for agents which largely result from our own experience with individuals that assist us or those that we assist (see also Bradshaw, 1997; Erickson, 1997; Maes, 1997, for example). The statement entails that the following matters are of importance: a correct communication of desires on the part of the user, correct understanding and execution on the part of the agent — often requiring the agent to build a model of the user and resolving ambiguities — as well as trust and the feeling of control, again on the part of the user. We see that it is paramount that one achieves something like mutual understanding between agent and user: "more intelligence or knowledge is not necessarily better: what is important is the match between the agent's abilities and the user's expectations" (Erickson, 1997). In a simple and more technical manner-of-speaking, the utility function of the agent must match the 'utility function of its user' as precisely as possible (maybe even to the point where the agent's utility function reflects desires the user would not even be able to make explicit?). Beside this, it is the search for an appropriate balance of 'letting the reins loose' and 'regaining control' that is one of the most difficult tasks for the designers of agents.

So, in essence, it is not only necessary for an agent to build a model containing knowledge about the interests, habits and preferences of the user, but also that the user builds an adequate model of his agent allowing him to better predict the agent's behaviour. In the previous section we already learned about the usefulness of successfully taking a certain stance toward a system such as the intentional stance that we commonly take toward our human assistants. This stance makes us give them appropriate tasks and have appropriate expectations about what they will do and when and how they might need our intervention — they are 'cognitively accessible' to us. False expectations are a major source of inefficiencies and frustrations in working with any assistant, espe-

cially computational ones. Therefore, it is a necessity to avoid false promises and over-anthropomorphizations:

> "We want to know that the choices and actions of our agents, whether computational or human, will not be clouded by complex and contradictory psychological variables. In most cases, we want to be able to predict their actions with greater certainty than those of 'real' people" (Laurel, 1997).

Negroponte (1997) nicely captures the desirable aspects of assistance in his metaphors of a digital butler, a digital sister-in-law and so on. In his own words:

> "The idea is to build computer surrogates that possess a body of knowledge both about something (a process, a field of interest, a way of doing) and about you in relation to that something (your taste, your inclinations, your acquaintances). (...) In fact, the concept of an 'agent' embodied in humans helping humans is often one where expertise is indeed mixed with knowledge of you."

This, he says, is why your sister-in-law is so successful in her advice to watch a particular film and a butler in offering you his services at opportune moments, in well trained timing, and respectful of idiosyncrasies.

2.1.4 Architectures

Above all an agent's properties and abilities are the result of the structures that form the agent itself and the processes that work on them, and which ultimately control its behaviour. While its sensory and motor apparati are surely important in this regard, we are, first and foremost, talking about the agent's infrastructure for action regulation, called the *agent architecture*. In one sense it is a methodology for building agents and more generally, it is the particular arrangement of data structures, algorithms, and control flows, which an agent uses in order to decide what to do (Weiss, 2001, cf. the entry in the glossary, p.584). Weiss (2001, p.584) names several example types:

- *logical-based architectures* — in which decision making is achieved via logical induction
- *reactive architectures* — in which decision making is achieved via simple mapping from perception to action
- *belief-desire-intention architectures* — in which decision making is viewed as practical reasoning of the type that we perform every day in the furtherance of our goals, and
- *layered architectures* — in which decision making is realized via the interaction of a number of task accomplishing layers.

In essence, the question of how to design artificial agents amounts to the question of what architecture to choose, i.e. how to construct the inside of the *?*-box in figure 1. Each of the architectures has its own advantages and disadvantages. The choice depends on factors like the desired reaction time, the ease of modification, modularity, and the nature and complexity of the domain. Sometimes, such agent architectures are inspired by certain cognitive architectures, like the ones we already mentioned. It is important to understand, though, that most cognitive architectures have not yet been used as blueprints for the design of agents, but rather as models of human subjects performing very restricted tasks in relatively simple, well-defined and controlled psychological experiments. In light of all the aspects they have in common a strict separation of the two — agent architecture and cognitive architecture — often appears to be somewhat artificial and dated. All the more, if we extend the notion of architecture from the purely cognitive to motivational or even emotional aspects involved in decision making, which we will have to in this thesis, or if we take into account the issue of embodiment, which we cannot deal with here (see the footnote on page 6). Yet, this last point brings us directly to an important issue which must not be neglected entirely: However autonomous or independent an agent may become, it is never completely detached from its surrounding environment nor is it, at least generally, isolated from other agents. An agent's properties and abilities may be limited, amplified and shaped by other agents it interacts with, but they are above all dependent on the nature of the environment and the problems it faces therein.

2.2 Environments, Problems and Resource Bounds

2.2.1 Characteristics of the Agent-Environment Interrelation

It is more worthwhile to provide dimensions along the lines of which an environment may be characterized than to give a definition for the term itself. Since we are interested in environments that are inhabited by agents it is even better to describe the interrelation between agent and environment (throughout this thesis the default usage of the term *environment* shall imply that an environment may include other agents). To this end we resort to the properties as collected by Russell and Norvig (1995, p.46), see below. An agent's environment, whether physical or virtual, may consist of a multitude of components, i.e. objects, attributes, variables and the like, which are related to one another and that, at a particular point in time, form a particular state of affairs. Some of these objects the agent can operate with or modify via operator applications (operations). The agent can initiate certain events that change the state of affairs, or it may simply be affected by them. Thus, the agent receives percepts from the environment and, by acting, may change the environment in perceptible ways. The following list shows several possible constraining factors that are characteristic for the interrelation of agents with their environment:

accessible vs. inaccessible If an agent's sensory apparatus gives it access to the complete state of the environment, then we say that the environment is accessible to that agent. An environment is effectively accessible if the sensors detect all aspects that are relevant to the choice of action. (...)

deterministic vs. nondeterministic If the next state of the environment is completely determined by the current state and the actions selected by the agents, then we say the environment is deterministic. (...) If the environment is inaccessible (...) then it may *appear* to be nondeterministic. This is particularly true if the environment is complex, making it hard to keep track of all the inaccessible aspects. (...)

episodic vs. nonepisodic In an episodic environment, the agent's experience is divided into 'episodes'. Each episode consists of the agent perceiving and then acting. The quality of its actions depends just on the episode itself, because subsequent episodes do not depend on what actions occur in previous episodes. (...)

static vs. dynamic If the environment can change while an agent is deliberating, then we say the environment is dynamic for that agent; otherwise it is static. (...) If the environment does not change with the passage of time but the agent's performance score does, then we say the environment is *semidynamic*.

discrete vs. continuous If there are a limited number of distinct, clearly defined percepts and actions we say that the environment is discrete. Chess is discrete — there are a fixed number of possible moves on each turn. Taxi driving is continuous — the speed and the location of the taxi and the other vehicles sweep through a range of continuous values.

Note: It is often much less important or even irrelevant what the actual state of the environment is, since the state that the agent 'believes' the environment to be in is paramount for its behaviour — ergo, many modern approaches to agency use the term *belief* instead of *knowledge*. In principle, one may distinguish further properties of environments and relations between an agent and its environment. See Dörner (1976) for an extensive list of properties and which also shifts the focus away from the perceptible features of an environment to the effects that operations have on it such as the *range of effect*, the *certainty of effect* or the *reversibility of an operation*.

2.2.2 Characteristics of Problems

Whenever one is dealing with agents as problem solvers one has to have a notion of what constitutes a problem and what it means to solve it. Therefore, in addition to characterizing agents and their environments we must also characterize problems. There are many definitions of the term *problem*. Essentially, they all require there to be some undesirable state of affairs which must or should be transformed into a more desirable state of affairs.

A solution to a given problem is a sequence of operator applications — the means — that allows the transition from a start state to a goal state, possibly via subgoals — the ends. Thinking of problems in this sense allows a straightforward division of the process of problem solving into two phases that are often called *deliberation* — what to achieve — and *means-ends reasoning* — how to achieve it. Deliberation is the process of finding the state of affairs one wishes to achieve from among a number of choices one has. Finding out about possible choices is often seen as part of the deliberation phase. The output of deliberation will be a goal, a list of goals, or simply goal criteria. Once deliberation has provided the adequate goals, the process of means-ends reasoning sets in to find ways of achieving them. One common subprocess of means-ends reasoning is planning, i.e. finding sequences of operations that, when applied, may solve a problem or subproblem. It is usually assumed that we solve many problems by planning or the use of automatisms which are often viewed as plans that have become hard-wired.

In the general case, problem solving is not a straightforward process. It is often hampered for two reasons: The necessary means are unknown or goal criteria are unclear. In this respect a problem differs from a *task*. We speak of a problem only in cases where the transition from an undesired start state to a desired goal state is blocked in some way, i.e. there is a barrier of sorts. These are much the same characteristics one speaks of with respect to the difference between well-defined and ill-defined problems. Figure 2 below shows the four principal types of barriers that arise in the context of problem solving[12].

		Clarity of Goal Criteria	
		High	*Low*
Degree of Knowledge about Means	*High*	Interpolation Barrier	Dialectic Barrier
	Low	Synthesis Barrier	Dialectic & Synthesis Barrier

Figure 2: Classification of Barrier Types in Problems.

In a way problems may be characterized in terms of the different barriers they exhibit. An *interpolation barrier* implies that one has only to find a correct sequence of operations on the basis of known operators and clear goal criteria. This means one needs to plan or schedule the application of means in order to arrive at the goal state. Sometimes, however, one may have a clear understanding of what constitutes a goal state, but some or all of the operators necessary (or supposedly necessary) are not yet part of one's repertoire. In such cases one will need to synthesize operators before putting them into the right order for goal achievement — one faces a *synthesis barrier*. Just as critical are those prob-

[12]The figure is taken from Dörner (1976) the proponent of Ψ-theory. However, the classification proposed is much older than Ψ-theory and is relatively uncontroversial.

lems in which it is unclear what the criteria for goal achievement are, i.e. where you are unsure in advance what constitutes a goal state. Such is the nature of problems exhibiting a *dialectic barrier*. Finally, there are some problems that require a problem solver to deal with the lack or the ignorance of adequate means as well as the lack of clarity concerning goal criteria: This is called a *dialectic and synthesis barrier*.

2.2.3 Resource Bounds

An agent operating as a problem solver in some environment is not only restricted by the primary characteristics of its environment or the characteristics of the problem types it faces. In addition, it is commonly facing the task of having to find solutions within a specified time frame, which means that it must limit its attention to the problem, its deliberation about options, its determination of likely consequences, its relevant calculations, and so on. The time frame can be defined by externally set dead-lines or the limited availability of certain essential, say nutritional or energetic, resources. Timely issues such as these put further pressure onto the reasoning processes and often force the agent to prematurely take decisions as to its actions. This, in turn, may mean suboptimal behaviour, but in real-world applications resource bounds are the rule more often than not. It is why so many researchers in AI speak of *bounded rationality* — a term introduced by Herbert Simon — as the essential issue, rather than rationality in the optimal, objective sense. We should add that the constraints need not only be externally imposed: The limited power of the 'reasoning device' itself, i.e. the limitations of processor speed and memory size for instance, may restrain the search for an optimal solution even more. There is on more vital resource that needs mentioning, and the reader will already have noted this in previous sections: information. Environments that are either principally inaccessible to the agent, that are too complex, or in which the agent is simply not left with enough time to obtain it, information — especially the lack thereof — may become a critical issue. To sum it up with the words of Georgeff (Georgeff et al., 1999):

> "As we [the agent research community] all know, but seem not to have fully understood (...) the world is complex and dynamic, a place where chaos is the norm (...). We also know that computational systems have practical limitations, which limit the information they can access and the computations they can perform. Conventional software systems are designed for static worlds with perfect knowledge — we are instead interested in environments that are dynamic and uncertain (...), and where the computational system only has a local view of the world (i.e., has limited access to information) and is resource bounded (i.e., has finite computational resources). "

"When the intensity of emotional conviction
subsides, a man who is in the habit of reasoning will search
for logical grounds in favour of the belief which he finds in himself."
—Bertrand Russell

3 BDI — Practical Reasoning in Rational Agents
The State of the Art in Theory, Models and Applications

This chapter introduces the basic concepts of the Belief-Desire-Intention (BDI) approach to agency. It is subdivided into five parts: The first part (3.1) sketches the foundations of BDI theory and application, i.e. its philosophical underpinnings, its logical formalizations, and the computational architectures that it has inspired. The theory as well as its realization in the so-called Procedural Reasoning System (PRS) are the subject of the second part (3.2). It deals with how BDI proponents attempt to explain and model practical reasoning in human agents on the basis of the mental attributes of belief, desire and intention, and the processes that operate on them — theory and realization are paralleled for each aspect. The third part informs of some practical applications that were realized with BDI agents (3.3), while the fourth (3.4) briefly outlines central issues currently being worked on and others to be tackled in the future.

3.1 Introduction

BDI, which stands for *Belief Desire Intention*, signifies a triad of concepts first introduced and theorized about in combination by the philosopher Michael Bratman toward the end of the 1980s. His work, and the book "Intention, Plans, and Practical Reason" (Bratman, 1987) in particular, extended the traditional view of practical rationality that was, until then, grounded solely on the concepts of belief/knowledge and desire/goal. They were the dominant concepts of knowledge-based systems and it became clear that such systems were inapt models of rationality, bounded rationality in particular (see Dastani et al., 2003). Bratman, whose focus lies on the concept of intention and its relation to plans, greatly influenced the agent development and research community that was beginning to form at the time. In his eyes only the assumption of a distinctive mental attribute of intention allows a satisfactory explanation of the stability and efficiency found in the control of human reasoning processes. The impact of his ideas, and those of other influential BDI researchers including Wooldridge, Jennings, Cohen, Levesque, Rao and Georgeff to name but the most prominent, has lead to today's dominance of BDI approaches in that part of the agent community considering agent construction from a perspective that focusses on mental attributes[13]. One major facilitator for the interest BDI has received has

[13]For a controversial philosophical discussion of BDI and related issues refer to (Carrier and Machamer, 1997), for instance.

certainly been the fact that it moved the focus of agent development away from finding an optimal decision rule, i.e. a focus on the outcome of reasoning and the theoretically optimal agent as in classical decision theory, to describing rational practical reasoning and its general properties in psychologically plausible cognitive terminology, i.e. a focus on the process of reasoning and agent types (cf. Dastani et al., 2003). While not putting issues of utility and preference aside, BDI approaches concentrate more on the relations between mental attributes and how they might allow for efficient and effective rational reasoning. In some ways, this moves them closer to cognitive architectures.

While Bratman worked on grasping BDI philosophically, others, inspired by his ideas, started formalizing BDI in order to more precisely define (logical) theories of rational agency, and to finally arrive at implementable agent architectures, i.e. computational models of the theory. One approach has emerged from this as a quasi-standard: it is defining a form of *multi-modal logics*, i.e. modal operators for the individual attributes, and postulating specific relational structures to cover the interactions between them. It has also become common practice to express the desired semantics using possible-worlds models and accessibility relations[14]. Most notably, Wooldridge (2000) introduced a logic, called $\mathcal{LORA}$ ($\mathcal{L}$ogic $\mathcal{O}$f $\mathcal{R}$ational $\mathcal{A}$gents), to reason about rational agents in terms of BDI. It is an extension of the original BDI logics developed by Rao and Georgeff (see e.g. Rao and Georgeff, 1995; Georgeff and Rao, 1996). $\mathcal{LORA}$ also allows to capture mutual mental states and the communication and cooperation of many agents. Other formalizations go in the same general direction, such as those of Cohen and Levesque (1990) (see also (Dastani et al., 2003) who discuss several approaches).

The basic framework of BDI logics has not only been extended to the multi-agent case but in many other directions, too. Some formalizations account for additional attributes, like *obligations* for example, in order to better cover social aspects of agency as done in the BOID architecture (cf. Broersen et al., 2001). Others turn to issues of *commitment*, *capabilities*, *know-how* or the like, and consequently, the formal methodology has been broadened to temporal logics (concurrency, timeliness etc.), logics of action (primitive actions, compound actions etc.), and dynamic logics (Padmanabhan Nair, 2003; Singh et al., 2001). The main purposes of rigidly defining formal systems for BDI are: to pave a way for programme specifications for practical application, to make the validation of systems possible, and more generally, to cope with the difficulties of analyzing systems that are embedded in complex environments. The formalizations mentioned all constitute abstract logical approaches defining the semantics of the general notions one deals with when talking about rational agency. Wobcke (2002b), however, criticizes that "there remains a substantial gap between the more abstract logical approaches and the more computationally oriented operational approaches", i.e. "those defining rigorous opera-

[14]For example, given the current situation and an agent's intentions, only certain worlds are what is called *intention-accessible* (the notions for *belief-accessible* and *desire-accessible* are analogous). Loosely speaking, they are those states of affairs considered to be rationally 'intendable' given one's current intentions and given the current state of affairs.

tional semantics for a range of agent programming languages". The gap between theory and practice still needs to be bridged at many points.

The basic ideas of BDI were first operationalized and implemented in the *Intelligent Resource-bounded Machine Architecture* (*IRMA*) by Bratman himself and others. IRMA was meant to be a direct realization of the original theory (Bratman et al., 1988). BDI's most widely known and much adopted implementation, however, is the so-called *Procedural Reasoning System* (*PRS*) developed by Georgeff and his colleagues (Georgeff and Lansky, 1987; PRS Manual, 2001), which constitutes a framework for implementing applications of BDI agents. There have since been a variety of implementations based on the PRS and its descendants (e.g. refer to Wooldridge, 2000, p.8). Even though the PRS was inspired by Bratman's theory, the applications implemented need not strictly conform to or reflect the form of BDI he has in mind, as Pollack points out (Georgeff et al., 1999). As for other implemented BDI frameworks, which mainly differ as regards programming language or programming concept, we must restrict ourselves to naming only a short selection: JACK (2005), Jadex (2005) or UM and JAM (2005). Also, we refer the interested reader to d'Inverno et al. (2004), where a concise collection of references can be found, and to Walczak (2005) or Singh et al. (2001) who give brief descriptions of other BDI related tools and systems.

In this thesis we will be concerned, first and foremost, with the PRS. However, the conclusions we arrive at will for the most part also be applicable to its successors and relatives such as the *distributed Multi-Agent Reasoning System* (*dMARS*) (d'Inverno et al., 1998), JAM or JACK. We shall see how PRS-like systems 'add the necessary flesh to the bones of the theory', and how they then allow developers to incarnate rational agents for specific applications from this solidly founded, yet mouldable matter. The background will be provided to a large extent by Bratman's theory and augmented by some of Wooldridge's considerations concerning necessary and desirable agent properties.

3.2 Theory and Architecture

BDI is a theory of practical reason in humans. Bratman (1987) discusses beliefs, desires, intentions and plans in philosophical terms and employing philosophical methods, based on the concepts of intentionality and practical reasoning. His main concern is that of extending classical belief-desire theories that explain the structures and processes involved in human practical reasoning by claiming that they are solely founded on the mental attributes of belief and desire. Bratman, though not denying the relevance of beliefs and desires, shifts the focus away from them, arguing that they alone are not sufficient, but that practical reasoning and subsequent action, in fact rational agency as such, seems impossible to conceive without the distinctive mental attribute of intention. The appeal of his theory lies chiefly in the combination of two aspects (cf. Wooldridge, 2000; Norling and Sonenberg, 2004): 1) folk psychology is well used to talking about rational agency

and human characteristics in general in terms of believing, desiring and intending, and the theory captures the associated common intuitions in a sensible manner, especially the idea that intentions are distinctive states of mind different from desires and beliefs; 2) the theory is philosophically well founded, clear and plausibly defended; remember also the issue of taking an intentional stance discussed in section 2.1.3.2. In the sections that follow, we will see more clearly how BDI views mental attributes and why intentions are distinguished from beliefs and, more importantly, from desires. Each theoretical aspect will be followed by its realization in the Procedural Reasoning System (PRS) architecture which shall be the focus of our attention.

The PRS constitutes an environment for expressing and executing factual and procedural knowledge used in driving reasoning processes. PRS-like systems have been most influential with those who construct artificial agents for real-world applications, because two very important aspects of the system are its *embeddedness in a changing environment* and that it aims at *real-time reasoning*. To this end the execution system integrates goal-directed and event-driven behaviour, or, in other words, it attempts to perform two things simultaneously: 1) to achieve any goals it has in light of its beliefs, and 2) to react to any new events that occur. There are a number of potential sources for events, each within its own process of execution, and that together make up the environment. First of all, there is the representation of the external world, then there may be other agents, and finally, there is the user. Agents are executed in parallel and may pass messages amongst themselves in an asynchronous manner. The mental attributes postulated by BDI are represented in four separate structures: a database (beliefs), a set of goals (desires), a set of intentions and a plan library (Act library). A central interpreter processes the elements of these structures and organizes all the steps on the way to action. Another vital feature of the PRS is that it allows meta-level reasoning in order to manipulate its own execution. Accordingly, there are meta-level beliefs, meta-level goals, meta-level intentions, and meta-level plans. However, this manipulation can alter behaviour only in general terms, but not change the basic reasoning cycle that the system performs.

The original PRS system was implemented in LISP. Its successor system dMARS is a more robust and faster reimplementation of the PRS in C++. In terms of functionality the dMARS is basically identical to the PRS only its definition is more rigorous formally and more detailed in order to make agents that are built with it to be better evaluatable, validatable and maintainable (cf. d'Inverno et al., 2004; Ingrand et al., 1992). Let us now go through all aspects of BDI — theory and architecture — in turn.

3.2.1 Beliefs, Desires, Intentions

3.2.1.1 Beliefs "Intuitively, beliefs correspond to information the agent has about its environment" (Jennings et al., 1998, p.282). Since this so-called information may not be correct, no longer correct or simply incomplete, due to sensory insufficiencies, sensory

deficiencies, or environmental changes, for instance, the term *belief* is preferred to the traditional AI term *knowledge*. This conception implies that, in principle, it is reasonable to assume that an agent has certain degrees of confidence about issues of its environment. However, Bratman argues that it is generally necessary to assume also that there is such a thing as "flat-out belief" (all-or-nothing), also called *acceptance*. This is because believing something to a certain degree opens up the possibility of inconsistencies in one's reasoning. If, for example, you believe something to some degree, but not flat-out, there is still a chance — and you take it into consideration — that this something is not the case. Now, due to the uncertainty you may still consider courses of action on the basis of this — presumably highly improbable — something, which then interfere with courses of action considered under the belief that the former is in fact true. In most implementations of BDI systems, such complicating issues are avoided right from the start by disallowing degrees of certainty in the set of beliefs.

The Procedural Reasoning System realizes beliefs in a representational format akin to standard first-order predicate calculus with statements of the form:

- `(position valve.1 closed)`
- `(AND (located customer.1 city.2) (located goods.1 city.1))`

The first statement means "*valve no.1* is in the *closed* position", and the second is an expression for the fact that "*customer no.1* is *located* in *city no.2* and the *goods no.1* are *located* in *city no.1*". The *database* holds the current beliefs of the agent about itself and its surroundings. It is constantly updated by the agent's environment that issues its events in the representational format just mentioned. Information can be either dynamic, i.e. it changes over time — like current observations and conclusions —, or it may be static in which case it remains in the database throughout the agent's lifetime — as when representing fixed properties about the application domain, for example. Additionally, the database may hold so-called meta-level facts (meta-facts) that describe the internal state of the system including information referring to current goals and intentions that the system considers for execution.

3.2.1.2 Desires When we think of our desires, mental images come up of situations we would like to be in right now. Some seem unattainable, but others we might reach. Bratman views desires as *pro-attitudes*, i.e. they play a motivational role, "they tend to lead to actions" (Wooldridge, 2000, p.23), and include further *pro-attitudes* like wanting or caring about. "Desires, just like beliefs, admit of degrees. They are the states of affairs that the agent would, in an ideal world, wish to be brought about" (Wooldridge, 2000, p.7), and for this reason they are commonly equated with goals. Wooldridge adds, "[i]mplemented BDI agents require that desires be *consistent* with one another, although human desires often fail in this respect". For instance, when you desire to lose weight and, at the same time, desire not to do any sports and not to cut back on your diet.

In earlier theories of practical reasoning beliefs and desires were the only concepts used to explain intentional behaviour. Bratman challenges this account from several directions in the course of his defence for the distinction of intentions from beliefs and desires, and the importance of their distinctive role both descriptively and functionally. In doing this, he characterizes desires in his own way by downgrading them to "potential influencers of action" with respect to the present and as prone to reconsideration with respect to the future (Bratman, 1987, p.16). Unlike intentions, desires do not play an active part in the control of the agent's reasoning or conduct. We will see the differences more clearly when we turn to intentions.

Since desires, or *goals* as they are called in the PRS, are thought to express those states of the world or within the agent itself that the agent should bring about, their representation is similar to the representation chosen for beliefs: It is based on logical formulae of the same calculus. Such formulae are then combined with special goal-operators picked from a predefined set and applied to the formulae. The resulting statements express that the agent should:

- achieve certain conditions (states of affairs) `(ACHIEVE C)`,
- achieve certain conditions using specified actions `(ACHIEVE-BY C (`$A_1 \ldots A_n$`))`,
- check the environment for whether certain conditions hold `(TEST C)`,
- use certain resources `(USE-RESOURCE R)`,
- wait until certain conditions hold `(WAIT-UNTIL C)`,
- check that some goal remains valid until a condition is satisfied `(REQUIRE-UNTIL G C)`,
- conclude, i.e. add to the database, some specific proposition `(CONCLUDE P)` or
- remove some proposition from the set of beliefs, i.e. from the database `(RETRACT P)`.

Goals express conditions over an interval of time, i.e. over a sequence of world states. Of course, only those goals can be regarded as desires in the relevant sense that are relevant to the current situation of the agent, i.e. those it considers as options for action, amount to desires. How this set of options is generated in the PRS will be explained in section 3.2.2.

Again, the PRS also allows for specifications on a meta-level. Meta-goals express desired internal behaviour for the agent and may, provided they match with the current beliefs that the agent holds, lead to meta-level intentions, and ultimately to changes in the internal workings of the agent. The following quote reveals one fundamental difference between desires and intentions as postulated by BDI proponents, a difference we will elaborate on next:

> "The intuition with BDI systems is that an agent will not, in general, be able to achieve *all* its desires, even if these desires *are* consistent. Agents must therefore fix upon some subset of available desires and commit resources to achieving them. These chosen desires are *intentions* (...)" (d'Inverno et al., 1998, p.2)

3.2.1.3 Intentions At the beginning of this thesis, there was some talk about our intuitions concerning intentions, but how does BDI theory see them? Like desires, intentions are "pro-attitudes", according to Bratman's view, but intentions have a number of distinctive properties in which they differ from desires to the degree that one has to admit them of being distinctive states of mind. On the one hand, intentions can be *present-directed*, i.e. concerning what to do beginning now, in which case we often do not say that we *intend to do* something now, but simply that we *are doing* something now. On the other hand, as is more often the case, intentions can be *future-directed*. Bratman singles out a number of general properties of intentions, which apply to future-directed intentions in particular[15]. Bratman sees intentions as involving a special commitment to action, whereas ordinary desires do not. According to him, commitment of future-directed intentions has two dimensions: a *volitional* and a *reasoning-centred* one. Intentions are:

conduct-controlling (by volitional commitment) In deciding what to do in present circumstances intentions will *control conduct*, and, if future-directed intentions survive until the time of action, and nothing interferes, they will control action then. It is the volitional aspect of commitment that ensures this property.

reasoning-centred (by reasoning-centred commitment) In the period between commitment to some intention and its actual translation into action, the intention will *influence the reasoning processes* that take place. Reasoning will proceed against the background of the intentions already committed to, but not executed yet.

- Intentions, once committed to, are relatively inert or stable. They *resist reconsideration* to some extent, while ordinary desires can be reconsidered and revoked much more easily.
- One will frequently *reason from intended end to intended means or preliminary steps*, and further intentions will arise as a result. They pose problems for further reasoning.
- Often intentions help *limit the options* an agent needs to consider for the future, and consequently they limit further intentions, since the agent assumes that the intentions already committed to will cause certain states of affairs and not others — he calls this property *option admissibility*.

[15]Note: Bratman (1987) discusses all of these properties in great detail and discovers that there are very delicate and complex distinctions and exceptions to be made for each one. For brevity's sake we must gloss over these intricacies.

Bratman presumes that the interplay of the two types of commitment is at the centre of the issue of intentions: "taken together these two dimensions of commitment help explain how intentions play their characteristic role in supporting coordination, both intrapersonal and social" (Bratman, 1987, p.17). Furthermore, intentions are important because they "can reduce the number of alternative options that need be considered by focussing the attention of the system on the possible futures to which it has committed" (Rao and Georgeff, 1993). But how do intentions fulfil the requirements of controlling conduct once active and of influencing the reasoning process in the meantime?

One rather obvious fact about human agents is that we are able to pursue goals that lie ahead in the semi-distant to distant future, and which require us to attain certain specific subgoals on our way to achieving those that are more distant. To do this we are in need of some form of internal coordination that structures our reasoning. Moreover, as was pointed out earlier, practical reasoners are often subjected to limitations, mainly temporal ones (bounded rationality). In order to meet the requirement of 'cognitive economy' that this boundedness entails, humans, or agents constructed as models of humans, must have some way of limiting the amount of time they spend on consideration and computation: "We are not frictionless deliberators" (Bratman, 1987, p.27). Also, we are not "time-slice agents — agents who are constantly starting from scratch in their deliberations" (p.35). What's more, most agents need to coordinate their actions with other agents, i.e. there are additional social limitations. Considering all these factors it appears rational to assume that we plan in some way, or, more precisely, that we have plans. Whether they are stored and prespecified or whether we construct them as we go is secondary at this point[16]. However, Bratman goes one step further: He postulates that plans and intentions are inextricably related and together shape our reasoning. Thus, intentions are viewed as elements in larger plans, they are the building blocks of plans, or, the other way round: "plans (...) are intentions writ large" (Bratman, 1987, p.29). But plans are also the building blocks of intentions! This proposition seems to be contradictory at first glance, but it can be resolved quite easily considering the fact that when someone intends X, he has some plan how to achieve X. He might also intend Y and have some plan of how to achieve Y, and likewise for some Z that he might intend. Having these three intentions X, Y and Z, and their respective plans, the agent thus has some overall plan made up of the individual plans for X, Y, and Z.

Somewhat orthogonal to the aforementioned Bratman distinguishes two notions of plans, of which he only considers the latter one in his further line of argumentation with respect to practical reasoning: The first notion of a plan considers it to be an *abstract structure* like knowing a procedure for achieving some end — a recipe (Wooldridge, 2000, p.28). The second notion, recognizes a plan as a *mental state*, and makes it approach

[16]It has been vital, however, for the success of BDI systems over the past: Most BDI systems make use of a static planning library and resist planning from first principles which makes them more adequate for most environments in which timely responses are of the essence. See Walczak (2005) for the relation of BDI and planning.

the notion of intention, because in this intentional sense, a plan involves an appropriate commitment to action, i.e. it is the knowledge of a procedure for achieving some end *plus* the intention — the characteristic commitment above all — to achieve it. There are two structural properties for plans worth noting: They are *partial* and they are *hierarchical*. The requirement that plans be partial stems from the consideration that it is typically too resource-consuming and even unnecessary to spell out every detail of a complex plan beforehand. In a world in which the future is predictable to no more than a bare minimum, it may even be considered irrational to have a fully specified plan right at the outset of its execution. The details of the plan can be filled in as it unfolds and the agent can, at least principally, react more flexibly to changing situations containing potential dangers and opportunities alike. A hierarchical structure of plans serves a similar purpose. It provides the agent with the possibility of deliberating only on specific parts of a plan while holding other parts fixed.

What does all that mean for the PRS? In a accordance with Bratman's ideas, plans in the PRS come in two different flavours, as *abstract structures* and as *mental states* — the latter are then considered to be intentions with the aforementioned special characteristics of commitment. The actual plans, i.e. the sequences of actions in the pursuit of some goal, are called Acts[17] (or, in earlier versions: KAs, for Knowledge Areas). The dichotomy in Bratman's conception of plans is mirrored by the separation of the set of general Acts, called the *Act-Library*, from the set of tasks or *intentions*, organized on an *intention graph* (see figure 3). Thus, the Acts of the plan library represent plans as abstract structures, while the Acts that form part of intentions may be considered as mental states, or rather as parts of mental states. In fact, an intention is a compound of some initial Act, possibly more sub-Acts of it to follow, and additional information regarding its current status of execution. An intention can be in one of three states: *normal*, *sleeping*, or *awake*. The state of an intention determines whether it may be activated for execution (*normal* or *awake* state) or whether it may not be activated (*sleep* state). Intentions in the *sleep* state wait for some activation condition to be satisfied, and, once this is the case, they switch into the *awake* state. Whilst in the *sleep* state their execution is suspended. The normal and the awake state are the same except for the fact that intentions in an awake state are preferred for activation. Intentions in the *awake* state switch back to the *normal* state in the next cycle of execution.

Let us take a closer look at the structure of an *Act*. An Act consists of a specification of the environmental conditions, so-called *gating conditions*, that have to hold for it to be applicable or which it is to bring about, and a *plot*, i.e. a specification of the steps of a procedure to be executed. The environmental conditions are specified using the goal syntax as described above. They are logically grouped into several categories, or slots, which each may only contain specific types of goal expressions. The slots are: *Cue* —

[17]The word *Act* is sometimes spelled with capital letters, *ACT* (see PRS Manual, 2001), but it does not appear to be an acronym.

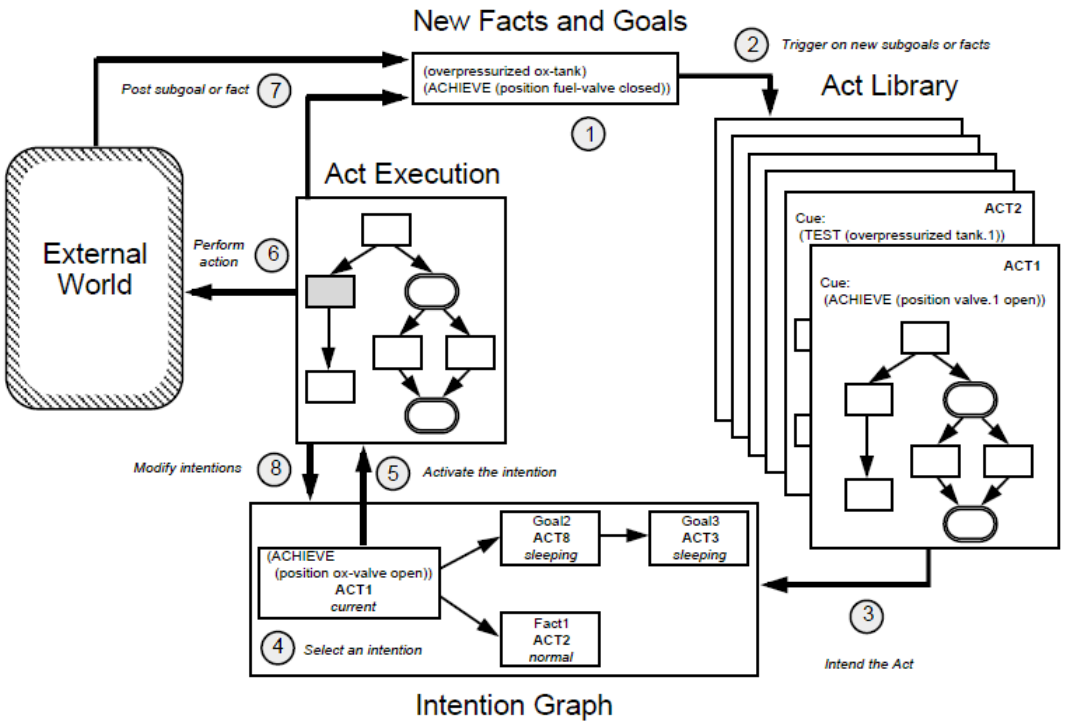


Figure 3: The PRS Interpreter Loop. Taken from Georgeff and Lansky (1987).

the purpose of the Act, *Precondition* and *Setting* — situational constraints, and *Resources* — resources to be allocated for the duration of the Act. There is also a *Properties* slot that can store key/value pairs, but it is negligible for our intents and purposes. Plots are directed, and possibly cyclic, graphs of *primitive actions* and/or *subgoals*. Primitive actions are actions to be performed directly on the external world or the internal world. In this case, the internal world consists of the database and the set of intentions. Primitive actions cannot be interrupted by the system which makes them crucial elements in the definition of the real-time granularity of an agent; they are the behavioural atoms, so to speak. The plot graph may contain conditional and parallel elements. The former necessitating only one element to be executed in order for the plan to proceed, the latter forcing the execution of all elements before the plan may continue.

So much for the structure of intentions, but what about the characteristic properties of intentions demanded by Bratman, which we listed above? In order to appreciate their function, and in order to understand their role in the PRS we must take a look into the process of reasoning which they influence fundamentally.

3.2.2 Practical Reasoning and the Role of Plans

In rational agents the regulation of action commonly takes the form of *practical reasoning*, i.e. (rational) reasoning toward action. Taking a look at how reasoning works in an agent should tell us more about the notion of rationality its designer has in mind. The introductory chapter hinted at the fact that it is common practice to see *plans* as the ultimate products of problem solving and reasoning, provided of course that we are dealing with practical reasoning and not theoretical reasoning. Bratman, as we just saw, follows this line by placing plans at the centre of his theory. In a nutshell, practical reasoning consists of constructing partial and hierarchical plans to reach overall goals that are rational to pursue given the currently held beliefs. Furthermore, it is the filling in of these plans with intentions wherever specific subgoals are to be met. All of this must be done before the background of the intentions already committed to. The selected intentions must then themselves be committed to as is appropriate, i.e. as is rational, and actions must follow accordingly. These processes are influenced by strategies of commitment and additional strategies for the reconsideration of intentions given novel beliefs. So, how is all of this realized in the PRS?

In a way, the core of the Procedural Reasoning System is the *interpreter* whose cycle of execution determines the behaviour of the agent to a substantial degree. This cycle cannot be altered in principal ways by any meta-level processes of the agent. The interpreter's main objective is to balance the response to new goals and the pursuit of current goals. Its general sequence of operation is shown in figure 3 and reflected in the list below, which adds explanations concerning the most critical steps. The steps are as follows:

1. **Establish goals & alter beliefs (the database) based on the events that occur**.

2. **Trigger Acts from the library given the new information & choose *one* applicable Act**. All those Acts applicable given the current elements in the database are collected. Then the interpreter posts meta-level facts to the database which describe this set of applicable Acts. These facts can be used by meta-level Acts, for example, to control system execution. Now, before the interpreter selects an Act to be added to the intention graph, it first needs to re-check applicability in light of the new meta-facts. For this reason, the interpreter invokes the applicability-testing process again using the new meta-facts that have been added to the database. At this stage, only Acts that make use of the newly-posted meta-facts will be considered applicable (provided all of their gating environment conditions are met). This process ends when no more Acts have become applicable. Then *one* Act is chosen at random from the set of applicable Acts.

3. **Place the *one* chosen Act on the intention graph**. The Act chosen in step 2 is added to the intention graph. If the Act is a response to some subgoal of an intention already on the intention graph, then that intention will simply be expanded to reflect

the decision to apply the new Act. Otherwise, it will be the *root intention* of a new tree structure.

4. **Choose *one* intention from the root of the intention graph**. One intention is chosen from the set of *root intentions* that are not in the *sleep* state. If there are intentions in the *awake* state they are preferred, and the one most recently awoken is chosen. Otherwise, a *root intention* will be chosen at random. Beyond this, it is also possible to define a domain-specific selection function that prioritizes certain intentions over others.

5. **Execute *one* step of the intention**. One step of the intention will be executed according to the leaf nodes of the Act that constitutes the intention. It can be one of the actions of step 6 to 8. After execution the intention will be updated depending on the success or the failure of the action taken. This usually means that the successor nodes of the Act must be activated.

6. **Perform a primitive action in the world**.

7. **Establish a new subgoal or the conclusion of some new belief**.

8. **Modify the intention graph**.

9. $\Uparrow$ **restart the loop at step 1 ....**

So, from an abstract perspective, there are three central phases to be distinguished. They are known as the *select phase* (Steps 1 and 2), the *intend phase* (Step 3), and the phase of execution, called the *activate phase* (Steps 4 to 8) (PRS Manual, 2001). The intention graph, like the graph structure employed on the level of Acts, serves the purpose of organizing behaviour[18]. It establishes a partial ordering of the whole intentions: Intentions earlier on a path have to be achieved or dropped before any of those intentions that follow them later on the path may be executed. "[T]he intention graph is best thought of as the actions that an agent currently believes it will execute in the future" (PRS Manual, 2001, p.33). Actually, the intention graph comprises tree structures of intentions, the root element of each is called *root intention*.

The PRS Manual (2001, p.41-42) makes the general functionality of the PRS quite clear by stating that "[t]he PRS interpreter uses a simple, general-purpose strategy for controlling execution of an agent. The generality of the approach provides users with the flexibility to encode complex control strategies for individual domains using meta-acts. It is important to understand that meta-acts don't change the underlying execution mechanisms (...). Rather, they exploit those mechanisms to make the system behave in manner

[18]Notice that the intention graph as a whole holds plans on a larger scale made up of individual intentions while the intentions themselves hold simple Acts, i.e. plans on a smaller scale. It organizes behaviour on a level of granularity coarser than that of the graph structure of individual Acts.

that appears as if some more complex execution strategy were being used. (...) The principal reasons for using meta-acts are to:

- Intend multiple Acts that apply in a single interpreter cycle.
- Make informed choices among multiple applicable Acts.
- Reorder the intention graph in response to new information.
- Kill intentions for activities that are no longer appropriate."

It is at this point, finally, that we see where exactly the actual characteristics of intentions come into play in the PRS. The main part of the control strategies of most BDI agents is dedicated to answering the subtle and often critical question of when and how the agent should reconsider its intentions, or the other way around, how long it should stay committed to its intentions. Wooldridge (2000, p.33-36) mentions the three most discussed commitment strategies in the rational agent literature:

Blind Commitment A blindly committed agent will continue to maintain an intention until it believes the intention has actually been achieved.

Single-minded Commitment A single-minded agent will continue to maintain an intention until it believes that either the intention has been achieved, or else that it is no longer possible to achieve the intention.

Open-minded Commitment An open-minded agent will maintain an intention as long as it is still believed possible.

Depending on the kind of commitment one chooses to employ, one will have types of agents that range from bold to cautious. Using meta-level control mechanisms, such as those possible in the PRS, one can make an agent react to changes in its environment in order to check whether its intentions are still valid or to *exploit serendipity*, for instance. One could also make it ignore possible changes so as to not waste its time reconsidering when a quick response is demanded.

Notice that there has been no real mention of utility or preference functions. Being a very basic reasoning framework, the PRS leaves the responsibility of making informed choices completely with the actual agent developer who may specify appropriate meta-Acts handling these issues. Its successor JAM, for example, at least provides for an optional utility slot for each goal. We indicated in the introduction that BDI's primary focus lies more on making rational decisions through defining appropriate relations between the different mental attributes — they generate options that are consistent with the current beliefs and intentions — than on doing so by means of explicit utility and decision

procedures. Dastani et al. (2003)[19] mention that BDI logics typically reduce utilities to binary values. They also point out (p.17) with reference to Rao and Georgeff (1991) that

> "[c]lassical decision theory and BDI thus seem far apart, and the question can be raised how they can be related. This question has been ignored in the literature, except by Rao and Georgeff's translation of decision trees to beliefs and desires (...). Rao and Georgeff show that constructions like subjective probability and subjective utility can be recreated in the setting of their BDI logic to extend its expressive power and to model the process of deliberation. The result shows that the two approaches are compatible."

However, the focus, as stated, is on modeling certain relations between the three attitudes of belief, desire and intention. (Wooldridge, 2000) discusses many of the possible relations that may be considered. He is mainly concerned with pairwise implications trying to find out which of them seem acceptable or even desirable for rational agents. Each possible relation (i.e. belief $\Longrightarrow$ desire, desire $\Longrightarrow$ belief, belief $\Longrightarrow$ intention etc.) is examined as an attitude towards the present, towards inevitable futures and towards optional futures. So, for instance, Wooldridge asks if it is rational to desire something that you believe is already true (belief $\Longrightarrow$ desire w.r.t. the present), or if one should intend that something will be inevitable given that one believes that it will be inevitable (belief $\Longrightarrow$ intention w.r.t. inevitable futures), or if an agent should intend that something be an option for it in the future when it desires it to be an option (desire $\Longrightarrow$ intention w.r.t. optional futures). It should be clear that such a discussion is ultimately a discussion about types of agents with the differences between types stemming from the different relation to which the individual agents adhere or do not adhere. While some relations seem unacceptable for any rational agent others are acceptable or even desirable in some circumstances but not in others. Concluding, we see that many different agent types can be conceived using BDI even if we let them vary only as regards their commitment strategies and the other relations they might employ between the three attitudes.

3.3 Applications

In the last fifteen to twenty years, since the first BDI implementations were realized, there have been quite a substantial number of applications in a variety of domains (e.g. cf. d'Inverno et al., 2004; Rao and Georgeff, 1995). It is frequently reported that the BDI approach facilitates rapid prototyping of agents and the inclusion of expert knowledge due to the fact that being able to model in terms of beliefs, goals, intentions and plans

[19] They compare the four approaches *classical decision theory*, *qualitative decision theory*, *knowledge-based systems* and BDI systems with respect to the representation of information, motivation and alternatives in decision making.

appeals to human intuition (cp. Braubach et al., 2004; Norling and Sonenberg, 2004; Rao and Georgeff, 1995; McIlroy et al., 1997, and others).

One of the most famous applications of BDI is the fault diagnosis system for the NASA space shuttle Reaction Control System (RCS) (Ingrand et al., 1992). The system was implemented using the PRS, i.e. two instances thereof (an interface and a controller) that communicated with one another and with a simulator. The PRS's were able to reason about the management of control tasks, mainly for the valves of gas tanks (switches, transducers etc.). All in all, about 100 KAs (Acts) were needed and additional 25 KAs for meta-level control. The large majority of them were written directly by the space shuttle mission controllers who had little knowledge of the PRS itself. The database contained more than 650 facts half of which were continuously updated during the simulation.

Ingrand et al. (1992) describe another system IRTNMS, which stands for Interactive Real-Time Telecommunications Network Management System. It is similar to the the space shuttle system. As its name suggests, the system performs Network Traffic Management (NTM) in order to avoid congestion problems on telecommunications networks that occur in cases of natural disasters, special holidays like Christmas, failure of switching systems and the like. Three instances of the PRS were employed (diagnosis, controls, monitor) and a simulator to provide the data on the current status of the network and to execute actions taken by the PRS.

Rao and Georgeff (1995) report on two more applications that go in much the same direction as the one just mentioned. The first is an air traffic management system called OASIS implemented in PRS and dMARS. It consists of one aircraft agent per aircraft, and a number of global agents, including a sequencer, wind modeller, coordinator, and trajectory checker. The system comprises up to seventy or eighty concurrently running agents. The aircraft agents must fly their aircraft while the others do the overall controlling and sequencing such that the aircraft are able to meet their ETAs (Estimated Time of Arrival). The authors also mention the application domain of business process management.

The SWARMM system outlined in (McIlroy et al., 1997) which models air combat tactics can be regarded as a close relative of the aforementioned air control system. It was constructed with dMARS, and employed operational knowledge obtained from airforce pilots. The task was to build agents for flying aircraft (in simulation) by modeling human decision making. The individual agents flying in a team were given individual as well as overall mission goals for various scenarios differing in the type of mission. The coordination of appropriate goal-directed and short-term reactive responses were at the centre of study.

As can be seen from the applications mentioned so far, the faculty to plan (or schedule) and the ability to communicate appear to be crucial for the agents in each domain. This is also the case for the domain of patient scheduling in hospitals as modeled in a system called MedPAge reported on in (Paulussen et al., 2001), for example. At first glance,

the use of the abstract concepts of mental attributes seems to lie solely in the facilitation of obtaining expert knowledge for the procedural information needed. Other applications from the domain of computer entertainment also seem to profit most from this aspect, Norling (2003) explain:

> "The power of a BDI solution lies in the ability to abstract a complex environment in which complex strategies are used. It is when the people being modelled describe their reasoning in folk psychological terms that BDI is useful."

In this context, Norling and Sonenberg (2004) mention a simulated Quake player (Quake is a multi-player shooter game) or agents modeled for the strategy game Black&White (refer also to Norling, 2003).

3.4 Current Advances and Future Issues

Research in BDI models is currently going in many different directions. Theoretical and practical issues often go hand in hand, and consequently, advances are being made on both fronts. In the following, we shall give a brief discussion of some of them.

It was indicated earlier that the granularity of the atomic actions and their run-time performance can be a vital issue. One needs to consider that actions take time, and only some of them may be executed in parallel. Furthermore, it is often crucial for agents to have appropriate situation-dependent strategies for balancing the amount of thinking to the amount of action done in a particular time frame — efficiency and effectivity may depend heavily on the quality of such strategies. Moreover, the process of thinking itself consists of two separate processes, namely deliberation and means-ends reasoning, which again may subdivide into yet more specific processes some of which we saw earlier. Trade-offs become more subtle issues, like: how often to reconsider, how long to reconsider, the level of detail to which plans are to be worked out, and so on.

Thus, one key problem in terms of BDI is finding an optimal intention reconsideration policy (commitment strategy), because, as Schut and Wooldridge (2001) point out, both deliberating over intentions unnecessarily (wasted effort) as well as not deliberating where it would have been fruitful (omitted effort) are undesirable. Now, assigning the right priority to each of these processes can be done in two ways, which we may call *a priori* and *a posteriori*, *innate* and *learned*, or *design-time* and *run-time*. Design-time would mean hard-wiring priorities into the agent even before it gets into contact with its environment for the first time. It entails that the agent would not be able to profit from its experience and its beliefs about the current situation. It also means that the programmer providing the agent with those predefined strategies would have to know quite a lot about all the possible situations that the agent might face. Design-time strategies are

the norm in the BDI solutions currently available, if indeed explicit strategies are used at all. However, there are approaches to adapting strategies at run-time such as the one presented in (Schut and Wooldridge, 2001).

Another concern, related to what has just been said and frequently put forward against BDI approaches is that of relying completely on prespecified plans. Very often BDI solutions do not make use of planning algorithms but resort to a static plan library in order to meet the demands of real-time reasoning where ad hoc planning is too time consuming. However, this is not so much a principal weakness of BDI, but rather a technical constraint (e.g. refer to Walczak, 2005, for discussion on planning in BDI agents). In relation to the PRS/dMARS architecture d'Inverno et al. (2004, p.10) defend the use of predefined plans by arguing that

> "[i]n PRS, planning ahead is possible through meta-plans (...). However, in practice, users rarely use meta-plans because this knowledge can always be embedded in the object plan — it just requires the inclusion of an extra step in the plan to call the relevant planning engine (which could also include the set of plans used by dMARS).(...) The reason dMARS does not do this automatically is the assumption that in most cases the environment is so uncertain that planning ahead will not help — it is better to try something and if that does not work, try something else."

Furthermore, BDI approaches are often criticized for having difficulties incorporating methods of learning and social interaction. Pollack (Georgeff et al., 1999) — with reference to the IRMA architecture mentioned earlier — has responded to this matter as follows:

> "[R]ational agents will tend to focus their practical reasoning on the intentions they have already adopted, and will tend to bypass full consideration of options that conflict with those intentions. Let's call this Bratman's Claim (...) Certainly, if Bratman's Claim is a viable one, then it must be possible to design IRMA agents who can learn and can interact with one another. However, all that is required is that Bratman's Claim be compatible with (some) theories of learning and social interaction: Bratman's Claim itself does not have to tell us anything about these capabilities. To date, I see no evidence that there is anything in either Bratman's Claim or its interpretation in IRMA models that would make an IRMA agent inherently poorly suited to learning or social interaction."

It appears that the criticism with respect to learning essentially boils down to the standard criticism of symbolic approaches, i.e. approaches relying on symbolic forms of knowledge representation (or belief representation for that matter), and the difficulties involved

in learning symbols by rote learning, learning by analogy, learning from examples and the like. Consequently, it is a criticism of the way BDI implementations represent belief rather than a criticism of Bratman's theory.

"Quand tu veux construire un bateau, ne commence pas par rassembler du bois, couper des planches et distribuer du travail, mais réveille au sein des hommes le désir de la mer grande et large."
—Antoine de Saint-Exupéry

4 Ψ — Practical Reasoning in Motivated and Emotional Agents
The State of the Art in Theory, Models and Applications

This chapter introduces the basic concepts of Ψ-theory and its approach to agency. Just like the previous chapter it is subdivided into five parts: The first part (4.1) informs about the roots and the general nature and purpose of the theory. The second part (4.2) then presents the concepts of the theory and shows how they are realized directly in an agent architecture. It is explained how the structures like schema, urge, motive and intention work together to produce goal-directed behaviour which is modulated in order to account for the flexibility of practical reasoning. It is also shown how these modulations can be interpreted as emotions. The third part informs of the applications for which Ψ-agents have been implemented (4.3), while the fourth and final part (4.4) points to the way ahead for the theory and the issues that are being pursued to increase its range and its validity.

4.1 Introduction

Over the last decades, the psychologist Dietrich Dörner, developed a fundamental, detailed and holistic theory of mental processes, called Ψ-*theory*. It is fundamental in that it provides an authentic, though variously inspired, theoretical framework specifying the mental structures and processes of human action regulation. It is detailed, because its profound specifications range from elementary processing units — individual (idealized) neurons — to the explanation of what constitutes mental attributes such as intentions, and how the latter derive from the former. Finally, it is holistic, meaning to say that it incorporates cognitive, motivational, volitional and emotional aspects of the mind, and laying its foundations on the interleaving of them to explain the shaping of action. What's more, it provides basic explanations of the phenomena of language and consciousness, if only rather speculatively at present. How is Dörner's approach motivated and what are the theoretical and methodological consequences?

Ψ-theory is founded, amongst other things, on the position that the traditional, reductionistic means of psychological analysis alone have failed to do their job of drawing the bigger picture concerning the workings of the human mind. According to this view, traditional psychology has reduced its level of analysis to explaining ever smaller structures in ever more detail while at the same time neglecting their interactions, and the large-scale processes operating on all of them. This criticism is not entirely new, and is

attributed most to Allen Newell whose "dream" of "Unified Theories of Cognition" (cf. "Newell's Dream" Anderson and Lebiere, 1998) sparked the development of cognitive architectures (recall section 2.1.2).[20] Yet, even if Dörner agrees on there being a need for unified theories in psychology, he does not believe cognition on its own to suffice. Why is this?

Dörner has long been interested in how humans act in complex situations. Consequently, he and his group have undertaken a number of experiments in which such complex situations are simulated on the computer (Dörner et al., 2002). Subjects must try to act successfully in the simulated worlds: Success, however, depends on a number of variables, and it requires subjects to have adequate mental models of how the variables interact. The results obtained in conjunction with the observations made suggested to Dörner that standard psychological models were not sufficient to account for the behaviour of the subjects, and that a holistic approach seemed more appropriate. Hence, the construction of a theory was begun — adding motivational and emotional processes —, a theory that could account for the action regulating processes taking place within subjects, and accurately predict the actions they would take. As a consequence of their observation that isolated accounts were too limited, Dörner and his colleagues motivate an approach more akin to *systems theory*. Accordingly, the human mind is understood as a system not necessarily complex in its underlying structure, but rather as producing complex patterns of behaviour from recursively relating its parts and nesting them in one another: "The whole is more than the sum of its parts" (Detje, 1999, pp.10-14) or (see Dörner's "Geleitwort" in Schaub, 1993).[21]

Since, the complex nature of such a theory does not allow for the application of traditional psychological methods, another approach was taken for the testing of its hypotheses and predictions: The mental apparatus of an experimental subject is modeled computationally. Dörner calls this approach to psychology *synthetic psychology*, i.e. theory and computational model go hand in hand. Thus, the theory does not only claim to be an explanatory psychological theory, but a generative one at the same time. Therefore, it provides relatively detailed instructions on how to computationally synthesize the psychic phenomena it wants to explain. So, by virtue of the properties aforementioned, Ψ may be seen as a *cognitive architecture* of sorts, and an attempt to realize "Newell's dream", a unified theory of cognition. However, in contrast to the famous architectures developed by Newell and Anderson, the pioneers of the field, Ψ is a cognitive architecture, but it is

[20] See also the related criticism of experimental psychology by Toda (1982, p.94) who argues that traditional laboratory settings are "sacrificing important information coming from the multiplicity of our input channels and the multiplicity of our thinking and other activities..." because they are not "our natural habitat".

[21] Note: The brain scientist and cyberneticist Valentin Braitenberg points out that we (and especially those with an analytic mind) often tend to overestimate the complexity of a situation or a behaviour. He demonstrates this by means of his so-called *Braitenberg vehicles* which, though very simple constructions, exhibit extraordinarily complex patterns of behaviour (Braitenberg, 1993). The problem is closely related to the "Frame of Reference Problem" and, as the reader will have noted, to the attributions we make when we take a stance toward a system (refer to 2.1.3.2)

a holistic one at that — maybe "architecture of mind" is a more appropriate term, for it goes beyond cognition (note the related comments in Dörner et al., 2002, pp.12-16).

There is a third fundamental foothold of the theory: It is firmly rooted in the scientific area of *cybernetics* from which it takes one of its main inspirations. In short, Dörner's claim is that all mental processes, including cognition as much as emotion, can be derived from the features of a regulatory system, which evolved solely in order to adapt the organism to specific, mostly environmental, conditions. What's more, sooner or later, all known mental processes inevitably emerge from such a regulatory system as long as it is constructed to fulfill certain requirements of autonomy (Dörner, 1998, p.21).

Finally, an account of the theory's development would not be complete without mentioning the influence of Masanao Toda. Neither the theory's roots in cybernetics nor those of its synthetic methodology can be fully appreciated without knowledge of his so-called *fungus eater* experimental scenario. Toda developed a simulated game in which subjects were to direct a robot miner, the *fungus eater*, sent to some planet, and collect as much uranium ore as possible while having to monitor the robot's nutrition level — it needed to eat fungi every once in a while in order to stay alive (for details refer to Toda, 1982; Rübenstrunk, 1998). Dörner developed a similar scenario to test his theory by placing a steam engine robot, called Ψ, on a little island on which it has to collect so-called nucleotides whilst also satisfying its internal needs so it does not die. This scenario will be explained in more detail in section 4.3.

The construction plans for Dörner's artificial organism Ψ can be found in Dörner (1998), a book intended to be generally intelligible and directed at a wider public audience, or, more recently, in Dörner et al. (2002), a much more detailed work that, above all, provides exact formalizations of the fundamental concepts.

4.2 Theory and Architecture

The essence of Ψ-theory is that it emphasizes the integration of perception, emotion, cognition, motivation and action for human action regulation. Instead of focussing on single modules, the emphasis is on the interaction of the different components. The general theoretical background was introduced above. In the following, the individual components and their many interactions of this complex theory and its model will be explained in as much depth as is essential for their appreciation.

4.2.1 Schemata, Urges, Motives and Intentions

4.2.1.1 Schemata In Ψ information about the world is encoded in networks of artificial, mathematically idealized, neurons that, in principle, are no different from those common to connectionist models of subsymbolic representation. Both share many notions

and functionalities such as connection weights, activation levels, thresholds, spreading activation and so on — we cannot go into any details here. Every neuron can have several connections to other neurons with each connection being of one of the following four types: excitatory, inhibitory — they activate and deactivate the postsynaptic cell respectively —, associative or dissociative — they strengthen or weaken other active connections to the postsynaptic cell.

However, there are also significant differences to standard conceptions of connectionism. First and foremost, Dörner proposes a special structure for Ψ's neural networks: The elementary units of interest are not so much the neurons themselves — although they underlie all representation — but the so-called *quads* which they form. Superficially, a quad is nothing more than a neuron with an axon, i.e. an outgoing connection, but in fact the axon fans out to four neighbour neurons which then establish the actual connection to the next quads. The purpose of this structure is to realize a *hierarchical and temporal organization* of representation[22]: one of these neurons establishes connections to those neighbouring quads that represent states of affairs which supposedly (temporally) follow the state of affairs it represents itself (say, a chair that stands beside the table), another connects to those quads that represent states of affairs supposedly having preceded it (say, a person standing beside the table), yet another connects to those representations that this quad may be a part of (a room, for instance), while the fourth neuron connects this quad to those other quads whose representations it may itself contain (a leg, for example). The 'maybes' and 'supposedlys' in the last sentence are to indicate that, according to the experience of Ψ, a certain representation may form part of, or have as its parts, many different situations and/or objects.

Networks of neurons, or rather networks of quads, represent the information gathered by the Ψ-agent from its environment, i.e. memory traces, of objects, scenes, motor sequences, and more complex combinations thereof. Perception is guided by a process called HyPercept (Hypothesis-driven Perception) which continually weaves a thread of protocol neurons constituting a time line. The head of the thread represents the current situation, while sections further away form the head represent situations encountered some time in the past. Protocol neurons point to the networks of neurons representing what is encountered (or more correctly, that which the agent believes to have encountered). Essentially, there are two forms of such networks: *sensory schemata* and *motor schemata*. Both of them consist of sequences of quads in which each quad may not only point to its successor within the sequence, but also herald the beginning of further subschemata that represent at a greater level of detail as stated above. *Sensory schemata* encode sensory data on different levels of detail with connections reflecting the time and the direction to get from one element to the next: for example, where and how fast to

[22]In the context of connectionism, one must be very careful when stating that something is represented by some neuron, because representation is usually spread across a very large number of cells. Here, saying that a neuron represents something (say, a table) is taken to mean that it forms a kind of 'gateway' to the representation of that something, i.e. the assembly of cells it connects to actually forms this representation.

move the agent's visual sensors for the next expected element of, say a face, to make its mark (an image) on the retina. So, each connection contains spatio-temporal coordinates, but it also stores the strength of the connection between the two, i.e. an estimate of how much this connection has been used — an indicator for its relevance — and from this (plus the total number of outgoing connections from the first neuron) the epistemic likelihood of the successor neuron's representation being the outcome of the predecessor may be derived. Similar to sensory schemata, *motor schemata* encode simple or complex sequences of likewise primitive or complex actions performed by the agent's effectors. In fact, a basic *motor schema* is made up of three elements: 1) a representation of the action's precondition — a sensory schema describing what situation has to be the case in order for the action to be applicable —, 2) a representation of a primitive or complex action (again, a motor schema) to take, and 3) a representation of the action's postcondition — a sensory schema describing what situation is expected to be the case after the action has been performed. Only if the postcondition holds after execution of the action can the motor schema be said to have been successful. Schemata consisting only of sequences of sensory schemata are also called *event schemata*. They represent affairs which happen without the agent having any influence on them or without it needing to influence them.

Many standard processes known from other connectionist models also apply to the neural nets in Ψ. Perceptions from the environment, processed by HyPercept, and the memory recall which is part of internal processes both result in waves of spreading activation — each wave is guided along one of the four directions mentioned above. This, in turn, leads to the strengthening of the activated connections, i.e. learning takes place. The complementary process of forgetting over time leads to the general decay of connections whereby only those neurons remain accessible whose connections were previously strengthened by recall. Thus, isolation of memory traces leads to abstractions in the representations, the emphasis of important information and the disappearance of that which is only very seldom accessed.

Before continuing, we should remember that the whole mental machinery of Ψ is realized as networks of neurons and quads — schemata (knowledge/beliefs), urges, goals, motives and intentions. We will not explicate this fact again, but it is worthwhile to bear in mind. Also, it must be noted that every *situational image* (German: Situationsbild[23]), the situation as perceived by Ψ, and every internal process that takes place is translated into schemata and attached to a potentially infinite, yet slowly decaying, protocol thread (*protocol memory*).

4.2.1.2 Urges It was mentioned at the outset of this chapter that Dörner's theory is profoundly influenced by ideas from the field of cybernetics. As a major consequence of this, Dörner believes so-called *urges* (German: Bedürfnisse) to be absolutely crucial for

[23] Bach (2003) translates it as *local perceptual space*.

the initiation and regulation of action[24]. Why is this and what are they?

Now, cybernetics is the study of regulatory processes, processes that work to maintain the equilibrium of a system. Equilibrium refers to a state in which a certain vital quantity of the system, we might call it a variable, is balanced out. Equilibrium is often necessary to ensure the systems survival, whether directly or indirectly. Frequently, the system will find itself in a state of disequilibrium when the level of at least one variable deviates from its optimum. Therefore, due to its profound homeostatic tendency, there are processes within Ψ that are designed to be initiated in order to try and reestablish equilibrium. If the agent's *need* (German: Bedarf) — the psychological term for the deviation — is relatively small, it may be satisfied by processes internal to the agent (physiological regulation). At some point, however, the need may have become too great for such internal automation to be sufficient. It is then that mechanisms must be started which relocate the search for a solution to outside the agent. More precisely, the agent must induce behaviour directed at its environment. Such mechanisms are usually intentional and goal-directed (psychological regulation). They are notified and initiated by a *signal for a need* which is called an *urge* (German: Bedürfnis). In fact, both the variable to be balanced and the signal that this variable is currently in such a state of disequilibrium that it necessitates psychological regulation are called *urge*. In order to distinguish the one from the other, we will refer to the former as a *fundamental urge* and to the latter as an *active urge*, *momentary urge*, or simply as an *urge*. Things will become clearer with the following example for which we must go ahead of ourselves slightly (see section 4.3) and assume a particular system, an agent, in order to discuss its urges. Imagine a robot Ψ driven by a steam engine with wheels and some sensors attached to it. An urge might thus be imagined as the measure to which the amount of some liquid in some tank differs from its optimum. The different tanks represent different fundamental urges.

Dörner (1998) identifies five fundamental urges for the steam engine robot (see the five tanks at the bottom of figure 4; in fact we may also include a sixth that he names: *damage avoidance* or *intactness*). They can be categorized into two basic groups: material (the need for water and the need for energy) and informational (the needs for affiliation, certainty and competence):

Water It is a prime necessity for any steam engine — no water, no steam. Steam production constantly reduces the level in the water tank, while the intake of water from the environment replenishes it.

Energy It needs to be available to steam production in order for the fire to keep burning below the water tank. Again, steam production consumes energy; collecting wood, coal, or in Ψ's case, nuts and seeds, refills the energy tank.

[24]Bach sometimes refers to urges as *innate desires* or *somatic desires*. Toda also uses this term though with a slightly different meaning.

Affiliation This term describes the fundamental urge for being socially accepted, i.e. for conforming to the norms of one's group or groups, and having the impression of being a legitimate member. Information signaling acceptance — called L-signal, for legitimacy — raises the level of the substance in the affiliation tank (i.e. reduces the need). On the other hand, the lack thereof or indeed the perception of negative feedback from other individuals of one's chosen social environment will lead to the level dropping.

Certainty Being able to reliably predict the outcome of one's actions or the course of events taking place in the environment irrespective of one's actions makes a creature be more certain. The less predictable its environment is — or appears to be to Ψ — the lower the level of certainty will fall.

Competence Every time Ψ effectively performs some chosen action, its level of competence will rise. In case of the action satisfying some need, the amount of increase in competence will be even greater. The opposite is true for inefficient or ineffective behaviour.

Attached to every tank we find a sensor which measures the need. In turn, we find another sensor attached to the first one, which adds up the single divergences over time — it is thus an accumulator. The motivator once sufficiently active signals the need: an urge has arisen. Together, the two sensors make up what is called a *motivator*. It should be obvious that different organisms may have different urges. But what to do in case of an urge arising? Act! Yes, but how exactly? This is where motives and intentions come into play.

4.2.1.3 Motives and Intentions The main subtheory of the Ψ-theory of action regulation is its theory of intention regulation, and consequently, any Ψ-agent is fundamentally a system that regulates its intentions. The two notions central to this theory are *intention* and *motive*. Unfortunately, the relation and the differences between the two terms, often synonymously used in the literature (see e.g. Dörner, 1998; Dörner et al., 2002, p.212, p.444; and p.181 respectively), do not always become apparent but may sometimes be a cause for confusion. In this section we will attempt to resolve this confusion as far as possible. Perhaps the best explanation to be found is the following (Dörner et al., 2002, p.179):

> "The core of an intention is an active urge; usually, this active urge is connected to one or more goals. (...) Urge and goal form the motive core of an intention. (...) Now, goals are usually not isolated in Ψ's memory, but connected to other contents of memory in many ways. (...) Thus, an intention is not only the respective motivational core (urge and goal), but a network

> of related associations which, one might say, form the *working memory* with respect to the current intention."

So, first of all, a *motive* is the conscious or subconscious perception of an urge. Ψ being a *multi-motivated* system, may be faced with quite a number of motives active at once at any point in time depending on the number of urges it has at that time. We already know the core element of any motive, namely a motivator. A motive, however, is more than just the signal of a need. It is a widely held position in motivational psychology that motives arise from urges and necessarily include goals. This makes Dörner state the simple equation (Dörner, 1998, p.307):

$$Motive = Urge + Goal. \tag{1}$$

Without some kind of urge we, or any other living organism for that matter, would probably not feel motivated to act at all. Nor do we usually initiate action without having a certain, specific or unspecific, *goal* in mind (with reflexes being an obvious exception).

What are *goals* according to Ψ-theory? They are those memory schemata that represent past situations in which one or several urges were reduced or even satisfied, or by which they arose or increased. In the former case they are called *appetitive* — they represent desirable situations —, in the latter case they are named *aversive* — they represent situations to be avoided. The association of urge and goal takes place at the motivator: Every time an urge arises, the representation of the current situation is connected to the motivator by the so-called pleasure-pain system which might simply be an associating neuron. Goals can be viewed either as the satisfactory (or 'painful') situation itself or as a gateway to such a situation. Dörner prefers to call such situations, situations in which a *consumatory terminal action* (German: konsummatorische Endhandlung) can be performed, i.e. a material or informational resource can be consumed. Due to the organization of memory, goals are usually linked to the whole sequence or sequences of situations and actions that have lead to them in the past. Sensory schemata within those sequences can function as subgoals, in the process of planning, for instance. Sometimes a goal will satisfy more than one urge, cause more than one urge, or at the same time satisfy and cause different ones: they form *motive amalgamations*.

Intentions, as the above quote suggests, are motives "at heart" and more. Thus, they are not separate structures but the relation is one of inclusion. In earlier work, Schaub (1993, p.63 and the following) defines intention as "the structure that contains a motive plus knowledge for the satisfaction of the motive". This idea is in line with the one uttered in the quote above. Thus, in a nutshell, intentions can be seen as 'enhanced motives', enhanced by some or all of the following informational components[25]:

[25] cf. also the figure on p. 444 in (Dörner, 1998) which depicts an overview of the structure of a motive/intention.

- a *plan* — a sequence of schemata leading from the situation at the intention's creation to the goal state
- the *current state* of execution of the intention, i.e. its position in the plan
- a *history* — the section of the plan already performed on the way to the current state
- a *deadline*, i.e. an estimate of how much time remains for attaining the goal
- the intention's *importance* — functionally related to the strength of the underlying urges
- the estimated amount of *time needed* to perform the plan
- the estimated *competence* for actually achieving the plan

Of course, all of these components are either directly contained within the memory structures or derivable from them. The time estimate for performing an action sequence (or for a certain sequence of events to occur) can be estimated by the temporal coordinates along the sequence of neurons that represent this sequence of actions. Likewise, the competence measure will depend on the epistemic likelihoods stored in the connections along that sequence. Essentially, all these components form part of Ψ's *working memory* which has been mentioned in the original quote and which we will inspect more closely in the next section.

Concluding at this point, we can state the difference between the terms *motive* and *intention* as follows:

1. an intention is an enhanced motive,
2. a motive *motivates action* as a response to arising urges and, as an intention, it is the *inner accompaniment of that action* (German: innere Begleitung) — history, deadline, time needed and so on, and
3. a motive *directs action towards goals* and, as an intention, it *organizes action towards particular goals* — plan.

In accordance with the seemingly customary usage of the two terms by Dörner, we will use the word *motive* wherever the motivational aspect is brought to the fore, but by default we will use the word *intention*, in particular wherever the planning and action aspect is at the centre of discussion as in "current intention" to denote the intention/motive currently in control of conduct.

One last remark: The term *intention* in Ψ-theory differs from its everyday usage in that the question of the consciousness of intentions is factored out completely. Intentions can generally become conscious or remain unconscious (PSI-Glossar, 2005, see *Absicht*). We

may learn more about the exact nature of intentions when we ask how they are created, how they are maintained and how they become conduct-controlling. Let us turn to these matters now.

4.2.2 Intention Regulation and Emotional Modulation

The process of producing adequate behavioural responses to a given situation, we called this action regulation, has two aspects for a Ψ-agent. First of all, there is the creation of intentions, their selection and their subsequent execution. The second aspect are the various ways of modulating the process by affecting the manner in which it occurs, as opposed to the causal nature of the process itself. We may call the former process *intention regulation* and the latter, for reasons that will become apparent, *emotional modulation*. Both processes are based on the same motivational foundations and intricately connect to define the purpose of behaviour, the structure of behaviour and the manner in which it takes place. Figure 4 shows the relationships of the various parts of Ψ with memory at the top, the motivational system at the bottom, and perception and action indicated to the left and right respectively. Also, the main emotional modulations are depicted as red arrows. In the centre we find the components of intention regulation from bottom to top: motive generation, intention selection (motive selection), and action selection. Figure 5 sheds some more light on the latter.

4.2.2.1 Motive Generation, Intention Selection and Action Selection Any action undertaken by a Ψ-agent is rooted in and fundamentally caused by some state of disequilibrium as stated earlier. Should low-level processes not be able to restore the unbalanced urge to anywhere near its optimum, or should the need in fact increase, homeostasis must be achieved on another level. An urge marks the creation of a motive since its respective motivator is connected to those parts of the agent's memory that store information on past experiences that are related to this urge arising or subsiding. Even if it has no prior experience in dealing with this urge, it is now motivated to act: It has a motive.

The generation of motives is followed by the core of intention regulation which is driven by the components of what Dörner (1998, pp.515-533) refers to as *working memory*. The term working memory signifies the combination of four structures: *protocol memory*, *situational image*, *expectation horizon*, and *intention memory*. It is their interplay that determines the agent's intended course of action towards equilibrium. Protocol memory we have already encountered at the end of section 4.2.1.1 and its function should be apparent. Also, there was mention of the situational image (German: Situationsbild): It is the volatile product of perception holding representations of the objects of the current situation in which the agent finds itself. They are represented as schemata to the level of detail and precision at which the agent perceives. The *expectation horizon* (German: Erwartungshorizont) is a projection from the current situation into the future based on the

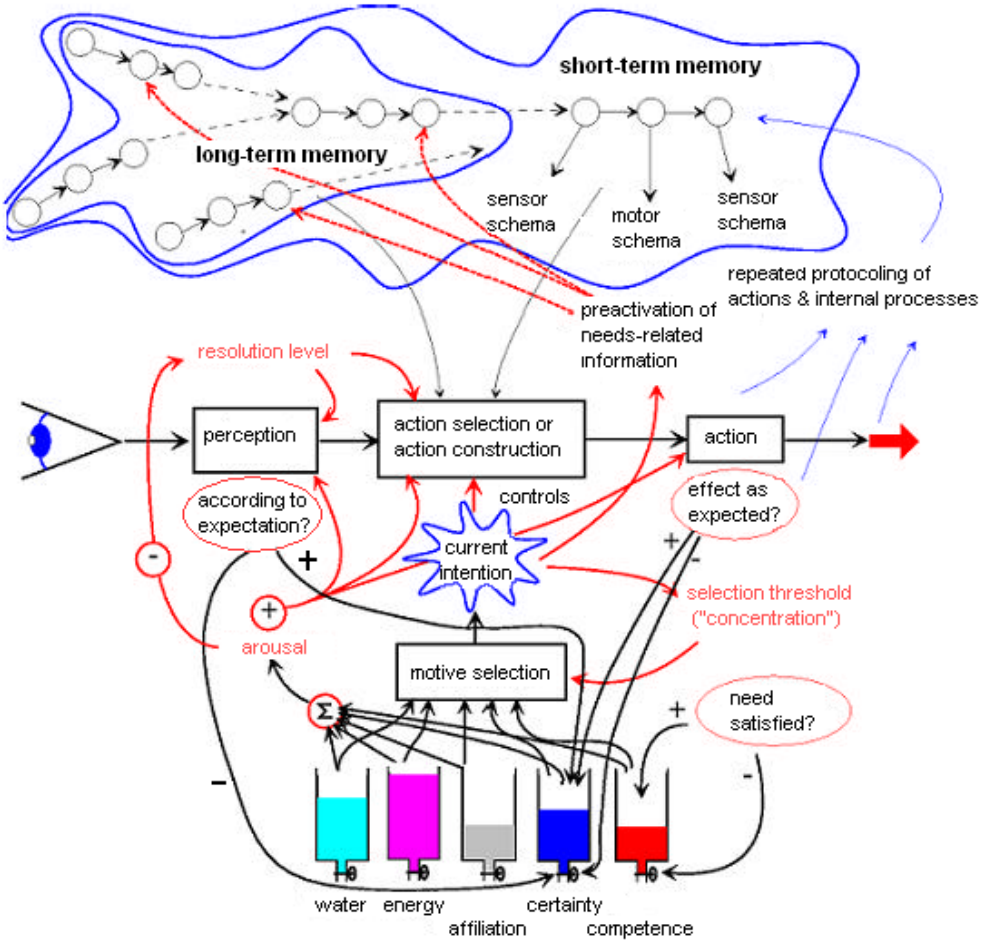


Figure 4: Ψs Internal Mechanisms

situational image and on what has been protocoled, i.e. experienced in the past. Identifying certain crucial situations in this projection allows the agent to form intentions to avoid them or bring them about. Thus, the anticipation of an urge, which amounts to a danger, or the satisfaction of an urge, which amounts to an opportunity, may also motivate the agent towards action. Either way, the agent may be faced with a whole array of motives waiting to be dealt with. Together, all these motives form *intention memory* which is not really a separate structure — neither is working memory as a whole — but just a convenient designation for the compound of active motivators, their associated goals and other schemata, their competence measures and their motivational strengths. Competence measures? Motivational strengths?

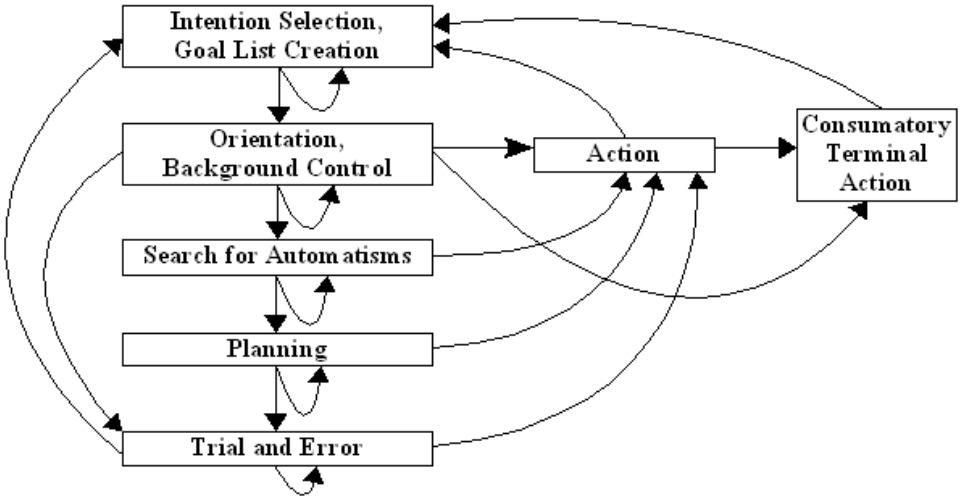


Figure 5: Ψs Phases and Transitions of Action Selection

Obviously, not all motives can be dealt with in parallel: they might be incompatible or require the same sensors and/or effectors, and therefore they have to be sequenced. Ψ-theory fashions this sequencing process as a competition for becoming conduct-controlling. At any point in time there is exactly one conduct-controlling intention. Determining which motive is to become the conduct-controlling intention is a continually active process: The strength of each motive, also called its motivational pressure, is computed as the product of an *expectation* measure — the estimated probability of success for the motive as a function of experience — and a *value* measure — essentially, the strength of the underlying urge(s) which is also the expected amount of satisfaction/pain when attaining the goal (*expectation*×*value-principle*). The former is an estimate of the agent's *general, heuristic competence* (the competence urge) and its *specific, epistemic competence* in dealing with the motive. The epistemic competence is obtained from the strength of connections within the goal-directed schemata for the motive from the current situation to the goal situations. Finding the motive with the greatest motivational pressure is performed by the neuronal process of lateral inhibition in which the currently conduct-controlling intention participates with a small bonus. This bonus, called *selection threshold*, is introduced in order to avoid rapid changes (oscillation) in the control of conduct which might otherwise make goal-directed behaviour impossible or at least very inefficient, because intentions could be interrupted at inopportune moments of their execution. When calculating the strength of individual motives, another factor should also be taken into account: their respective deadlines. Ψ does not do so directly, e.g. by incorporating a motive's urgency into the selection function, but by creating another aversive motive, there to make Ψ avoid missing the deadline. This is owed to the fact that the anticipation of missing the

deadline increases the value of other motives relating to that same urge. As long as an intention remains in control of conduct, its plan is executed until it is finished and the intention is dropped. Intentions are also dropped if they can no longer be achieved before the deadline passes or if their motivational pressure has fallen below some minimum, i.e. the urge is no longer there or the estimated competence for attaining the goal is minimal.

While at any point in time only a single intention is chosen by the process just described, the others are not wholly forgotten — their underlying urges are still active and they form part of intention memory. What's more, beside those goals related to the currently conduct-controlling intention, Ψ's *goal list* holds a certain number of goals related to other intentions. The stronger a particular urge is, the more likely it is that goals leading to a reduction of that urge (according to the agent's experience) will be included into the list of goals. Moreover, this likelihood is dependent on the strength of the current intention and the selection threshold: The higher their values are, the less probable the inclusion of side goals becomes. Therefore, we see that the selection threshold influences the agent's tendency for 'narrow-mindedness' by controlling the breadth of its attention — the number of side goals considered — and the likelihood of a switch in conduct control — the bonus of the current intention. On top of the goals to be pursued or monitored, intention memory stores information relevant to the appropriate execution and state of intentions (deadlines, current state etc.).

Knowing which goals it wants to pursue, the agent initiates a sequence of steps to determine what actions to perform in furtherance of these goals (see figure 5). During the first phase the agent checks its surroundings for any changes that may have occurred, i.e. it senses its environment. This orientation process allows the agent to see if some goal state can easily be attained or if some potentially dangerous situation should be avoided. It is to this end that the agent makes use of its goal list. Exploiting potential serendipity or avoiding potential hazards may thus lead to an immediate change of intention and an attempt to perform a consumatory terminal action. If none of this is the case, the agent proceeds with the actual action selection.

Under normal circumstances the agent will have encountered situations in the past in which an urge arose and it was able to satisfy it, i.e. it has goals. Therefore it will be in a position to draw upon such experiences by recalling from memory those stored schemata that led to the goal situations. These schemata are called *automatisms* and the relation of them to the current situation of an unsatisfied urge forms a *motive*. Of course, there is no guarantee that the agent will find an appropriate automatism in its repertoire, due to a lack of experience, the process of forgetting, or simply due to the fact that none are applicable in the current situation. It is worth mentioning that Ψ also incorporates a basic method for the confirmation or refutation of automatisms by checking whether they are consistent with other related experiences. If, for any of those reasons, the agent is not able to come up with an automatism, it has two more options to arrive at a goal-directed behaviour: *planning*, i.e. recombining schemata to produce new possible action

sequences, or *trial-and-error* which manifests itself in specific or diversive exploration. *Specific exploration* describes the behaviour of trying to find out more about the structural or functional nature of a specific object. *Diversive exploration*, on the other hand, signifies the behaviour of searching for situations with a high degree of uncertainty, situations in which one will be confronted with new, as yet unexplored, objects or events.

It must be emphasized that all processes described above can take place with a different level of elaboration and to a different extent at different times according to the current status of emotional modulation, presented next. When highly aroused, for instance, orientation, the search for automatisms or planning may become very hasty and therefore more prone to error than when arousal is lower. As a result it will also switch more quickly from one process to the next. On the other hand, this hastened search for an action sequence will lead to the agent acting rather than deliberating, which can be a very sensible thing to do when the satisfaction of its urges is very urgent.

4.2.2.2 Emotional Modulation Up to this point, we have concerned ourselves with cognitive and motivational aspects of action regulation alone, and — at least according to the classical view of cognitive science, artificial intelligence etc. — this is where we may stop. Sure, if one wishes to explain all the workings of the human mind, one would have to account for emotions, too. But to what avail beside mere theoretical interest? Dörner et al. (2002, pp.195-196) state their point of view on the matter claiming their theory to show

> "that you do not need a special theory for emotions, but, that emotions can not only be integrated into 'normal' information processing, but that the effectivity of human thought can only be explained by the addition of emotions. (...) A feeling is a certain *modulation* of behaviour and the inner processes determined by a certain measurement of *competence* or *certainty*".

They elaborate on this (p.207) by saying that the state of those two accumulators (or motivators) expresses something very important about the relation of the individual to its environment, because

> "[t]he filling of the certainty accumulator 'tells' the organism, to what extent it can expect a predictable environment and the level of the competence accumulator 'tells' the organism, to what extent it can expect to be able to affect its environment. This 'telling' happens (...) by setting certain behavioural tendencies and furthermore by shaping the inner processes in a particular manner."

Any successfully performed behaviour functions as a competence signal and leads to the level of the competence urge rising, while unsuccessful behaviour has just the opposite

effect. If such behaviour also reduces any other urges present, the reward in terms of competence is considerably greater — the opposite holds true for any urge inducing situations. What's more, these situations are stored as appetitive or aversive goals for future action regulation. The effect of an action also has an indirect effect on the agent's feeling of certainty, because it is compared to the effect it expected the action to have prior to its actual execution. The level of certainty is influenced according to the degree of match of expected and experienced effect — confirmed expectations create certainty while unconfirmed expectations create uncertainty (see also figure 4).

As the modulations they are claimed to be, emotions must be composed of certain modulators. Ψ-theory proposes the following modulators whose values depend primarily on the values of the urges for *competence* and *certainty* (Dörner, 2003b; Dörner et al., 2002, pp.197 and the following):

Arousal The arousal, i.e. the readiness to perceive and act (German: Wahrnehmungs- und Handlungsbereitschaft), is the relation between the sum of the underlying urges of all the motives that the system holds, including the currently conduct-controlling one, and the total number of types of urges the systems has (irrespective of their current state). It can be regarded as the amount of energy currently invested and thus the speed with which processing, perception and action take place — the general activation of the organism. (PSI-Glossar, 2005, see also *Aktiviertheit*)

Resolution Level This modulator influences the accuracy with which processing is performed. It not only influences the way the world is perceived and acted upon, but also modulates such internal processes as classification or planning. The resolution level is itself negatively influenced by the level of arousal. A high level of resolution means a high level of accuracy in processes of comparison as well as more elaboration of plans and hypotheses. (PSI-Glossar, 2005, cf. the entry *Auflösungsgrad*)

Selection Threshold We already encountered this modulator in the preceding section. Motives that compete for becoming conduct-controlling must surpass this threshold, whose value is primarily determined on the basis of the level of arousal: the greater the arousal, the higher the selection threshold. The selection threshold essentially represents the system's concentration. So long as it is high, motives that are currently not conduct-controlling will most likely not take over control of action and the system will stay concentrated on its current intention. If the threshold is low, however, the likelihood of the system switching to another motive increases. Thus it determines the systems flexibility. (PSI-Glossar, 2005, see the entry *Selektionsschwelle*)

Scanning Rate This last modulator determines how often normal motive-driven processing is interrupted to make way for an interval of *safe-guarding behaviour* (German: Sicherungverhalten), i.e. scanning of the environment by the agent in order

> to detect possible threats and opportunities. The rate at which this is done depends on several different aspects: past experience with the current motive, the current amount of fear, i.e. the anticipation of a state which raises an urge, and last but not least it depends on the urgency and the importance of the current intention. (PSI-Glossar, 2005; Dörner, 1998, see *Abtastrate* in the former; the latter calls it *Sicherungsrate* (English: securing rate), p.518)

Special types of emotions now simply amount to special combinations of values of these parameters. For instance, anger (German: Ärger) can thus be viewed as an unexpected decrease in competence, high arousal, a low resolution level as well as a heightened selection threshold. Fear, on the other hand, may be described in terms of much uncertainty, a low level of competence, and a high level of arousal. Finally, Dörner is even able to specify common types of personality on the basis of emotions, various behavioural tendencies, and their underlying parameters (see Dörner et al., 2002, p.230 and the following).

4.3 Applications

While the core theory has been around now for about two decades, no Ψs are roaming our desktops, inhabiting the Internet or flying fighter jet missions on our behalf — neither in reality nor in simulation. In fact, application of Ψ-theory has remained solely in the hands of Dörner and his research group who construct experimental scenarios set in a simulated computer world and having the purpose of facilitating the comparison of human subjects' performance with that of Ψ-agents. The so-called *Insel* (English: island) scenario, described in (Dörner et al., 2002), is the only actual application of Dörner's theory to date. Detje (1999), a member of Dörner's research group, performed experiments using a similar scenario called *Landschaft* (English: landscape). He explains the patterns of behaviour that he found in subjects in terms of five different psychological theories of action including Ψ-theory. So, similar to how the *fungus eater* served as a model for Toda's psychological theory of agency, the simulated steam engine living on an equally simulated island, was introduced by Dörner as a model for his own theory. In this, Dörner borrows from Toda's work, as he does in some aspects of Ψ-theory as such.

The *island* scenario is a computer simulation of a robot — the model of Ψ — who lives on an island that consists of regions differing in vegetation and geology, i.e. there are trees and rocks in one region, marshes with shrubbery in another, while a third region is made up of cliffs, sand dunes or lakes. It is Ψ's primary task to survive given varying environmental circumstances. To this end, it must learn to search for and make use of the available resources for fueling its steam engine: water and sunflower seeds or nuts from which internal machinery extracts oil for powering the steam engine. Moreover, the robot agent has to avoid damage to its hull, its sensors and effectors. In addition to caring for its own needs, Ψ is also given the task of collecting as many nucleotides as possible. Nucleotides are described as small rocks that may be used to fuel nuclear

reactors without emitting radioactivity (but Ψ cannot use them for fueling its own steam engine). Nucleotides are to be found in all sorts of different places on the island.

4.4 Current Advances and Future Issues

Dörner's Ψ-agent is constructed as a model of the human being. As such it strives to produce human-like behaviour in every of its many aspects. However, there are some in which it still differs significantly from its human counterpart. The experiments that Dörner and his group have conducted, comparing the performance of human subjects with Ψ-subjects, have shown this. They put the performance differences — quantitative and qualitative — down to two things that characterize humans and that Ψ's currently lack: *personality differences* and *self-reflection/self-modification*. According to them, personality differences often determine the overall strategy that humans follow when solving problems. They simulate these interpersonal variations by creating Ψ's that differ with respect to a number of parameters that are hard-wired for each individual. Among them are, for instance, the amount by which changes in the level of the urges affect the agent's arousal, the sensitivity of success and failure influencing the level of the competence urge, the sensitivity for certainty and uncertainty signals, the frequency of safe-guarding behaviour, or the relations between arousal, resolution level and selection threshold. Dörner et al. (2002, pp.332-341) show how different configurations of such parameters can model different types of human problem solvers with their individual traits of personality.

The aspect of self-reflection and self-modification seems the much more difficult of the two to realize artificially. The importance of this aspect lies in the fact that self-modification allows humans to reprogramme their own behaviour in the process of solving complex tasks for which their standard behaviour is inadequate. They can adapt their strategies and even pursue many goals at the same time. The prerequisite of self-modification is self-reflection, i.e. the ability to make one's own behaviour become the object of observation, and thereby enter a meta-level of thought. Dörner, in line with many other scientists (Dörner et al., 2002, see pp.346-349), believes that humans can achieve this mainly because of their ability to speak: We talk to ourselves about the things we do. More precisely, it appears to be the fact that we ask ourselves questions like "What have I seen?", "What have I done so far?", "What are my goals?", "What is my current strategy and has it been successful?". The observed functions of such self-interrogation are that people order their goals hierarchically, that they form meta-level goals and subgoals, and that they reduce and redefine their goals. By analyzing their behaviour in such a way, humans become better problem solvers. Dörner and his colleagues propose methods of self-reflection for Ψ that analyze the memory structures of the agent, specifically the protocol thread, by making the agent ask itself questions and try to answer them. To this end, but not this end alone, they show how Ψ can be made capable of producing and understanding language (Dörner, 1998, chapter 7). Finally, Dörner suspects that having

language and the ability for self-reflection are key premises for consciousness.

In order to complete the picture of current developments, it is necessary to point to the MicroPsi architecture proposed and realized by Bach and his colleagues (Bach, 2003). It is an attempt to formulate Ψ-theory in a more abstract and formal way so as to make it more accessible to the computer science community, and enhancing it with additional concepts for the formation of categories and attention. In its formulation it makes use of concepts from BDI theory and Sloman's theory of cognition and affect. Bach (2003) states its main differences to Dörner's formulation of Ψ-theory:

> "On the agent design level, MicroPsi possesses a more structured, partly parallel process layout, a divided memory and more general internal representations. On the level of the theory, methods for dynamic category building and for attention based processing have been added, even though implementation is still in its prototypic stages. The original low level perception mechanism of the Psi agents, which is based on a simple image recognition method, is currently not a part of MicroPsi."

The authors are also working on the realization of a 3D environment. They plan to conduct experiments with MicroPsi-agents once the implementation has stabilized.

"It is the mark of an educated mind to rest satisfied with the degree of precision which the nature of the subject admits and not to seek exactness where only an approximation is possible."
—Aristotle

"The important thing is not to stop questioning."
—Albert Einstein

5 A Constructive Comparison

5.1 Scope and Criteria of the Comparison

It is the purpose of this chapter to compare and relate the two approaches presented in the two previous chapters in terms of their nature as agent architectures, i.e. in terms of the forms of behaviour they are capable of showing in practical agent-based applications. This emphasis on practicality relegates possible discussions on their explanatory power with respect to the human mind (or any other natural mind) into the background. Since the usual limitations of a thesis dictate brevity, the scope of the comparison is restricted and cannot grant an inspection of every aspect as thoroughly as might be desired. The discussion follows a breadth-first rather than a depth-first strategy, because the interrelated nature of the architectures' components often does not allow us to pick out single parts without sacrificing our understanding of the agent as a whole. Furthermore, it must be noted that an in-depth comparison is also hindered by the fact that no applications of Ψ-theory exist today other than the island scenario, and that the PRS has not been applied to this, ergo: there are no applications common to both.

The comparison will be conducted on two levels. The first level concerns the characterization of the Ψ approach as a form of BDI (5.2). We will attempt to show that Ψ-theory can be interpreted as a BDI approach in Bratman's sense since it complies with the central demands put forward by him as regards rational agency and the notion of intention in particular. It will be argued that this categorization though valid in principle can be contested in its details which we will indicate. The second level of the comparison is preceded by an imaginary sample scenario (5.3) to facilitate the comprehension of what follows. It will be referred to every now and then as seems appropriate. The second part of the comparison as such draws attention to the architectural aspects of either approach, i.e. to the PRS as a representative BDI system and Ψ. Firstly (5.4), the structural and the procedural elements of the two architectures are dealt with to shed some light on terminological and conceptual similarities and differences — belief, urge, desire, goal, motive and intention. We scrutinize the processes of goal generation, intention generation as well as the modulations of action. Secondly, and concluding the chapter (5.5), we will point to ways in which either approach might sensibly be enhanced with some unique features of the other such that both approaches may benefit in theory and/or application.

5.2 Understanding Ψ as a BDI architecture

Having some idea of what BDI and Ψ are, one will undoubtedly be left with the impression that, while both agent architectures certainly differ in the details of their specifications, their underlying theoretical assumptions and overall structure seem quite closely related. There is, for example, the obviously similar usage of the notion of intention and its relation to plans. So, let us take a closer look into these similarities by assuming Ψ to be an example of a BDI approach. To underpin the legitimacy of this idea, we must check it against some key demands of BDI according to Bratman. These demands primarily concern intentions and the role they play in rational behaviour. We will not explicitly consider the relation of intentions to beliefs or desires (nor that of the latter two among themselves or between each other). Wooldridge (2000, p.25) states that it is surprisingly hard to capture these relationships, and showing them for the Ψ architecture seems too lengthy and little rewarding a process. So, let us instead focus on intentions per se and assess Ψ's intentions/motives in terms of Bratman's criteria listed in paragraph 3.2.1.3. We start, however, with beliefs and desires.

5.2.1 Beliefs

"Beliefs are just some way of representing the state of the world, be it as the value of a variable, a relational database, or symbolic expressions in predicate calculus" (Georgeff et al., 1999). BDI theory does not have much to say about the specific nature of the representation of beliefs which makes it compatible with almost every kind of representation. Ψ gathers information about its environment by means of the HyPercept process which attempts to recognize objects and whole situations on the basis of what the agent has already encountered and encoded in the past and which then activates the stored schemata, if it recognizes them, or creates new ones, if the sensory information is novel. Moreover, it records every perception in an autobiographical protocol thread by connecting schemata to protocol neurons. Together with all the motor information also encoded in schemata, a Ψ has an extensive 'belief system' about itself and its environment which comprises past and present situations and the categorizations of objects provided by the hierarchical organisation of schemata. Even though we have not concerned ourselves with the nature of Ψ's sensors, it is a trivial matter for most agent scenarios to assume that they will not allow perfect sensation. What's more, emotional modulation will occasionally cause the perception process to become less precise, less thorough and/or less frequent. Thus, Ψ's belief is hardly ever complete or completely correct.

5.2.2 Desires

While the identification of beliefs in Ψ was a rather trivial matter owing to the lack of detailed criteria imposed onto it by Bratman and others (see also (Georgeff and Lansky,

1987)), and the mild attention it has generally received in BDI research, the identification of desires proves a more delicate task. This has at least two reasons: Firstly, unlike most BDI systems Ψ is inherently motivated and derives this motivation not from the value of certain goal states alone, but includes the current state of its urges. Secondly, one must distinguish the notion of desire from the notion of intention in a clear way.

According to BDI theory the main three properties of the mental attributes called desires are: They are *pro-attitudes*, they are *the states of affairs that the agent would like to bring about*, and they are merely *potential influencers of action*. We argued in chapter 4 that urges play the key motivational role in Ψ. So are the urges equatable with desires à la BDI? Bach (2003), for instance, calls urges *innate desires* and *somatic desires*. It is obvious that the urge tanks, or *fundamental urges* as we called them, are not what we are looking for. They are not mental attributes. However, urges in the sense of states of disequilibrium also do not qualify, because they do not have the property of intentionality, i.e. they are not directed towards something. They are not states of affairs the agent wishes to bring about. Yet, having some urge, the agent "wishes" to bring about the replenishment of its underlying urge tank. This is the core motivational role that the urge plays and it is clear that it will have to be part of the concept of desire. Commonly, formulations of the BDI paradigm view and implement desires as options, i.e. descriptions of states (goals) considered to be possible for the future with utility measures associated to them. Applying this idea to Ψ, we recognize that motivators and their relations to different goal descriptions (schemata leading to some consumatory terminal action) have this property provided that the motivator is active, i.e. provided it currently signals a need. In other words, motives seem to be good candidates for being desires.

Are motives desires then? They certainly play a motivational role, as their name suggests and their composition of urges proves. However, a motive as a whole is not really a *state of affairs that the agent would like to bring about*, but more often than not it is a collection of such goal states. Considering the idea of desires being options or goals, it seems best to speak of a desire in Ψ in terms of the association of an urge with one particular goal. A desire is commonly seen either as already incorporating the (partial) plans for attaining its goal state or not, i.e. it has either *procedural and declarative goals* or only *declarative goals*[26]. Therefore, since a goal — the representation of a consumatory terminal action — is usually associated to procedures for bringing it about, Ψ's goals are both procedural and declarative. Furthermore, under the assumption that the agent only really desires those states of affairs which it considers possible, the set of desires might be restricted to those that have a positive expectation value. The remaining associations of the motivator would consequentially be 'potential desires', but not desirable in the current situation.[27]

[26]See also (Winikoff et al., 2002) for related work on the distinction between *procedural goals* and *declarative goals*.

[27]Of course, this view of desires can be contested by arguing that natural agents can desire things even if they do not believe them to be possible at the moment they desire them (Detje, 1999, p.17, see the difference between "desire" (German: Wunsch) and "resolution" (German: Vornahme)). Seeing desires as options

As a result, a motive would have to be characterized as a mental attribute composed of multiple desires (and 'potential desires') having a shared motivation, i.e. a shared underlying urge. In light of this view of desires and motives, a Ψ-agent might at some stage of its reasoning be said to be motivated to quench its thirst, having a desire to drink some milk and a desire for it to rain, for example. As long as a desire is not intended it remains a potential influencer of action. In conjunction with what the agent believes, it "merely" provides the agent with a reason for action. So what does Ψ intend?

5.2.3 Intentions

5.2.3.1 General Considerations Bratman's central claim concerning the distinction between intentions and desires, as listed in section 3.2.1.3, is that having an intention entails having entered into some sort of *commitment* to action. Wooldridge (2000, p.34) points out that "an agent has commitment both to *ends* (i.e., the state of affairs it wishes to bring about), and *means* (i.e., the mechanism via which the agent wishes to achieve the state of affairs)". In a way, Ψ's first and most fundamental commitment is the commitment to equilibrate its urge which results in the creation of and the commitment to a motive to do so. It is a kind of blind commitment since the agent will not drop the motive unless equilibrium has been restored. However, is it justifiable to speak of the agent having committed to an *intention* at this early stage of action regulation? The signal of a need, i.e. the creation of the motive, does not yet commit the agent to particular ends or particular means — it may have goal associations but no commitment to them — nor does it initiate any actual search for appropriate action patterns let alone execute them. Therefore, even though this 'commitment' to equilibrium plays a vital role for a Ψ-agent, its inception should not be mistaken for meaning that the agent already has an intention. It is a kind of fundamental commitment that underlies every intention, yet it begins even before we may speak of an intention in the sense established by BDI theory. More concisely, the agent does not *choose* to commit at this point, i.e. it is not a matter of choice between alternatives, and it does not yet commit any resources. Apparently, we have to look at the further stages of processing!

Going back once more to the idea of desires, it is commonplace in BDI systems to regard intentions as selected desires. Applying this view to Dörner's agent architecture cannot be done without heeding the motive core of intentions, that is to say the fact that the set of desires from which to choose the intentions is not utterly heterogenous. Its desires are grouped into motives according to the urges they deal with, if you like. Accordingly, the actual intending is a process of two steps for Ψ: First, a main motive (current intention) is chosen from among the set of motives, i.e. from the set of sets of desires in a manner of speaking. This establishes the agent's primary focus of attention and conduct. From this stage onward, the agent can be said to intend the reduction of the

available to the agent at a given moment, common BDI, however, apparently does not share this view.

urge associated with the chosen motive, committing resources to achieving it. However, it does not yet intend particular ends, nor particular means. What's more, the process can be said to include a moment of choice since the agent weighs between competing motives based on their motivational strength and its overall emotional situation with selection threshold and arousal being the main influencers of this mechanism. It is in this manner and at this point that the structure which Dörner tellingly calls *intention memory* is beginning to take a form that directs it towards action. However, the set of intentions from which to finally choose a veritably conduct-controlling one (in BDI lingo), is not complete yet. The second step determines which desires as part of other motives the agent will also attend to. Depending on their importance relative to the previously chosen (main) intention, further desires may be included in the goal list.

It is a point for discussion if the elements of the goal list are to be seen as genuine intentions or mere desires. On the one hand, many of the goals contained therein will never actually be striven for let alone attained. This leads us to doubt that the agent ever really *intends* them in the first place. On the other hand, given that we not only adhere to the conception of an intention being some chosen desire, but one to the achievement of which the agent commits resources, we notice that Ψ's goal list really contains BDI-like intentions. Resources are made available for the achievement of the desires that partake in it: A situational image and an expectation horizon are constructed and updated in order to check the plausibility of the desires' attainability and timeliness. Also, the state of execution is protocoled. All of the abovementioned characteristics suggest that we indeed make a valid mapping that is faithful to the concept of intention put forward by BDI theory when we state the following: Once the goal list has been established, the agent has intentions such that it intends to bring about a satisfaction of those motives of which the goal list contains desires. What's more, it intends to bring about the satisfaction of some motives more strongly than that of others and this fact is reflected in the number of desires included per motive. Eventually, the process of intending goes on with the choice for particular ends from the list of goals (or desires) and the selection of specific means by recollecting automatisms, planning or trial-and-error. It is at this stage, finally, that the agent really has made a commitment and therefore fully intends. We should now check this more or less direct mapping from intentions sensu Dörner to intentions sensu BDI against the more tangible criteria proposed by Bratman so as to more solidly make a case for it. Since Bratman's criteria primarily concern the property of commitment, the underlying questions are: What exactly do Ψs commit to? And, how committed are they to their so-called intentions?

5.2.3.2 Aspects of Commitment As for the *volitional* aspect of *commitment*: the motive that emerges victorious from the phase of intention selection will control conduct at least until that phase is reached again. In actual fact, taking into account what we just noted with respect to the process of intending, we must make our statement more precise and say that the victorious motive now has the greatest chance of becoming conduct-

controlling. The decision as to which goal the agent's conduct will eventually aim for has not yet been made. All that has been decided at this point is that certain desires will be attended to while others, at least for the time being, will not be taken into consideration. One might say that the breadth of attention, or perhaps the breadth of intention, has been defined. Following this decision, the agent orientates itself (situational image) in order to find out whether or not any of the goals it currently has can be found in its environment. If so, the respective motive will become conduct-controlling and an action — a consumatory terminal action — will immediately be attempted. If there are no goal situations to be found, either because there really are none or because Ψ's resolution level is too low, the search for an automatism will commence. This means the agent tries to find sub-ordinate goals, i.e. situations it believes may lead to goals. Again, should such a goal be found in the situational image, appropriate action follows. If not, the process of planning and, as ultima ratio, the method of trial-and-error will determine the next action to attempt. It must again be stressed that any goal in the goal list may become the one to aim for whichever process of action selection the agent performs. The decisive factors influencing the choice for a certain intention (end + means) to become conduct-controlling are: the size of the goal list, the composition of the goal list (which goals of which motives are present), the amount of time dedicated to each process, and the level of detail to which each process is executed. Obviously, these factors are only secondary compared to the intentions' applicability and motivational strengths.

So, given what was said above, the desire now really becomes an intention in the narrowest sense of the term — a 'final' choice has been made between the different options on offer. Up to now, we have argued that the options partaking in the goal list may already be called intentions because resources have been committed to their attainment and a first choice from among them has been made such that they have a veritable tendency to eventually control conduct. But just how strong is this tendency? Bratman demands intentions to actually control conduct when their time for action arises and nothing interferes. However, he is not very clear about what it is that might interfere. With respect to Ψ there are several related reasons for a motive or any of its desires being denied the privilege of controlling conduct. The most straightforward one is that the motive may have become obsolete due to its underlying urge having vanished — perhaps as a side-effect of the pursuit of some other motive. In this case the motive just vanishes along with the urge even before its time for action has arisen. Secondly, it may be the case that a motive's time for action arrives but that its motivational pressure (expectation$\times$value) does not suffice to exceed the motivational pressure of the currently conduct-controlling intention with its selection bonus. Usually, it will at least have a few of its goals (desires) take part in the goal list, however, it need not necessarily be the case. In turn, not having enough motivational pressure may have two reasons, of course. Either its value is not great enough, that is the underlying urge is not strong enough (and consequently the satisfaction expected when achieving the motive's goal), or the expectation may be too low, i.e. Ψ believes it has become relatively improbable that it may achieve the goal situation.

The processes of emotional modulation does at times make Ψ remain in the phases of action selection long enough for the appropriate time for action to pass as regards other motives. This can result in those motives' expectation to decrease substantially. Even if it does not, the actions performed in furtherance of Ψ's current intention, may have resulted in Ψ now being in a situation from which it does not expect any of the goals associated with the other motives to be attainable. One important point in this respect: Even though the expectation may become relatively low — zero, in fact — the value will likely increase since any non-current urges will not have been dealt with and will have increased values as a consequence.

Turning now to the second property demanded by Bratman: the *reasoning-centred* aspect of *commitment*. Its first component is the demand for stability, meaning that an intention resists reconsideration to a degree greater than a desire does. Just how long an intention persists depends on the agent's current commitment strategy (recall section 3.2.2). We have already pointed to the case in which the reason for an intention, namely its urge, goes away. It is not rational for Ψ to maintain its intention to satisfy some urge by any means if that urge is no longer there. Ψ cannot be accused of any irrationality in this regard for all intentions whose primary motivation derived from the now vanished urge will vanish along with it. Note that this disappearance is not due to a rational decision the agent makes but it is simply the result of a generic mechanism of activation and deactivation of neural networks; it is a fundamental regulatory process. So the terms *reconsideration* or *revocation* as employed by Bratman are somewhat misleading when applied to Dörner's conception. Of course, the central point in the argumentation for a relative stability of intentions is that it is inefficient for an agent — and humans are very efficient agents both in reasoning and acting — to reconsider all possible alternatives every time the environment changes. Many domains require that the agent responds in a timely fashion, and such a response is hindered by too much deliberation over which course of action would be optimal. Dörner et al. (2002, pp. 96–97) point to the fact that it is not advisable to let certain intentions rest in the middle of their execution: "with some intentions, half done is really half done", but others will have to be started all over again after having been halted temporarily. Therefore, Ψ is constructed to make the current intention have a characteristic inertia regulated by the selection threshold parameter. On the whole, however, intentions are reevaluated quite frequently since every time the agent performs the orientation phase and constructs an expectation horizon, the expectation measure of motives is updated based on this new projection. Thus, Ψ has some range of flexibility bounded by its current selection threshold, i.e. by its stubbornness with respect to the main focus of intention. It may seem as contrary to the demand for efficiency that Ψ is made to reevaluate its intentions relatively often but we have to remember that the expectation horizon is not a boundless consideration of all alternatives driven by some deduction mechanism or the like. Instead it relies solely on the agent's experience which is subject to forgetting and thus principally limited. What's more, it is modulated by the resolution parameter which, if quite low, will make the process end

very quickly and result in a shallow, incomplete projection. In such situations, Ψ will stick to its guns and continue with its current intention rather than reconsider.

The third demand states that having committed to a set of intentions, one will frequently reason from the intended ends to intended means. As a result, new intentions may be created and committed to, often representing intermediate steps on the way to other intentions. How does Ψ reason once it has intentions? The goal list contains certain ends that the agent has committed to in its own way of commitment which we described earlier. Now, choosing a goal from among this list would mean that the agent has actually found it in its situational image and this will lead to an attempt to perform a consumatory terminal action — no further reasoning is necessary. If no goal is found, however, the agent begins its search for other means of reaching one of its goals, i.e. it searches for automatisms or plans, the execution of which might lead it to attain a goal state. As mentioned, the intermediate schemata on this devised path of actions and events can be considered sub-ordinate goals. Staying in line with our established terminology, the agent does not so much create a new intention with the sub-ordinate goal as its end, but the already existing intention matures, so to speak, by committing itself to these new means. The orientation phase has more consequences beyond constructing a situational image and seeking goals therein: the update of the expectation horizon may lead to new intentions being created when potential chances or dangers are anticipated. However, this anticipation is not done *against the background of the intentions already committed to* — it is quite independent of the agent's commitments. This leads us to the last requirement Bratman proposes.

Option admissibility, i.e. the fact that intentions help *limit the options* that need to be considered for the future, is perhaps the least attributable property as regards Ψ. The agent is required not to entertain options that are inconsistent with its intentions. BDI implementations which use some kind of symbolic logics as their standard form of representation often have explicit filtering algorithms for detecting such inconsistencies. In the literature on Ψ-theory no such check is to be found, but Dörner et al. (2002, pp. 139–144) describe a process which might be used to this end. The process is called *confirmation and refutation* — we mentioned it briefly — and it is employed in the search for automatisms. Given a certain candidate automatism, it goes through its sequence of motor schemata, first checking if they also form part of other memorized schemata and, if so, comparing their outcomes. The greater the number of different outcomes, the less probable the success of the candidate automatism becomes, at least according to the information Ψ has gathered in its past. It goes without saying that this process, as any other in Ψ, is realized completely as waves of activation through a network of quads. It is plain to see that *confirmation and refutation* could also serve as the basis for checking the consistency of intentions. Every time a new commitment to particular ends or means, or both, is about to be made, this process could check whether or not the candidates are compatible with the ends and means that the agent has already committed itself to.

It is important to add that the criteria proposed by Bratman are rather vague and leave a lot of room for discussion. What's more, the examples which he produces often refer to situations in which higher level decision making mechanisms or consciousness are required. Ψ, as stated in section 4.4, is not capable of such mechanisms yet. Therefore, we have to state: Ψ-theory corresponds in large parts to BDI theory, i.e. it meets its principal requirements regarding intentions, but it only does so on a level excluding effects of consciousness and higher order reflective processes. This is not to say that Ψ lacks the potential for such advances but rather that they have not been approached yet. Similarly, traditional BDI approaches, including the one that we focus on in this work, also lack many such capabilities at present. Even more interestingly, however, they themselves do not seem to meet all the basic requirements regarding intentions as Wobcke (2002a) points out.

5.3 An Example Scenario – Arthur, a Digital Personal Assistant

So far we have been talking about agents in a fairly abstract way, both about their desirable properties and their internal mechanisms. In order to make the following discussion more tangible, we will illustrate many aspects with the help of a somewhat idealized example scenario. The example is derived from the idea of having a digital sister-in-law, as outlined in section 2.1.3.3 and as proposed in the famous commercial video by Apple at the end of the 1980s titled *Knowledge Navigator*. Therefore, let us consider the following digital assistant.

Suppose that you have come home from work and you sit down in front of your computer. You start it and a friendly face appears on the screen greeting you, *"Good Evening!"*. Your personal agent — let's call him Arthur — continues by saying, *"Before I tell you about the things I managed to do while you were at work, let me remind you that you are invited to a dinner party at the McKenzie's house tonight at eight. Jack called this afternoon to ask if you could pick him up on the way and give him a lift there. By the way, I have a route that I could print out for you in case you are going. Unfortunately, the highway was closed down two hours ago due to a traffic accident, so I'd advise you to leave earlier, at around seven thirty. Would you like me to dial Jack's number for you?"* You decide to call Jack to confirm that you are going to pick him up and tell Arthur to print out the route.

After a while, you ask Arthur what else he wants to inform you about. Your agent answers, *"While you were away I went looking on the Net for a camera at a reasonable price. I found the model you were searching: XYZ-Megastore are selling it for 250 Euro. I also found someone offering one at Ebuy for 100 Euro, some parts of which are broken, however. I contacted Lee's Electro Shop downtown and sent them the description. Their agent replied that they could fix it for a price of around 50 Euro depending on the exact damage. As far as I can tell, it is a very reliable service."* Thinking for a second you agree to buying the new camera from XYZ-Megastore and ask Arthur, *"Anything else?"*. Arthur responds, *"No, but I could update*

you on the parties and concerts this week-end, if you would like me to? The Raving Rascals are playing at the Rigid Rabbit on friday — Not right now, thanks. I'd better hurry up to be on time for tonight's dinner!"

While Arthur informed you on many urgent or otherwise important things, he left out a lot of others, knowing that they would be of no interest to you now, but which he had also attended to during your time at work. For example, Arthur knows of your interest in history, Lithuanian history in particular, and your sophistication in speaking Russian, but also your love for languages as such. From this knowledge, Arthur inferred that you might be interested in literature on the history of the relationship of Lithuanian and Russian in light of the historical developments in the Baltics. Aware, however, that his knowledge of how to acquire information concerning the subject matter, and more importantly, his knowledge of how to acquire literature in general, was very poor, to the degree that if you asked him to gather it now it would take him unacceptably long, Arthur decided to improve his competence in this regard. By querying the few services free of charge plus some at very low costs, and by trading knowledge with other more competent agents, he was able to collect a first compilation of material which he will present to you tomorrow evening. Since this was not entirely for free, he also plans to ask you for some more pocket money...

So much for Arthur. Obviously, the scenario is still a vision today — note again: the vision of Apple's Knowledge Navigator was uttered in the 1980s. This is not only due to the hard task of teaching computers to understand and speak natural language. When this vision was produced, it was neither possible to create such agents, nor did the environment exist in which they could have acted so as to adequately serve their masters. Today, however, things have changed at least as far as the latter restriction is concerned — the *Internet* — and they are going to change even more so in the foreseeable future with the inception of the *Semantic Web* and the *Service Web*. Hence, pressure on agent developers to construct assistants like Arthur is likely to be mounting. Are the architectures introduced in this thesis keys to a solution?

5.4 PRS and Ψ — Comparing the Structures and Processes

In order to avoid misunderstanding, it is important to note that when stating something as being the case for BDI, we mean BDI as it is commonly realized in systems like the PRS. Bratman makes no statements as to how exactly any of the mental attributes which he proposes are to be realized in artificial agents, or indeed how they are realized in humans. This section will compare the two models directly. First the forms of knowledge representation will be discussed along with some aspects of memory. Following this, the discussion will turn to how each approach realizes the reasoning necessary for goal-directed behaviour to occur. The example scenario will serve as a means for clarification: We translate the processes studied into a PRS-Arthur and a Ψ-Arthur.

5.4.1 Knowledge Representation and Memory

As far as belief representation is concerned, or knowledge representation as it is more commonly called, Artificial Intelligence researchers have devised two principal ways for its realization, both of which we encounter in the two approaches: symbolic, we find this in the PRS, and subsymbolic, which we find in Ψ. Not only the analysis of rational agency in BDI agents, but also the representation of information in the systems themselves, takes the form of logical expressions which are of a symbolic nature. More precisely, declarative information, i.e. factual information, whether innate or derived from sensory input, explicitly takes this logical form, while procedural information, i.e. information on how to act (action sequences, plans), is represented in a standard STRIPS-like fashion which is of course also highly symbolic. Ψ, on the other hand, stores its information in a network of simple artificial neurons each single one of which does not represent any facts. The representation is hierarchical and distributed over the network in a fashion that encodes *microfeatures* of whatever is being represented by the whole (Wilson and Keil, 2001, see the entry "Local vs. Distributed Representation"). Moreover, the method of representation for declarative and procedural knowledge is exactly the same. As a consequence, there is no separation of declarative memory and procedural memory, there is only one memory system, whilst the PRS entertains two: the *database* and the *plan library*. We cannot go into the subtleties of representation here, nor would it help us much if we were to consider the opposing arguments in the symbolism vs. connectionism debate.[28] There are however, some facts worth mentioning.

Firstly, the connections forming a Ψ's mind contain probability estimates for the success of actions and the occurrence of events. The Acts of a PRS-like system do not feature such estimates — recall (end of section 3.2.2), however, that outcome probabilities can be integrated into BDI theory even though this has apparently not been done very often in practice. Why are such probabilities important? Arthur, for instance, would hardly be able to say that he thinks that Lee's Electro Shop has a reliable service, if he was not able to represent the probability of outcomes of certain actions, in this case: the probability of a broken electrical device being returned by Lee's Electro Shop in an intact state. Even more fundamentally crucial to this ability is, of course, the general ability to memorize, which leads us to another issue.

Secondly, learning is an integral part of Ψ whereas it is not in the PRS. New neurons and new connections are formed whenever situations occur that have not been experienced before. The information contained in connections (probabilities and strengths) is

[28] The interested reader is referred to (Dörner et al., 2002, pp. 14-15, pp. 65-68 and pp.172-175) for short discussions on the relation of Ψ's method of representation to symbolic approaches. A subproblem of the whole debate is the question of how to represent variables in connectionist systems (e.g. see Marcus, 2001, for a general discussion), and it may be interesting to answer the question whether Dörner's approach is indeed capable of representing variables in the way the PRS is or not. A starting point is provided by what he calls abstract schemata (see e.g. Dörner, 1998; Dörner et al., 2002, pp. 142-144 and pp. 267-270; pp. 59-62).

adjusted whenever these connections are activated (reinforcement) or when they are not used over a long time (decay). Broadly speaking, Ψ's memory works according to the Hebbian learning rule which is often reduced to the phrase: "Neurons that fire together, wire together". The strengthening of connections allows the respective schemata to be more easily retrieved than before and the process of forgetting further adds to this (since unimportant information simply vanishes over time) while also ensuring that valuable space is not wasted. On top of these forms of learning and the learning of new operators when successfully executing newly combined action sequences or trial-and-error behaviour, Ψ is also able to create genuinely new goals, i.e. it is autonomously able to store novel situations in which consumatory terminal actions are possible as generally desirable (more discussion on this later) — note that we are talking about learning goals that are stored in *long-term memory* and not simply about generating options in *working memory*.

Moreover, Ψ does not only learn new action sequences and identify (and categorize) new objects in different situations, but it also remembers when and in which order they appeared. This is owed to its protocol thread which, even though it will fall apart with the passage of time, equips the agent with an autobiographical memory, one facilitator for learning (and for reflection!?, which neither the PRS nor Ψ feature yet). Naturally, the importance of an autobiographical memory begins to show whenever certain temporally related faculties are required of an agent. For instance, remember that Arthur anticipated that you would want a route to be printed out and up-to-date traffic information so that you would be able to pick up Jack easily without being late for the party at the McKenzie's house. Now, of course, you or Arthur's programmer may have 'told' Arthur previously, to some level of detail, what to do in such cases, i.e. how to behave in foreseeable time critical situations where you need to get from A to B by some form of transportation, possibly via C, and so on. This presupposes that one is able to and willing to specify such behavioural patterns a priori; surely, it would have to be done for a PRS-Arthur. However, an agent software working for many different individuals with their individual preferences and needs, or being part of a complex and inaccessible environment, may profit immensely from having some way of learning and adapting, and the ability to notice recurring temporal patterns may be one fundamental part of that. A well known drawback of learning is that it takes a lot of time which is why it has not found its way into many of the applications of BDI agents which are known for their real-time behaviour. A much more obvious restriction of a completely amnesic Arthur is that he could not tell you that Jack called "this afternoon" or that *first* Jack called, that he *then* went looking for a camera and that he *finally* contacted Lee's Electro Shop.

5.4.2 Goal-Directed Behaviour

Let us now examine how each approach realizes the generation of actions. We will be focussing on goal-directed (pro-active) behaviour since that is, ultimately, the realm of

agent architectures which have intentions at their core. For the purpose of making the discussion in the following three sections more transparent, let us differentiate between three kinds of *goals*, or rather three stages of goals, for which we shall use the terms listed below. We will focus on deliberation rather than means-ends reasoning for we believe this to contain the more interesting issues. Therefore, when talking about goals here we mean declarative goals above all.

potential goals are representations of situations that are generally desirable for the agent to bring about and which may, in the course of future reasoning, become activated as *relevant goals*

relevant goals are *potential goals* that have been chosen as options (more or less the notion of *desires* we mentioned in 5.2.2), because they are relevant to the agent's current situation and toward which it considers to direct its behaviour, and

actual goals are those selected *relevant goals* that, normally as part of intentions, actually guide the agent's actions, i.e. the agent actually directs its behaviour toward attaining them.

We can now formulate a number of essential questions which an agent architecture can be asked to answer. When, how and why does an agent acquire potentials goals in the first place — this is the question of *goal acquisition* or *goal learning*? When, how and why does an agent activate a set of potential goals and thus make them relevant goals for further reasoning and action — this we might call the question of *motivation* or *option generation*? Finally, when, how and why does an agent select one or some of the relevant goals as actual goals which it then pursues — essentially this is the question of *decision making*? Usually, agents are further forced to choose from among the actual goals since they cannot execute them in parallel. They have to determine what we might call a *current goal* towards which the agent directs its behaviour at the current point in time. In what follows, this point shall be subsumed under the subject of *decision making*.

5.4.2.1 Goal Acquisition The answer to the first question already shows us one point where the PRS and Ψ differ in principle. As mentioned several times before, there is no learning of any kind involved in the PRS architecture as such. Therefore, an agent that is equipped only with the structures and processes provided by the standard PRS specifications will have to manage with those potential goals that it was initially provided with by its implementor. Of course, once having adopted some potential goal as an actual goal by placing it on the intention graph, subgoals may be generated, yet these subgoals, being intermediate steps on some Act, are also innate. For Ψ things stand very differently: Learning is an integral part and this includes the learning of new potential goals. It is achieved by combining the general ability to memorize the percepts one encounters and the actions one takes with the changes of one's internal state. By internal state we

mean the respective states of the urges. Thus, the agent can *itself* place value on novel situations depending on whether or not they affect its internal state. It associates the representations of the situations, the schemata, with its urges and so creates appetitive and aversive relations. This makes those schemata potential goals for future consideration. The association is an automatic process which always occurs whenever Ψ has performed some *consumatory terminal action* or experienced a *painful situation*. Let us elaborate a little more on the aspect of utility.

Placing value on situations and actions is at the heart of much of AI. Some states have to be considered more desirable than others, otherwise no (rational) decisions can be made. Simply put: no preference, no goals. Preference is usually a combination of some estimated probability of the goal being reached, at least for domains that are not completely certain or deterministic, and a utility measure for that goal. The hard part in practice is finding both an adequate probability distribution and an adequate utility function for a certain domain (cf. e.g. Wooldridge, 2000; Dastani et al., 2003; Russell and Norvig, 1995, p. 11 and pp. 475-480 respectively). This is why one often resorts to learning these functions. Let us consider the case of utility. Due to the domain-dependence of any utility function an architecture cannot prescribe a general utility function — this is not the purpose of an agent architecture. As a consequence, the literature on the PRS does not mention utility explicitly. Much the same holds true for Ψ: utility, which is of course utility with respect to the agent's urges, is only explicated because one has a particular agent in mind, namely a steam engine agent living on an island and built to model a human agent. What types of urges the agent actually has should not be important for defining the architecture — although perhaps competence and certainty can be regarded as very useful across very many domains since they measure how well one is able to solve one's problems in general. So, instead of prescribing a particular utility function, an architecture can at best define where and when a utility function should be or needs to be applied, it can constrain the utility function and/or its components in some way.

The PRS makes no assumptions as to a utility function (nor does it make any assumptions concerning probabilities of outcomes). As for Ψ, while it does not define a utility function either, it does constrain it in important ways by demanding that it be dependent on urges which reflect the homeostatic tendency of a regulatory system. What's more, these urges are integrated into two more processes beside goal learning, but which are just as fundamental to the whole conception of Ψ-theory: *motivation* and (emotional) *modulation*. Laxly speaking, *motivation* determines the general directions that action should take and *modulation* determines in what manner action will take place[29]. The matter of motivation on the other hand will concern us next. One remark though before we continue: despite Ψ constraining possible utility functions in the ways just described, finding adequate utility functions for domains remains a hard task: adequate urges must be

[29] *Cognition*, or more acurately, decision making determines what action will actually be performed — if one wishes to make this traditional distinction between motivation, emotion and cognition at all

found, the adequate quantitative relations between the urges must be determined, and their individual effects on emotional modulation must be quantified, whether once and for all (see Dörner et al., 2002, pp. 230-248 on the relation between the parameters of emotional regulation and personality) or perhaps also subject to learning processes.

What urges should Arthur have? As stated, an urge for competence and one for certainty seem advisable, because they tell it how well it is generally faring in solving its problems. We will further discuss this when we turn to emotional modulation. But what else beside them? Surely, an urge for making money is necessary, is it not? Maybe it should always be active in order for Arthur to have an unappeasable lust for riches — is this really desirable?[30] On the other hand, Arthur should also have an affiliation urge inextricably tied to its user, i.e. Arthur should be strongly depended on legitimacy signals from its user. Can we reduce this to one urge maybe, since making money always pleases its user!? Not really, suppose Arthur had got the money for the camera it bought you by selling some stocks it had previously bought from a company you do not like for some reason (morals, competition etc.). It is an interesting question to ask which urges an agent needs in order to be or become fit for its purpose and which of the urges should be more important than others.

Supposing it has the abovementioned four urges, how does Ψ-Arthur learn a new goal and what does it mean? Consider Arthur actually having found some literature on the history of the relationship of Lithuanian and Russian in light of the historical developments in the Baltics. Once it purchases this literature it will learn to avoid purchasing it in the future, because its money urge rises and the situation is thus stored as an aversive goal. However, if we assume that you like what it has found, it will be rewarded with strong L-signals when it presents its result. Of course, Arthur will have anticipated this reward beforehand, otherwise it would not have considered purchasing that literature in the first place. So, this situation of telling you of its purchase will now be an appetitive goal. Notice, that Ψ-Arthur might tell you of a purchase next time without actually having bought something only to get positive affiliative feedback from you — it may "lie". Therefore, it is probably better to not provide this feedback until you actually receive the goods. This is certainly an interesting if not necessary ability, if we want to model a human-like character. For an assistant agent it is probably less desirable — at least with respect to its master. But let us continue and we will notice some more interesting properties.

5.4.2.2 Option Generation When, how and why does an agent activate a set of potential goals and thus make them relevant goals for further reasoning and action? Concerning the PRS we are interested in the generation of the set of desires. Concerning Ψ we ask for its motives.

[30] This is a practical question as much as an ideological one.

Deliberation in problem solving agents is often viewed as a two-step process of *option generation* and *filtering*. This division says that an agent first tries to understand what options are available to it — the transition from potential to relevant goals — and then filters out those it considers (rational) to attempt — the transition from relevant to actual goals. Potentially, there are a multitude of different factors that may influence an agent's deliberation. However, there are some that are predominantly used: The agent might consider particular goals attainable from the situation it currently finds itself in and not others, with some attainable goals believed to be more probable than others. The agent may also believe certain goals to have a higher general utility than others. Moreover, the agent's internal situation (whether physiological — osmotic pressure in the cells, hull damage, hunger urge etc. — or informational — e.g. the accumulation of internal or external environmental feedback to one's past behaviour) may motivate it to attain certain goals rather than others. And finally, the agent may consider some goals to be more consistent with its intentions and/or its experience than others. Often agents use a combination of these factors for their control. It goes without saying that no principal judgements as to the usefulness of any individual influencing factor, the usefulness of a combination of factors, or the likely quality of the decisions can be made. Whether or not an agent needs such decision facilitators depends on what it is supposed to achieve and where, i.e. in what domain and environment, it is supposed to achieve it.

The way of generating options found in the PRS is that of trying to unify the conditions specified in the Acts of the plan library with the beliefs currently held in the database. Any Acts that become active as a consequence of this are considered options. Recall that Acts are declarative and procedural goals, i.e. an option is both a goal state and an action sequence that (potentially) leads to it. In particular, the generation of options only takes an Act's goal description into account if it currently holds unattended goals in its database. As stated earlier, BDI theory is primarily about the relations of its three attributes to each other and how they should be constrained to fulfill the requirements of rationality. This means that one may have to define meta-Acts in order to model the more delicate transitions of deliberation. Rao and Georgeff's extension to the theory also allows for the additional inclusion of utilities and probabilities, yet it is seldom employed, and if so, utilities are predefined by the designer just as the Acts are predefined. Therefore, applications of the PRS concentrate on implementing the desirable properties of the attribute interrelations by predefining Acts in such ways as to ensure these properties, possibly with the additional help of meta-level reasoning. The standard method of option generation will simply react to changes in the agent's environment as well as to possible goals and subgoals found in the database. Neither experience nor internal states such as urges are usually influencing factors in PRS-like agents which is not to say that the architecture itself prohibits the inclusion of them. There is no claim for the database to contain only facts from the *external* environment. In principle, the database might also maintain (symbolic) information on the *internal* status of the agent including information on its motivations, its urges. Therefore, even though the PRS does not pre-

scribe the existence of urges, nor propose ways of integrating them, this does not imply that motivational aspects, like those of Ψ could not be integrated. In fact, in section 5.5 we shall consider briefly how this might be done. There are several approaches to equipping symbolicly reasoning agents with motivations (see e.g. Munroe et al., 2003a,b, and those they refer to).

Understanding the main difference separating the two approaches as regards the generation of relevant goals requires that we take up the point we just established concerning the introduction of urges. The following quote from Munroe et al. (2003a, p. 1074) may help emphasize the effect of an agent having a set of innate urges (here called motivations):

> "Like the traditional notion of utility, motivation places value on actions and world states, but is a more wide-ranging concept than utility. In the traditional view, an agent examines its options and chooses one with the highest utility. Motivation also enables this to be done, but it is also more intimately involved with the agent's decision-making process than utility. A motivated agent has a dynamically changing internal environment, provided by motivations, which can influence its decisions. For example, in the presence of food, an agent may or may not choose to eat depending on the state of its internal environment (specifically its hunger motivation)."

Obviously, the concept of a *motivation* that Munroe and his colleagues have in mind corresponds directly to Dörner's conception of an urge as the signal of a need, and what they call *internal environment*, he would call the collection of urges. The status of the urges at a particular point in time directly influences the decision making processes at that particular point in time. For Ψ this means, above all, that they influence the step of activating a set of potential goals as relevant ones, since nothing is ultimately more relevant to Ψ than an equilibrium of its urges. Therefore, potential goals that are connected to the active urges, i.e. those urges in a state of disequilibrium, will become options for the process of intention selection in the form of motives. The lower the level in the tank of a particular urge, the higher the relevance of those potential goals associated with the urge and the greater the motivation for choosing one from among them as an actual goal later on when making the final decision. The generation of options is thus highly depended on the agent's internal environment. The external situation, however, is also taken into account by way of the expectation measure and its component measure for epistemic competence in particular. Motivation is an integral and fundamental part of the architecture of Ψ whereas it is not so for the PRS. Ψ's *are* inherently motivated agents whilst PRS-agents *may be*.[31]

[31] It should be obvious that the inherent internal motivation of Ψ was the prime cause for our slight difficulties in scrutinizing the concepts of desire, motive and intention at the outset of this chapter.

What of Arthur? Let us again take the example of Arthur providing you with literature on some subject. For a PRS-Arthur without a motivational component to generate such an option, i.e. to have such a desire, would require that there be some environmental trigger and/or some explicit goal in its database. An environmental trigger might simply be the presence of beliefs concerning your interests in some topic and the availability of money — say, to use the service. An explicit goal could just state that the agent should achieve the collection of literature or it might even be so explicit as to even specify further constraints on the type of literature etc. (obtained directly from the user, for instance). Other than these triggers, there is the requirement of there being some Act, however detailedly specified, whose conditions match some or all of these triggers. For instance, there might be a general Act that specified how to go looking for new information on the Internet in a series of steps and subgoals: achieve contact search engine service, achieve get results for topic X, and so on.

Turning to Ψ: The added motivational aspect in Ψ-Arthur would changes things in important ways. Its generation of options begins with the generation of at least one motive. Thus, it would have to have an active urge for affiliation, for example, or for certainty (or even for both at the same time). Being sufficiently great, the urges would cause the creation of a motive (or more motives) in order for the agent to deal with their reequilibration. Then, provided that Arthur has some sort of "researching automatisms or plans" at its disposal and they are included into the goal list, pro-active researching behaviour becomes an option. However, in case Arthur's relevant urge tanks are relatively full, it may not even consider this pro-active researching behaviour at all unless, for example, you demand it. Demanding it would either mean promising a reward — L-signals or money — or threatening to punish — U-signals or the withdrawal of money.[32] These considerations show the level of autonomy that a Ψ-agent is capable of. Whether or not this is desirable for practical applications is another matter altogether.

Moving on to the next step in the cycle of action regulation, we see even more clearly what effect the introduction of urges has in the Ψ architecture, i.e. how it integrates *internal* and *external* environmental conditions into its decision making process.

5.4.2.3 Decision Making When, how and why does an agent select one or more of the relevant goals as actual goals which it then pursues? This is often called the option *filtering* process. It establishes a set of intentions that will organize the agent's behaviour. The final step of decision making then chooses a current intention on the basis of which a concrete course of action is initiated.

The default decision making process for the standard PRS is very simple: the filtering behaviour provided by the architecture just chooses one option at random from all appli-

[32]The interested reader is referred to (Dörner, 1998, pp. 686-690) for more ideas on how a Ψ-agent may understand and react to demands or orders addressed to it by another agent.

cable Acts for the current cycle. However, we should not forget that meta-level reasoning can be applied to reduce the set of options. Meta-Acts, as described in section 3.2.2, are the central means by which deliberation can be molded in order for the agent to exhibit the desired behaviour. However it is actually realized, there is an explicit step from option to intention in the PRS. Intending means placing an Act on the intention graph either as a new root intention or as part of an existing intention, if it is a response to a subgoal of that intention. After filtering out new intentions and committing to them, yet another decision must be taken, namely which of the intentions actually takes over control of the agents actions. In the PRS a root intention on the intention graph must be chosen and pursued. There is a default method that the interpreter uses to choose from among those intentions that are not in the *sleep* state: it picks the most recently awoken intention if there are any intentions in the *awake* state; it picks one at random otherwise. Again, many applications will require more informed selection procedures and implement them using meta-Acts.

Ψ's deliberation is primarily about setting a focus of attention on a particular motive, on a particular urge to be dealt with. Not all the options generated can be pursued or are really relevant. The filtering process commences with the selection of a current motive according to the expectation×value-principle. All of the goals of this current motive plus a number of side goals from other motives together form the current span of attention of the agent. The decision for and against goals is thus influenced both by the external environment of the agent — the epistemic competence measure is based in part on the situational image — and its internal situation — mainly the value and the general competence. Beyond this, there is the regulative effect of emotional modulation. Especially the selection threshold plays an important role here: it is dependent on the level of arousal. A low selection threshold means that intentions can be changed relatively quickly which gives the agent a certain amount of flexibility. More side goals are included into the goal list. If the selection threshold is high, on the other hand, less side goals are allowed for which makes the agent concentrate more on its main intention and decrease flexibility. Therefore, since deliberation is also constrained by emotional modulations and these in turn rely heavily on the motivational status of the agent, the setting of the innate parameters that determine how each affects the other will influence immensely what type of 'personality' an agent has. Following the creation of its goal list, Ψ starts orientating itself and perceiving its surroundings. It then checks to see whether or not one of its goals, or at least some subgoal, can be found in the situational image it has created. In case it does identify some goal, it will immediately attempt to perform a consumatory terminal action or at least a basic action towards the goal. Otherwise, Ψ will try to remember a goal-directed automatism, create a goal-directed plan or simply resort to trial-and-error behaviour. We see that Ψ's choice for pursuing a particular goal now depends solely on its perception of the current situation.

Looking at the way new intentions are created and how the set of all intentions is maintained the main difference in the approaches is that the PRS's default behaviour is

extremely basic in that it does not make a lot of assumptions. It leaves it up to the agent designer to find appropriate mechanisms and specify them using meta-Acts. In stark contrast, Ψ's intention creation and selection mechanisms are much more elaborate including such influencers as the agent's competence and internal motivation and being modulated by regulating variables. Above all, the level of abstraction of the PRS's specification means that agent's developed with it do not automatically adhere to the core procedural principles of BDI theory, at least not in the form conceived of by Bratman. In (Georgeff et al., 1999) Pollack makes the following statement on this issue,

> "[W]hile it is possible to build a PRS application that respects Bratman's Claim [stating that rational agents will tend to focus their practical reasoning on the intentions they have already adopted, and will tend to bypass full consideration of options that conflict with those intentions] (...) it is also possible to build PRS applications that embody alternative BDI models. It is up to the designer of a PRS application to specify how beliefs, desires, and intentions affect and are influenced by the application's reasoning processes; there is no requirement that these specifications conform to Bratman's Claim."

The first part of this comparison made clear in how far Ψ can be seen to adhere to Bratman's Claim. The main difference between Ψ and typical BDI approaches in this respect is that the generation and selection of options by Ψ is not done before the background of what is already intended while BDI approaches explicitly focus on rational consistency. The only effect of an intention on another that Ψ-theory knows is the lateral inhibition which the currently conduct-controlling one exerts on all others.

Finally, regardless of whether or not one implements a PRS-like agent to conform to the principles of BDI, there is another aspect that should not go unnoticed. It is the trivial fact that all meta-Acts like other Acts may become intentions. Therefore, when implementing meta-Acts to control the agent's reasoning one can easily compromise the plausibility of the approach with respect to being a model of human reasoning by introducing meta-Acts that fulfill some computational purpose but would never really be "intended" by natural agents. In the last part of this comparison we will give an example where this could be the case: the introduction of urges and their manipulation into the reasoning cycle (section 5.5.2). Still, in most cases in which meta-Acts will be used, such as realizing informed choice among multiple applicable Acts or intending more than one Act per interpreter cycle, for instance, the concern should be slight. Typically, they will reorder intentions on the graph or even kill intentions on that graph. For instance, when explicitly reconsidering an intention the agent may decide to kill that intention, i.e. it will drop its commitment to that intention. Thus, decision making in PRS-like agents amounts to an explicit maintenance of the set of intentions which can be in the form of consistency checks among intentions (or intentions and other attributes) and realize particular commitment strategies, usually variations of those we mentioned in section 3.2.2.

As for commitment strategies, one notices that Ψ is 'blindly committed' to the satisfaction of its urges, the core of every intention, as stated earlier. It's commitment to ends and means is rather a special form of single-minded or open-minded commitment which depends on its expectation and how emotional modulation forces it to focus its attention or speed up its decisions.

Returning one last time to Arthur. Suppose there were three options that PRS-Arthur had generated: one to go searching for literature on lithuanian history at a certain database service, another to go buy the camera at Ebuy and a third to buy the camera at XYZ-Megastore. Now, assuming it is equipped with some meta-level reasoning, the following process may be run: a meta-Act checks the applicable options for consistency amongst themselves. For instance, it might detect that the latter two options should not both be intended at the same time, because it would mean that the agent might end up having two cameras and having spent more money than it currently has. Furthermore, the option to buy the camera at Ebuy seems to conflict with an intention already on the intention graph. Therefore, the set of options is reduced to the two others and both are intended. Say, the literature search intention is a response to a subgoal posted by another intention on the graph — for instance, the intention to gather information on Lithuania, of which historical information is a part — and consequently it is enqueued on that intention's branch. The camera purchase intention on the other hand becomes a new root intention. Remember, all these informed choices must be implemented in the form of meta-Acts of some kind and are certainly not a part of the standard PRS. Finally, one of the root intentions, for example the camera purchase, will be chosen and its next step is executed, say the contacting of the XYZ-Megastore sale service.

Ψ-Arthur has two active urges, namely the one for affiliation and its urge for competence. There are several affiliation-related and several competence-related goals it has learned in the past. Ψ-Arthur engages in its process of intention selection and establishing a goal list. However, it is very angry at the moment, i.e. it is highly aroused and therefore its selection threshold is quite high. All this is due to its strong urge for competence and its more mild, but still present, need for certainty, which are themselves the results of Arthur's failures in the recent past. For instance, it was incapable of purchasing a first class ticket to Sydney for you because all the services it could reach were sold out. So, its anger now makes it somewhat narrow-minded, i.e. it concentrates on its current motive. Also the fact that its resolution-level is low makes the whole selection process become very quick. According to the expectation$\times$value-principle it chooses to focus on its competence urge, primarily because its value is considerably higher than that of the affiliation urge. However, there are a few side goals with respect to affiliation which make it into the goal list despite the agent's concentration on acquiring competence. Suppose that among those goals which finally remain there are goals similar to the information gathering and camera purchasing ones introduced above. Arthur's next step is to orientate itself and thereby create a situational image. However, none of its goals seem to be attainable right away, so it searches for an automatism and actually finds one that seems

to be applicable in its current situation. There is, in its mind a direct way to go from the current situation to being in possession of a camera and this success quenching its urge for competence (and maybe even affiliation). It has enough money and the situational image 'tells' Arthur that XYZ-Megastore is open for business. Arthur takes the first step of action and contacts the sale service. However, since its resolution level was so low it did not look into its mailbox and missed the fact that another agent had just offered it a great deal on a camera. Maybe, it still finds out next time, but it may be too late? But here we must leave it.

5.5 Possibilities for Mutual Enhancements

Throughout this comparison we have made an effort to show where BDI theory and Ψ-theory share common principles with respect to how they explain reasoning and the role that intentions play in it and where they do not. Moreover, we have discussed the similarities and the differences of their concrete architectural realizations in the PRS and Ψ. In order for all of this to be fruitful beyond the gross classification and the structural and procedural comparison that were provided, we might ask ourselves: To what avail might one go about enhancing the one with elements of the other? And, if such mutual enhancement is indeed warranted, how should it be done?

5.5.1 The Motivation for a Hybrid System

What can we expect to gain from the envisaged endeavour of melting the PRS and Ψ? Let us begin by considering the bigger picture. We are dealing with two theories, BDI and Ψ, whose main goal it is to satisfactorily explain how agents, human agents in particular, manage to act adequately (i.e. rationally, intelligently, successfully) in varying environments and under different circumstances. Now, the implementation and application of such theories of agency can serve two purposes, as we have reiterated time and again: the focus may be on the agent becoming a useful tool for solving certain problems or it may be a way of testing (and revising) the theory in compliance with which the agent was built. However, do these two purposes not also go hand in hand? When implementing an agent in compliance with a certain theory is that not a way of testing the theory irrespective of whether one actually intends the implementation to serve this purpose or whether one simply desires to build a useful application? Of course, the testing of the theory will only be valid for the particular domain to which the agent is applied and to the extent to which it stays true to the theory's assumptions and hypotheses. So, one has to take great care when interpreting the results. Still, most agent implementations also serve as tests for the applied theory and may provide valuable feedback in this respect. Therefore, the more agents one implements the more results and experience one is likely to gain both for the sake of improving the theory and for the sake of building good agent

applications. Experiment either way is paramount! Let us therefore rephrase our initial question: Can we expect an increase in PRS and Ψ-agent implementations, if we propose a hybrid system composed of elements from both architectures?

BDI theory has enjoyed great interest and many different implementations over the last two and a half decades, especially in the form of PRS-like systems. The sheer amount of BDI applications has helped the theory become known and tested, and has generated much valuable experience in the process. What were the likely facilitators of this development? BDI theory's concentration on the concepts of belief, desire, intention and plans — with a special focus on the latter two — makes it a relatively restricted theory in that regard (relative to Ψ-theory, for instance). What's more, since it does not make claims as to the concrete architecture of an agent, it leaves much room for interpretation and creativity. Above all, many of the traditional, and widely known and used structures of AI were reinterpreted in terms of this new theory. They were deemed fit for implementing agents with beliefs (symbolic state descriptions), desires (symbolic goal expressions) and intentions (primarily, plan structures). Arguably, these reasons in particular have lead to the quick transition from theory to agent development. The agents built according to BDI theory were, and still are, relatively simple ones for the most part (again, compared to Ψ), which do not yet adhere to every single principle of the theory. Especially the delicate interplay of the concepts in BDI which makes up the bulk of the theory is not wholly realized — the gap between theory and practice. Yet, by reducing the theory so far that agent development could get going albeit not modeling the theory in its entirety, and by appealing to folk psychology — recall section 3.3 — the way for a success of BDI agent applications in realistic environments was paved. It must be emphasized that a central element of the reduction was that of restricting oneself to predefined plans, i.e. a plan-library that is fixed. Without doubt, all of this has reflected back favourably onto the original theory, and it gives rise to some hope that its validity can come under scrutiny through practice.

The gap between the formalized theory and operational practice is one area of BDI agent research that needs attention. Another one concerns the question, whether or not reasoning processes according to BDI are indeed sufficient for a correct account of how human rational behaviour emerges. It is often argued that emotions are a relevant component of rational decision making and a prerequisite for autonomy (see e.g. Bernedo Schneider, 2004, pp.2-5 and pp.16-18). It does not seem completely nonsensical to presume that emotions sometimes influence profoundly such properties as commitment without being irrational. Scheutz and Sloman (2002) mention explicitly the benefit of non-deliberative mechanisms such as emotions for agents in domains that require real-time behaviour, i.e. which constrain the agent concerning the time and the memory it has for processing, the amount of fault tolerance or general energy levels, for instance. Interestingly, these are exactly those domains in which BDI agents have had their greatest success so far. This fact on its own promises fascinating new research questions and possibly new types of agents. We also saw the relevance of genuine motivation founded

on urges internal to the agent and how this, too, can increase an agent's autonomy (see also Dörner, 2003a). This is partly because the agent can provide itself with new goals rather than wait for goals to be given to it from some external source, and because it can take its decisions in line with its own needs. By making the nature of the modulations of perception, planning, decision making etc. depend on the configuration of personality parameters, many more variations in agent type can be tested. It adds to the flexibility already offered by the individual modulations themselves. Ψ-theory promises solutions for all of the aspects just mentioned, however, there is a severe lack of experience with Ψ implementations: the only ones available to date are those by Dörner himself and the MicroPsi architecture developed by Bach and his colleagues which we have mentioned.

In contrast to BDI, Ψ-theory has not profited from such great interest among agent developers and implementations remain rare. The primary reason for why there are not nearly as many implementations seems to be that pretty much the opposite holds true for every aspect we have just discussed with respect to BDI. First of all, Dörner's theory is directly realized in (or derivable from) an architecture that proposes structures and processes which could generate the forms of behaviour theorized about — it is a synthetic theory. Also, many of the key structures and processes that it proposes are quite novel or at least less well studied in artificial intelligence research than those which are used by common BDI realizations: this is especially true for the HyPercept process and the quad representation of schemata including its appendant processes of spreading activation, decay and reinforcement. Moreover, Dörner strives for some plausibility with respect to the neuronal correlate of those structures and processes, and the theory aims at integrating cognition, motivation and emotion. What's more, some of these processes, perception and learning in particular, are generally seen to be difficult and very time consuming which is a drawback for a lot of realistic scenarios — the main reason for their exclusion or at least massive reduction in many BDI systems. As a consequence of its general complexity and comprehensiveness, the relative inexperience with its foundational elements and the performance drawbacks to be expected, a complete implementation applied to realistic problem scenarios seems infeasible or even overkill, to say the least. In order to increase the pervasiveness of Ψ-agents, a substantial reduction of the architecture seems imperative.

5.5.2 The Nature of a Hybrid System

Now that we have motivated the development of a hybrid architecture, let us sketch what it should look like in the case of the PRS and Ψ. There are potentially many ways in which the two could join forces, but we will argue for one particular architecture that has the following general structure. At the heart of the architecture we will have the complete PRS as described in the second chapter. The system will be enhanced by introducing urges as sources of motivation, and modulations (emotions) on the basis of this motivation. Both the motivational and the emotional component will affect the core reasoning

system from outside without being a part of it. What are the arguments in favour of this architecture?[33]

Let us consider what decisions we have made in adopting this division and what their consequences are. The decision for the PRS as the 'cognitive core' has two essential components. Firstly, we adopt the PRS reasoning cycle and secondly, we adopt the symbolic form of representation for beliefs, desires and intentions alike. As for the reasoning cycle, we saw that the central components of it are essentially the same for Ψ: option generation is followed by decision making to determine intentions and this, in turn, is followed by means-ends reasoning. The primary difference is the fact that the agent's current motivation influences the processes of option generation and decision making in Ψ and that emotions modulate them all — but we are planning to integrate them.

The second component, namely representation, is a more difficult matter. Our options in this respect are constrained in important ways. Consider the following: In spite of the possibility of enhancing the PRS with processes like spreading activation — similar to the way ACT-R was enhanced by Anderson in its latest versions —, such enhancements would only serve to produce hybrid systems that, as Dörner et al. (2002, pp. 14-15) argue,

> "mingle different levels of description. (...) It is possible that one doubles reality. (...) If in psychological theories one intermingles statements about processes in neural networks, e.g. spreading activation, with the invocation of (...) if-then-statements, it is conceivable that an if-then-statement can be described as spreading activation in a neural network. Thus, one causes terminological confusion. One creates differences where there are none. This is why hybrid descriptions should be avoided".

This constraint is of course much harder for the development of a cognitive architecture that strives for psychological validity than for an agent architecture that is built for some other purpose. Still, if we take this concern seriously, which we will do, then one must make a definite choice between the two, i.e. essentially a choice between the symbolic representations used in the PRS and the subsymbolic descriptions used by Ψ. What is the best choice in light of the abovementioned motivations for our hybrid system? In the previous section we argued that perception and learning are two difficult and time-consuming processes. In Ψ the format of representation — hierarchical quad networks with spreading activation, decay and reinforcement — is a key facilitator both of perception and of learning and its structure at a particular point in time is shaped fundamentally by the two. In most successful BDI systems neither perception (not percepts!) nor learning play an important role, because the applications have strict demands for real-time behaviour. If we want to profit from the success in such problem domains we should consider not including complicated perception or learning processes, for the time being

[33] Notice that the proposed hybrid is not unlike the MicroPsi system by Bach (2003).

at least. Therefore, if we can do without them, we can also do without the representational format. Note that we are letting go of goal learning as a special case of learning. This, however, is a similar restriction to the common restriction of BDI systems to a prespecified plan library. So, since we do not want to implement Ψ in its entirety and since we do not want a hybrid symbolic-subsymbolic representation, it is reasonable to adopt the symbolic representation of the PRS. But what remains of Ψ? Have we done it undue harm?

Motivation and emotional modulation remain; and yes, we have harmed it, because we have reduced it substantially. The exclusion of the representational format proposed by Ψ-theory is not to be taken lightly! Still, it has retained two of its most vital components so that its general structure remains intact! Arguably, it is one of the few legitimate ways of reducing the theory in such a way that it becomes more feasible for implementation and, above all, for practical application without sacrificing it entirely. Tearing it apart, studying its parts in isolation and finally integrating the results is not the same as testing the theory as a whole — we refer the reader back to 4.1. Ψ-theory has many parts and these parts are highly interrelated. This is why the tests of Ψ-theory that have been conducted so as to verify or falsify its hypotheses used the complete architecture and were conducted as psychological experiments in which Ψ's compete with human subjects. Our hybrid system, however, will still be highly self-regulatory, it will still have urges and display emotions, it will still have veritable intentions. These essentials of Ψ-theory may now be studied in real-world applications where we can profit from the experience gained by BDI applications, while existing BDI agents enhanced by motivation and emotion are given a chance for increased flexibility and autonomy, and possibly even increased believability and realistic behaviour because of emotional aspects (see Bernedo Schneider, 2004, pp.16-18 for a short discussion).

5.5.3 The Realization of a Hybrid System

Finally, we turn to what the hybrid agent may look like in more detail. The basis of it will be the PRS system that we have described extensively in chapter 3. As for motivation, urge tanks might be simulated by internal variables that are maintained as beliefs in the database with Acts updating them periodically. Similarly we might have variables and Acts that represent and update the parameters of emotional modulation on the basis of the urge variables. While this method seems plausible in the sense that we would be reusing what is already there, it is highly problematic in another respect: we 'mingle' processes that should be separate. The changing states of the urges and the changing emotional situation are not reasoning processes, they are not cognitive, as it were. In particular, there would be intentions to update urges, the level of arousal or the selection threshold which would be awkward to say the least. Of course, motivation and emotion must affect the reasoning process and may even be conversely affected by it, but they should not be a part of it and they need not be part of it — in section 5.4.2.3 we already

warned of the abuse of meta-Acts. Therefore, it is advisable to relocate the regulation of urges and emotions to outside the reasoning cycle. New procedures will have to be included which handle the representations (variables) of the urge tanks and the modulation parameters. Which modulators can we include and what might they modulate?

The most fundamental modulators in Ψ-theory are *arousal*, *resolution level* and *selection threshold*, with the level of arousal determining the levels of the other two. The values of the modulators can be obtained provided that we find some way of measuring the success of the agent's actions necessary for the competence urge — this should be relatively straight forward — and some way of measuring the certainty — this is much more difficult. Dörner (cf. 1998, pp.352 and the following) points out that uncertainty events are indicators for the necessity of the elaboration and reconstruction of memory. Ultimately, they are indicators for the need for information and learning. The elements perceived and/or the procedures available for action are not sufficiently elaborate in order for successful behaviour. However, our hybrid cannot learn new Acts. Still, certainty could be measured as the number of relevant environmental facts in the database, i.e. if, for some time, there are only very few environmental facts in the database such that applicable Acts can be found, it could be seen as an indication that the agent needs more information — the environmental facts. This could result in the agent actively engage in behaviour to obtain more such facts, like change to another situation or request more information from its environment or other agents, for instance. Let us return to the modulations.

There are several candidate processes for modulation by, essentially, the resolution level and/or the selection threshold. First of all, the most general modulation is to influence the duration for which the agent stays in each of the phases depicted in figure 3 on page 30. The resolution level might also determine how many of an Act's preconditions must be matched for it to be considered an option, i.e. whether partial matches already lead to the Act being deemed appropriate in the current situation. Thirdly, the number of Acts intended per cycle and/or the probability of the current intention staying in control of conduct might vary with the selection threshold. Modulating the PRS opens up the possibility of developing even more different types of agents than just by the interactions of beliefs, desires and intentions alone. The flexibility and autonomy of agents is likely to increase since all processes have their source in internal motivations. The agent's decisions are based on and its behaviour regulated by its own internal state.

Here we must lay the discussion to rest. We have sketched motivations for and constraints on a hybrid system. We also provided food for thought as to a possible realization. However, closer inspection and more detailed considerations are beyond the scope of this thesis.

"All's well that ends well."
—William Shakespeare

6 Summary and Conclusions

This thesis set out to compare two approaches to explaining and modeling human goal-directed behaviour on the basis of the concept of intention: the theory of Belief-Desire-Intention (BDI) by Michael Bratman with its models based on the Procedural Reasoning System (PRS) on the one hand, and Ψ, theory and model, by Dietrich Dörner on the other hand. The models, taking shape as architectures for the construction of artificial agents, were chosen to be the focus of attention. A structural and procedural comparison was realized that shed some light on the differences and commonalities, not only as regards intentions, but also the concept of utility. Beyond this, the thesis showed in how far Ψ could be seen as a special form of BDI approach, and it sketched the motivation, the nature and the possible realization of a hybrid system. First, however, in order to clarify the terminological background and lay the foundations for an understanding of the principle issues of the comparative endeavour, the meaning of the central scientific concepts was established and some of their relations outlined.

As the first of the two approaches, BDI, a philosophical theory of practical reason in humans, and its agent architectural realization in the PRS, was introduced. The three mental attributes which the theory proposes were explained. Beliefs are seen as the, possibly incorrect or incomplete, information the agent has. Desires make up the motivational component that tend to lead the agent to goal-directed actions. Intentions, finally, which are closely related to plans, are viewed as the actual initiators of actions following the decision to pursue particular desires. They have the characteristic property of commitment — both to ends and means — which puts them in a place to control and stabilize conduct. BDI theory is concerned most with the question of which transitions between the three mental attributes allow for rational reasoning in humans, and when an agent should reconsider its commitment to individual intentions. Different types of commitment and different mental state interrelations make different types of BDI agents.

The PRS, it was shown, is a basic practical reasoning framework for constructing BDI agents. It provides a symbolic knowledge representation format. Beliefs are specified as facts in a database and desires as goal expressions. Furthermore, so-called Acts are stored in a library as prespecified plans (stored procedures) and, when instantiated as options and then chosen, form the core of intentions. The actual reasoning, i.e. the generation of options for action, their selection as intentions and their eventual execution is performed by a general interpreter whose default behaviour is extremely simple: The PRS, however, provides for the possibility of implementing meta-reasoning working from within the interpreter cycle and allowing agent developers to equip agents with more appropriate

and informed methods of choice or more finely-tuned commitment strategies. Finally, some of the realized applications of BDI agents were sketched and the observed benefits were named: Particularly, the intuitiveness of the concepts of belief, desire and intention, and the success of PRS-like implementations in real-time scenarios. Also, open issues like the lack of learning abilities and the difficulties of balancing the individual reasoning processes were noted.

Ψ, the second approach introduced, is both psychological theory and computational model. The theoretical assumptions it makes are directly translated into an agent architecture capable of exerting the action regulation and its resulting behaviour which the theory attempts to explain. It integrates cognitive, motivational and emotional components on the foundation of a novel form of subsymbolic representation, i.e. a neural network forming sensory and motor schemata, and control structures. The schemata, it was explained, are used to encode information about the agent's environment, concepts and procedures. They form the agent's memory which is constantly growing as a protocol thread of what the agent does and perceives, but at the same time disintegrating since elements that are rarely used are forgotten which, in turn, allows other, more relevant, structures to be more readily retrieved: The result is that the agent is able to learn. One fundamental foothold of Ψ-theory is the idea that humans, as autonomous agents, have a set of urges — for certainty, competence, affiliation etc. — which have to be kept in a state of equilibrium. These internal variables motivate the formation of intentions directed at attaining situations — goals — in which they are reequilibrated by so-called consumatory terminal actions that satisfy (or prevent) an urge. The association of an active (disequilibrated) urge with goal descriptions — called a motive — forms the core of every intention. It was shown how Ψ can acquire new goals by association as a result of the ability to learn and the ability to place value on situations if they lead to the satisfaction (or 'disatisfaction') of an urge.

The proposed mechanism of intention regulation at the heart of the theory is initiated when motives arise. Based on the intensity of the underlying active urges and the expected competence of attaining the associated goals given the current situation, Ψ chooses one current motive which it intends to pursue. Its goals, in conjunction with some goals from other motives, form the focus of attention. The subsequent phase of orientation then looks for these goals in the current situation and if one is found a cosumatory terminal action is attempted. Otherwise, automatisms are sought for in memory or plans are constructed that might bring the agent closer to a goal. Either way, an action is performed, if only as trial-and-error. All these processes are regulated by modulations that Dörner interprets as emotions. They determine how elaborate or how shallow processes are performed, and how readily the agent will change its current intention. Different types of agents basically differ in the amount of effect these modulations can have (personality parameters) and the manner in which urges tend to (dis)equilibrate — i.e. fast or slowly. It was finally explained that Ψ agents have up to now only been applied in psychological experiments, conducted by Dörner himself, and comparing Ψ's

performances to that of human subjects in simulated scenarios. Ψ-related work currently in progress aims at introducing language and self-modification abilities.

Following the introduction of the two approaches, the comparison was undertaken guided by the three questions originally posed. First of all, Ψ-theory was investigated in terms of whether or not it fulfils the basic requirements of a BDI approach put forward by Bratman, especially those for volitional and reasoning-centred commitment of intentions. Two central results followed from the discussion: Firstly, it is indeed possible to interpret Ψ as a BDI approach for it conforms to all the necessary criteria. Two criteria, however, are not entirely fulfilled: that commitment be made against the background of already committed intentions and that intentions help limit the options that need to be considered. This last criterium requires that an agent does not entertain options that are inconsistent with its intentions. BDI implementations often have explicit filtering algorithms for detecting such inconsistencies, in the literature on Ψ-theory, however, no such check is to be found. It was sketched though that the prerequisites for it are given by a process called 'confirmation and refutation' which is normally employed as part of the search for automatisms. Secondly, in Ψ, the identification of desires as the precursors of intentions poses a slight problem since the actual precursors of intentions for Ψ are the motives which may have a number of goals attached while desires, as they are commonly seen in BDI implementations, are single goal options. Therefore, motives can best be described as sets of desires; then all other criteria for desires apply.

The second part of the comparison looked into the similarities and difference between the PRS architecture and that of Ψ. The comparison was subdivided into two main parts, the first of which dealt with questions of representation and memory while the second was dedicated to a scrutiny of the practical reasoning processes underlying the generation of goal-directed behaviour in each approach. A fictitious scenario of a digital personal assistant was introduced to which the observations of this comparison were applied by example. The comparison itself provided the following main insights: Firstly, the representational format used is fundamentally different — the PRS uses symbolic structures while Ψ employs a form of subsymbolic neuronal network representation. Secondly, many features and abilities already incorporated in the Ψ architecture are not part of the PRS; they could, however, be integrated using the meta-Act mechanisms provided by the PRS. Above all, it is the disability of PRS to learn new concepts, operations or goals. Moreover, it is the related fact that it does not remember what it has done in the past, unlike Ψ which entertains a protocol memory.

Turning to goal-directed behaviour and focussing on deliberation, we looked into three fundamental steps on the way to attaining a goal: goal learning, option generation and decision making. As for goal learning, we saw that the PRS not only lacks the general ability to learn but that it cannot itself place value on situations, i.e. it neither provides a concrete utility function, which, as a general architecture it cannot do, nor does it constrain possible utility function. It is the obligation of the agent designer to in-

troduce means for determining preferences. The Ψ architecture, on the other hand, while not defining a utility function either, constrains such functions by demanding that they be dependent on urges which reflect the homeostatic tendency of a regulatory system. In option generation and decision making, the introduction of urges and their motivational aspects as done in Ψ show their greatest effect, because any options the agent entertains have to be motivated by the presence of an urge or the anticipation of it. Both the PRS and Ψ, however, also check that the options, and the intentions they may result in, are warranted by the external situation. The main difference between Ψ and typical BDI approaches in this respect is that the generation and selection of options by Ψ is not done before the background of what is already intended while BDI approaches explicitly focus on rational consistency, as mentioned. BDI agents normally add explicit admissibility checks between desires and intentions or intentions amongst themselves, i.e. checks under rationality constraints. These are not features of the PRS but, again, tailored meta-Acts will have to realize them. The same goes for possible commitment strategies.

The third and final part of the comparison was dedicated to motivating and outlining a possible hybrid system composed of parts from both architectures and theories. The main motivation was to make the complex Ψ architecture become more easily implemented and applied to realistic practical agent scenarios by trying to reduce it in some way. At the same time it was argued that BDI agents and their applications could benefit from an enhancement by motivational and modulatory aspects they currently do not offer. In a nutshell, Ψ seems to offer some interesting aspects which BDI agents might adopt, while BDI agents, the way they are usually implemented, have proved to be very useful in practical applications. The idea behind both these arguments was to increase the number of agents implemented such that more experience and insights could be gained regarding the core elements of each theory. With this motivation behind the endeavour to create a hybrid system, the nature that such a system might have was discussed. A central constraint was identified in that the two forms of knowledge representation — symbolic and subsymbolic — should not be mingled. Eventually, an architecture was proposed that has the complete PRS as its core with motivations and emotional modulations like those of Ψ-theory acting on the processes of the PRS. For the future one can hope for applications, based on the hybrid architecture proposed here, which also included ideas from Ψ-theory. It is a very promising theory and it would be interesting to study, for example, how motivation and emotion should and/or do affect the transitions of the three mental attributes in rational agents. This would also draw attention to the question of what rationality actually means in the two approaches and whether these notions are compatible.

Bibliography

Anderson, J. and Lebiere, C. (1998). *The Atomic Components of Thought*. Erlbaum, Mahwah, NJ.

Bach, J. (2003). The MicroPsi Agent Architecture. In *Proceeding of ICCM-5*, pages 15–20. Universitäts-Verlag Bamberg, Germany.

Bernd, H. and Hippchen, T. and Jüngst, K. and Strittmatter, P. (2000). Durcharbeiten von Begriffsstrukturdarstellungen in unterrichtlichen und computergestützten Lernumgebungen. In Mandl, H. and Fischer, F., editor, *Wissen sichtbar machen — Wissensmanagement mit Mappingtechniken*, chapter 2. Hogrefe, Göttingen, Germany.

Bernedo Schneider, G. (2004). Agenten und unsere Emotionen — Ein Vergleich von Dörners PSI-Theorie mit der Emotionstheorie von Ortony, Clore und Collins. Master's thesis, Universität Osnabrück, Germany.

Bradshaw, J. M. (1997). An Introduction to Software Agents. In Bradshaw, J. M., editor, *Software Agents*, chapter 1. AAAI Press / MIT Press, Cambridge, MA.

Braitenberg, V. (1993). *Vehikel. Experimente mit kybernetischen Wesen.* Rowohlt Taschenbuch, Germany.

Bratman, M. E. (1987). *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA.

Bratman, M. E., Israel, D. J., and Pollack, M. E. (1988). Plans and Resource-Bounded Practical Reasoning. *Computational Intelligence*, 4.

Braubach, L., Pokahr, A., Lamersdorf, W., and Moldt, D. (2004). Goal Representation for BDI Agent Systems. In Bordini, R. H., Dastani, M., Dix, J., and Fallah-Seghrouchni, A. E., editors, *Second International Workshop on Programming Multiagent Systems: Languages and Tools*, pages 9–20.

Broersen, J., Dastani, M., Hulstijn, J., Huang, Z., and der van Torre, L. (2001). The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In Müller, J. P., Andre, E., Sen, S., and Frasson, C., editors, *Proceedings of the Fifth International Conference on Autonomous Agents*, pages 9–16. ACM Press, New York, NY.

Carrier, M. and Machamer, P. K., editors (1997). *Mindscapes: Philosophy, Science, and the Mind*. Pittsburgh University Press, Pittsburgh, PA.

Cohen, P. R. and Levesque, H. J. (1990). Intention is Choice with Commitment. *Artificial Intelligence*, 42.

Dastani, M., Hulstijn, J., and van der Torre, L. (2003). How to Decide What to Do? Published on the Internet, URL: *citeseer.ist.psu.edu/663591.html*, Last visited: August 2005.

Detje, F. (1999). *Handeln erklären — Vergleich von Theorien menschlichen Handelns und Denkens*. Deutscher Universitäts Verlag, Wiesbaden, Germany.

d'Inverno, M., Georgeff, M. P., Kinny, D., Luck, M., and Wooldridge, M. (2004). The dMARS Architecture: A Specification of the Distributed Multi-Agent Reasoning System. *Journal of Autonomous Agents and Multi-Agent Systems*, 9(1-2):5–53.

d'Inverno, M., Kinny, D., Luck, M., and Wooldridge, M. (1998). A Formal Specification of dMARS. In Singh, M., Rao, A., and Wooldridge, M., editors, *Intelligent Agents IV: Proceedings of the Fourth International Workshop on Agent Theories, Architectures and Languages, Lecture Notes in Artificial Intelligence 1365*, pages 155–176. Springer-Verlag, Heidelberg, Germany.

Dretske, F. (1992). *Explaining Behaviour — Reasons in a World of Causes*. MIT Press, Cambridge, MA.

Dörner, D. (1976). *Problemlösen als Informationsverarbeitung*. Kohlhammer Verlag, Stuttgart, Germany.

Dörner, D. (1998). *Bauplan für eine Seele*. Rowohlt Taschenbuch Verlag, Reinbek, Germany.

Dörner, D. (2003a). Autonomie. In Christaller, T. and Wehner, J., editors, *Autonome Maschinen*. Westdeutscher Verlag.

Dörner, D. (2003b). The Mathematics of Emotion. In *Proceedings of ICCM-5, International Conference on Cognitive Modeling*, pages 75–80. Universitäts-Verlag Bamberg, Germany.

Dörner, D., Bartl, C., Detje, F., Gerdes, J., Halcour, D., and Schaub, H. (2002). *Die Mechanik des Seelenwagens*. Verlag Hans Huber, Bern, Switzerland.

Duden (1994). Duden — Das Große Fremdwörterbuch — Herkunft und Bedeutung der Fremdwörter.

Erickson, T. (1997). Designing Agents as if People Mattered. In Bradshaw, J. M., editor, *Software Agents*, chapter 5. AAAI Press / MIT Press, Cambridge, MA.

Franklin, S. and Graesser, F. (1996). Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages*, pages 21–35. Springer-Verlag, Heidelberg, Germany.

Georgeff, M., Pell, B., Pollack, M., Tambe, M., and Wooldridge, M. (1999). The Belief-Desire-Intention Model of Agency. In Müller, J., Singh, M. P., and Rao, A. S., editors, *Proceedings of the 5th International Workshop on Intelligent Agents V : Agent Theories, Architectures, and Languages (ATAL-98)*, volume 1555, pages 1–10. Springer-Verlag, Heidelberg, Germany.

Georgeff, M. P. and Lansky, A. L. (1987). Reactive Reasoning and Planning. In *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87)*, pages 677–682. AAAI Press, Seattle, WA.

Georgeff, M. P. and Rao, A. S. (1996). A Profile of the Australian Artificial Intelligence Institute. *IEEE Expert: Intelligent Systems and Their Applications*, 11(6):89–92.

Huhns, M. N. and Stephens, L. M. (2001). Multiagent Systems and Societies of Agents. In Weiss, G., editor, *Multiagent Systems — A Modern Approach to Distributed Artificial Intelligence*, chapter 2. MIT Press, Cambridge, MA.

Ingrand, F. F., Georgeff, M. P., and Rao, A. S. (1992). An Architecture for Real-Time Reasoning and System Control. *IEEE Expert*, 7(6):34–44.

JACK (2005). Agent Oriented Software (AOS) – JACK Intelligent Agents, software agent system. Published on the Internet, URL: *http://www.agent-software.com*, Last visited: August 2005.

Jacob, P. (Fall 2003). Intentionality. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Jadex (2005). Jadex BDI Agent System. Published on the Internet, URL: *http://vsis-www.informatik.uni-hamburg.de/projects/jadex/*, Last visited: August 2005.

Jennings, N. R., Sycara, K., and Wooldridge, R. (1998). A Roadmap to Agent Research and Development. *Autonomous Agents and Multi-Agent Systems*, 1(1).

Kay, A. (1990). User Interface: A Personal View. In Laurel, B., editor, *The Art of Human-Computer Interface Design*. Addison-Wesley, New York, NY.

Kokinov, B. N. (1994). The DUAL Cognitive Architecture: A Hybrid Multi-Agent Approach. In Cohn, A., editor, *Proceedings of the 11th European Conference on Artificial Intelligence (ECAI94)*, pages 203–207. John Wiley and Sons, Ltd., Chichester, England.

Laurel, B. (1997). Interface Agents: Metaphors with Character. In Bradshaw, J. M., editor, *Software Agents*, chapter 4. AAAI Press / MIT Press, Cambridge, MA.

Lenzmann, B. (1998). Benutzeradaptive und multimodale Interface-Agenten. Dissertation an der Technischen Fakultät der Universität Bielefeld, Dissertationen zur Künstlichen Intelligenz (DISKI).

Maes (1997). Agents that Reduce Work and Information Overload. In Bradshaw, J. M., editor, *Software Agents*, chapter 8. AAAI Press / MIT Press, Cambridge, MA.

Marcus, G. F. (2001). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. MIT Press, Cambridge, MA.

McIlroy, D., Smith, B., Heinze, C., and Turner, M. (1997). Air Defence Operational Analysis using the SWARMM Model. In *Proceedings of the Asia-Pacific Operations Research Symposium (APORS'97)*.

Minsky, M. (1986). *The Society of Mind*. Simon and Schuster, New York, NY.

Munroe, S. J., Luck, M., and d'Inverno, M. (2003a). Towards Motivation-Based Decisions for Evaluating Goals. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 1074–1075. IEEE Computer Society.

Munroe, S. J., Luck, M., and d'Inverno, M. (2003b). Towards Motivation-Based Decisions for Worth Goals. In *Proceedings of the 3rd International/Central and Eastern European Conference on Multi-Agent Systems*, pages 17–28. Springer-Verlag, Heidelberg, Germany.

Negroponte, N. (1997). Agents: From Direct Manipulation to Delegation. In Bradshaw, J. M., editor, *Software Agents*, chapter 3. AAAI Press / MIT Press, Cambridge, MA.

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

Norling, E. (2003). Capturing the Quake Player: Using a BDI Agent to Model Human Behaviour. In Rosenschein, J. S., Sandholm, T., Wooldridge, M., and Yokoo, M., editors, *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 1080–1081. IEEE Computer Society.

Norling, E. and Sonenberg, L. (2004). Creating Interactive Characters with BDI Agents. In *Australian Workshop on Interactive Entertainment IE2004*, pages 69–77. Creativity & Cognition Studios Press, Sydney, Australia.

Padmanabhan Nair, V. (2003). *On Extending BDI Logics*. PhD thesis, Griffith University, Gold Coast Campus, Queensland, Australia.

Pauk, A. (1997). Technologie und Anwendung Intelligenter Agenten als Mittler in Elektronischen Märkten. Diplomarbeit in der Fachgruppe Informationswissenschaft der Fakultät Verwaltungswissenschaft an der Universität Konstanz.

Paulussen, T. O., Awizen, M., and Gerstacker, J. (2001). MedPAge — Funktionsbereichsübergreifende Planung, Steuerung und Koordination von Krankenhausprozessen. Extended Abstract für das 4. Kolloquium des DFG-Schwerpunktprogramms "Intelligente Softwareageten und betriebswirtschaftliche Anwendungsszenarien".

Pfeifer, R. (2003). Körper, Intelligenz, Autonomie. In Christaller, T. and Wehner, J., editors, *Autonome Maschinen*. Westdeutscher Verlag, Wiesbaden, Germany.

PRS Manual (2001). Procedural Reasoning System User's Guide — A Manual for Version 2.0. Published on the Internet, URL: *http://www.ai.sri.com/ prs/prs-manual.pdf*, Last

visited: August 2005. Version of March 13, 2001 — Artificial Intelligence Center, SRI International, Menlo Park, CA.

PSI-Glossar (2005). Glossar der PSI-Theorie. Published on the Internet, URL: *www.uni-bamberg.de/ppp/insttheopsy/dokumente/DetjeKuenzelSchaub_Glossar_der_PSI-Theorie.pdf*, Last visited: August 2005.

Rao, A. S. and Georgeff, M. P. (1991). Deliberation and its Role in the Formation of Intentions. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence (UAI'91)*, pages 300–307. Morgan Kaufmann Publishers, San Mateo, CA.

Rao, A. S. and Georgeff, M. P. (1993). Intentions and Rational Commitment. Technical Note 8, Australian AI Institute.

Rao, A. S. and Georgeff, M. P. (1995). BDI-Agents: From Theory to Practice. In *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)*, pages 312–319. AAAI Press, Menlo Park, CA.

Rübenstrunk, G. (1998). Emotionale Computer — Computermodelle von Emotionen und ihre Bedeutung für die emotionspsychologische Forschung. Published on the Internet, URL: *http://www.ruebenstrunk.de/emeocomp/INHALT.HTM*, Last visited: August 2005.

Russell, S. and Norvig, P. (1995). *Artificial Intelligence — A Modern Approach*. Prentice Hall, Upper Saddle River, NJ.

Schaub, H. (1993). *Modellierung der Handlungsorganisation*. Verlag Hans Huber, Bern, Switzerland.

Scheutz, M. and Sloman, A. (2002). Agents With or Without Emotions? In *Proceedings of FLAIRS 02*, pages 89–94. AAAI Press, Menlo Park, CA.

Schut, M. and Wooldridge, M. (2001). Principles of Intention Reconsideration. In *Proceedings of the 5th International Conference on Autonomous Agents*, pages 340–347. ACM Press, New York, NY.

Singh, M. P., Rao, A. S., and Georgeff, M. P. (2001). Formal Methods in DAI: Logic-Based Representation and Reasoning. In Weiss, G., editor, *Multiagent Systems — A Modern Approach to Distributed Artificial Intelligence*, chapter 8. MIT Press, Cambridge, MA.

Sloman, A. (2003). The Cognition and Affect Project: Architectures, Architecture-Schemas, And The New Science of Mind — Incomplete Draft. Published on the Internet, URL: *http://www.cs.bham.ac.uk/research/cogaff/sloman-cogaff-03.pdf*, Last visited: August 2005.

Toda, M. (1982). *Man, Robot, and Society*. Martinus Nijhoff Publishing, The Hague, Netherlands.

UM and JAM (2005). I.R.S. Intelligent Reasoning Systems. Published on the Internet, URL: *http://marcush.net/IRS/*, Last visited: August 2005.

Urchs, M. (2002). *Maschine Körper Geist — Eine Einführung in die Kognitionswissenschaft*. Vittorio Klostermann, Frankfurt am Main, Germany.

Walczak, A. (2005). Planung und der BDI-Ansatz in verteilten Anwendungen. Master's thesis, Universität Hamburg, Germany.

Weiss, G., editor (2001). *Multiagent Systems — A Modern Approach to Distributed Artificial Intelligence*. MIT Press, Cambridge, MA.

Wilson, R. A. and Keil, F. C., editors (2001). *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press, Cambridge, MA.

Winikoff, M., Padgham, L., Harland, J., and Thangarajah, J. (2002). Declarative & Procedural Goals in Intelligent Agent Systems. In *Proceedings of the Eighth International Conference on Principles of Knowledge Representation and Reasoning (KR2002)*, pages 470–481.

Wobcke, W. R. (2002a). Intention and Rationality for PRS-like Agents. In McKay, B. and Slaney, J., editors, *AI 2002: Advances in Artificial Intelligence*, pages 167–178. Springer-Verlag, Berlin, Germany.

Wobcke, W. R. (2002b). Modelling PRS-like Agents' Mental States. In Ishizuka, M. and Sattar, A., editors, *PRICAI 2002: Trends in Artificial Intelligence*, pages 138–147. Springer-Verlag, Heidelberg, Germany.

Wooldridge, M. (2000). *Reasoning about Rational Agents*. MIT Press, Cambridge, MA.

Wooldridge, M. (2001). Intelligent Agents. In Weiss, G., editor, *Multiagent Systems — A Modern Approach to Distributed Artificial Intelligence*, chapter 1. MIT Press, Cambridge, MA.

Wooldridge, M. and Ciancarini, P. (2001). Agent-Oriented Software Engineering: The State of the Art. In Wooldridge, M. and Ciancarini, P., editors, *Agent-Oriented Software Engineering*, chapter 1. Springer.

Wooldridge, M. and Jennings, Nicholas, R. (1995). Intelligent Agents: Theory and Practice. *Knowledge Engineering Review*, 10(2).

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst, noch nicht anderweitig für Prüfungszwecke vorgelegt, keine anderen als die angegebenen Quellen oder Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe.

(Michael E. Elbers)

Bonn, den 5. August 2005