The social cognitive actor

Helmhout, M.

CHAPTER 2

Multi-Agent Systems

THE first chapter has discussed the issue of how cognitive science can contribute to social sciences in order to explain interactive social behaviour. In chapter 5, we will develop a new cognitive agent-based computational social simulation model (RBot) in an attempt to satisfy the need for new theories and models. In this chapter, we will discuss Multi-Agent Systems (MAS) as one of the theories for the development of such a new model.

Multi-Agent Systems (MAS) is the name for a new approach of designing, analysing and implementing complex adaptive software systems (Jennings, Sycara, & Wooldridge, 1998). This approach has emerged out of a history of theory and applications in the area of (social) simulation models.

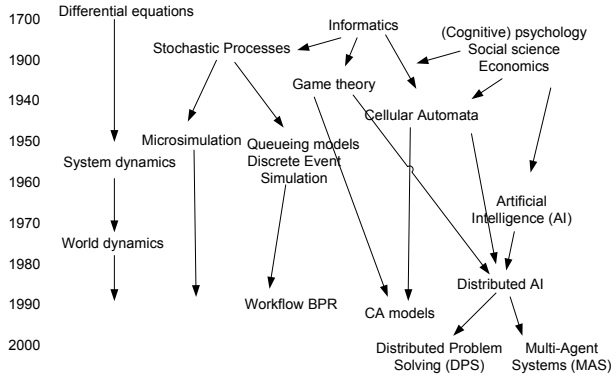

Figure 2.1: History of contemporary approaches of (social) simulation theories and techniques (adapted from Troitzsch, 1997; Gilbert & Troitzsch, 1999).

Figure 2.1 shows a selection of simulation theories and techniques that have been developed over the years and many of them have been applied in the field of MAS and social simulations. The early development of simulation techniques at the left side of the figure are primarily based on mathematics. Moving to the right, there are simulations and theories that exist because of the introduction of the computer and at the bottom right, we can see the influence of Internet on the upcoming popularity of Distributed AI and Multi-Agent Systems.

Mathematical theories, e.g. differential equations and stochastic processes, require specific skills for describing deterministic and stochastic processes in an attempt to predict or describe phenomena such as the dynamics of populations (Volterra, 1926). Game theory, discussed later on in this chapter, is a form of applied mathematics to understand economic behaviour based on interactions or games between rational agents who try to maximise their profits. Game theory still has great influence in describing outcome strategies in simulations.

In the 1960s, the first computer simulation models were developed with help of queueing models, discrete-event simulations (Birthwistle, Dahl, Myhrhaug, & Nygard, 1973) and system dynamics (Forrester, 1991). Around the same time, micro-simulation or micro-analytical simulation models (MSMs) emerged as a simulation technique.

> [MSMs have] been used to predict individual and group effects of aggregate political measures which often apply differently to different persons... [They] consist of at least two levels: the level of individuals or households (or the level of enterprises) and the aggregate level (for example, the population or national economy level). (Gilbert & Troitzsch, 1999, pp. 53–55)

Another approach that emerged was cellular automata whose behaviour emerges from properties of local interactions. The basic model of a cellular automata is a two-dimensional grid of cells—the structure of squared paper—in which a cell reacts only with its neighbouring cells. The simulation uses a time-step generator and at every time-step, every cell reacts to its neighbours according to a set of rules. In this way a stepwise reaction will flow through the grid and influences the overall behaviour of the system.

> To summarise, cellular automata model a world in which space is represented as a uniform grid, time advances by steps, and the 'laws' of the world are represented by a uniform set of rules which compute each cell's state from its own previous state and those of its close neighbours. (Gilbert & Troitzsch, 1999, p. 122)

In the 1980s, artificial intelligence, a separate field concerned with intelligence of the individual, became interested in the interaction and distribution of intelligence, known as Distributed Artificial Intelligence (DAI). Ferber (1999, p. 24) gives a short description about the history of the roots of DAI that today has resulted into Multi-Agent Systems. DAI started around the eighties with the model of the blackboard system (Erman, Hayes-Roth, Lesser, & Reddy, 1980); a model consisting of Knowledge Sources (KSs) organised in a star-topology with

in its core the 'blackboard'—a place where the KSs can share their knowledge—and which is managed by a separate control device that has the task to prevent and coordinate conflicts of access between these KSs. Around the same time another type of control system (cf. Lenat & Brown, 1984; Lenat, 1975) came up that solved problems with help of a community of specialists. Hewitt (1977) described control structures in which he did not consider processes as a sequence of choices, but "he tended to think in terms of distributed systems considering control structures as patterns of message passing between active entities called *actors*" (Ferber, 1999, p. 25) [1].

The roots of DAI are the foundations of today's two approaches in DAI: Distributed Problem Solving (DPS) and Multi-Agent Systems (MAS) (Bond & Gasser, 1988). DPS takes a pure *engineering* approach to distributed systems, i.e. it is concerned with "how to build functioning, automated, coordinated problem solvers for specific applications (Bond & Gasser, 1988, p. 4)" (as cited by Van den Broek, 2001, p. 21). The approach taken by DPS is a top-down approach in which tasks are decomposed into smaller tasks appointed to specialists.

> A DPS is... a top-down designed system, since agents are designed to conform to the problem-solving requirements specified at the top. Within this top-down task decomposition approach, the individual components are considered to be of *secondary importance* to the need of the overall system. The agents themselves have *limited autonomy* because their role in solving the overall problem is usually designed-in, with coordination rules included. (emphasis added: Van den Broek, 2001, p. 22)

On the other hand, a MAS is constructed taking the bottom-up approach. The agent itself is now the centre of attention and not the system as a whole. In a MAS, the agent has more autonomy, and control is delegated to the agent, i.e. the overall outcome of solving the problem is not determined by a strictly top-down organised control system, as in the case of DPS, but by (social) interactions between independent and rational agents making decisions that satisfy their own (be it social) goals.

[1] There is not a clear definition about or difference between the terms 'actor' and 'agent'. For instance, Hewitt gives the following description of the actor: "The actor metaphor for problem solving is a large human scientific society: each actor is a scientist. Each has her own duties, specialties and contracts. Control is decentralized among actors. Communication is highly stylized and formal using messages that are sent to individual actors" (Hewitt, 1977, p. 350). Apart from the description, he gives the following remark: *"The reader should keep in mind that within the actor model of computation there is no way to decompose an actor into its parts. An actor is defined by its behavior; not by its physical representation"* (Hewitt, 1977, p. 327).
In this dissertation, we will use the metaphor actor in a different way; as an entity that can be decomposed into (cognitive) elements that have their own behaviour. Besides that, we will also adopt the term 'agent', because of its wide adoption in the field of distributed systems. In section 2.1, we will explain the notion of agency that describes the term agent in terms of the weak and the strong agent. Although the distinction between 'actor' and 'agent' is not that sharp, and can often be synonymous, in this dissertation we try to distinguish both by referring to the term 'actor' as a being that is more autonomous and is a more general concept than the 'agent'. The agent often refers to a software entity that serves the person who designed the agent, e.g. an Internet search-agent, insurance agent, etc.

In this dissertation we adopt MAS as theory and methodology for simulating organisations for the following reason:

> ...[M]ultiagent systems applied within the natural systems approach[2] provide the required methodological position in simulating organisations. The multiagent system approach exploits the full potential of the DAI paradigm because of its emphasis on the emergence of a system level and a strong notion of [agency[3]] resulting in higher autonomy levels in agents. In this way, the agents become more socio-realistic in that they might compete and disagree with other agents, and generally act in their own best interest. The result is that coordination among the agents is necessary in order to achieve coherent collective behavior. Hence, multiagent models perceived in this way are bottom-up models of coordinated problem solving. (Van den Broek, 2001, p. 24)

In other words, MAS can supply us with the means to study what the aspects of actors are that plausibly explain interactive (social) behaviour (see chapter 1; research question 1). Hence, MAS is concerned with the behaviour of a collection of distributed autonomous (heterogeneous) agents aiming at solving a given problem. The characteristics of a generic MAS are (Jennings et al., 1998):

- Each agent has incomplete information, or capabilities for solving the problem, thus each agent has a limited viewpoint;
- There is no global system control;
- Data is decentralised; and
- Computation is asynchronous

Inferred from these characteristics, the main component studied in Multi-Agent Systems is the autonomous *agent* and its associated behaviour in an environment.

The approach taken in this chapter is to explain MAS starting with the individual agent and subsequently going towards a system of (cognitive) social agents that considers relations, coordination and interaction between agents. The purpose of the chapter is not to provide a detailed review of the field of MAS, but to accustom the reader with the paradigm MAS[4].

This chapter is structured as follows. First, in section 2.1, the notion of agency is discussed to clarify the term agent. The term agent is applied in many research fields, e.g. biology, economics, cognitive science, social science and so on, and therefore has become somewhat blurred. In section 2.2, we approach the agent as a human-like entity in which components, e.g. perception,

[2] "A natural systems approach... stud[ies] the strategies and representations that people use to coordinate their activities, in much the same way the cognitive scientists investigate individual cognition in people" (Van den Broek, 2001, p. 23).

[3] The strong notion of agency will be explained in the next section.

[4] See for additional information Bond and Gasser (1988), Durfee, Lesser, and Corkill (1992), Jennings et al. (1998), Sawyer (2003), Sycara (1998) or Wooldridge (2002).

cognition, action and so on, form the basis for constructing an agent architecture. Section 2.3 describes two agent typologies: the cognitive agent commonly used in AI and the social agent. Whereas the cognitive approach focuses on the internal working of the agent, the social approach studies how (socially) well an agent is embedded in its social-cultural environment. The environment plays an important role in the design of an agent; it, in part, defines how many degrees of freedom an agent possibly can have. In section 2.4, the distinction between a physical, communication, social and task environment is described. An environment allows agents to interact with other agents and at the same time, scarcity of space, time and objects in an environment can create conflicts between agents. In section 2.5, several types of coordination mechanisms are discussed that enable conflicting agent to negotiate and cooperate. Many situations in our daily life are solved with the help of coordination mechanisms. For instance, traffic lights function because we understand the meaning of the coordination mechanism. Such understanding solves problems that otherwise would occur when several cars want to occupy the same space (crossroad) at the same time. Section 2.6 discusses the applications that are implemented in the field of MAS. They vary from simple desktop applications for consumers, towards complicated distributed real-time systems in the industry. Applications of MAS are—especially on the Internet—widespread and have proven themselves a worthy competitor to alternative systems, i.e. due to the Internet, MAS applications are more wanted than ever before and can be the answer to many of today's coordination problems. Finally, section 2.7 closes the chapter with a discussion about what kind of agent and theory is important for our research, i.e. the type of agent will be selected from a variation of agents that differ in the way they are constructed, from simple to very complex. The complexity of such a construction depends on the philosophical ideas behind the design of the agent and its environment, and the amount of complexity we need in order to answer the research questions.

2.1 Agency

Understanding the field of MAS starts with the understanding of the agent as the main component of the system. In different fields, researchers apply the term agent differently and even within MAS, there are different understandings of the term agent, i.e. the context of application is necessary to understand the meaning of the term agent. Many definitions and classifications of agents have been proposed (Franklin & Graesser, 1996). In this dissertation we first introduce a general definition of an agent: "An agent is a computer system, situated in some environment, that is capable of flexible autonomous action in order to meet its design objectives" (Wooldridge, 2002, p. 15). Starting from this point, Wooldridge and Jennings (1995) distinguish two notions of the term agent: the first is the weak notion and the second the strong notion of agency.

2.1.1 Weak notion of agency

The weak notion is a general way to describe the term agent as a hardware or software-based computer system with the following properties:

Autonomy: agents operate without direct intervention of humans or others, and have some kind of control over their actions and internal state (Castelfranchi, 1995).

Social ability: agents interact with other agents (and possibly humans) via some kind of agent-communication language (Genesereth & Ketchpel, 1994).

Reactivity: agents perceive their environment and respond in a timely fashion to changes that occur in it.

Pro-activeness: agents do not simply act in response to their environment, they are able to exhibit goal-directed behaviour by taking the initiative.

The weak notion is applied in situations where behaviours, in contrast to the strong notion, are described at a high level of abstraction.

2.1.2 Strong notion of agency

In the field of AI, the notion of agency is stronger and apart from the characteristics of the weak notion, agents have additional properties, such as mental attitudes—knowledge, beliefs and so on—or representations (Jorna, 2002), i.e. an agent is "...an entity whose state is viewed as consisting of mental components such as beliefs, capabilities, choices, and commitments. These components are defined in a precise fashion,..." (Shoham, 1993, p. 52). Strong agents are often defined in more detail and can be described at the intentional or the functional level of description. Dennett (1987, 1978) introduced a distinction in various levels of description. He discerned a physical, a functional and an intentional level (or stance) of description.

> The physical stance explains behavior in terms of physical properties of the states and the behavior of the system under concern. For its proper functioning the human organism requires a complex interaction between its parts and with the external world... The second level of explanation takes the point of view of the functional design of a system. The behavior of a system is conceived of as the result of the interaction between a number of functional components or processes. The physical structure (architecture) of the system is not explicitly taken into account, although it may impose constraints on the behavior of the system. In the third place Dennett distinguishes the intentional stance. Complex behavior that is adapted to the prevailing circumstances, according to some criterion of optimality is said to be rational or intelligent. A behaving system to which we can successfully attribute rationality or intelligence qualifies as an intentional system. (Gazendam & Jorna, 2002, pp. 12–13)

The intentional level and functional level are considered as appropriate for applications in agent theory to ascribe mental attitudes and cognitive mechanisms. Two types of strong agents can be distinguished. The first type is the reasoning agent that is defined at the intentional level and whose implementation is based on logics and the practical reasoning theory (Bratman, 1987). The other agent is an agent with a cognitive architecture, or a symbol system and is defined at both, the intentional and the functional level. The former type of agent is used in primarily describing behaviour of agents, and the latter is applied when more details have to be known about the internal functioning (implicit behaviour) of the agent.

In order to better describe the strong agent, Newell (1982) developed a systematic description of an agent's behaviours known as the computer system level paradigm. The computer system level paradigm describes five levels of description, the device, the circuit, the register transfer, the symbol and the knowledge level. We are mainly interested in the knowledge level and the symbol level.

Newell applied the knowledge level to describe problems and the symbol level to implement and solve the problems. At the knowledge level, a problem can be described and solved by determining the goals of an agent and the actions that have to be performed to achieve these goals (Kalenka, 2001). At the symbol level, a physical symbol system (Newell, 1980) is applied for implementation and is necessary for exhibiting intelligent action.

The five levels of the computer system level paradigm can be connected to different timescales or bands (Newell, 1990). Whereas neurons may fire in 1 ms, and cognitive reasoning takes 100ms till 15 seconds, higher rational actions may take minutes or hours. Organisational and social behaviour can take up minutes, days, months, or even more than years.

Besides that, Newell (1982) introduced the behavioural law, the Principle of Rationality, which is applicable to the Knowledge Level: "If an agent has knowledge that one of its actions will lead to one of its goals, then the agent will select that action." (p. 102). However, Newell (1982)'s behavioural law only takes into account the individual and its desires, beliefs and goals.

As an addition for taking into account the society, Jennings and Campos (1997) defined a new system level above the knowledge level—the social level—in which the whole MAS or society is represented and consists of socially responsible agents. They define agents as being autonomous and balancing their individual needs with those of the overall system (Kalenka, 2001). According to Jennings and Campos (1997), the shortcoming of the behavioural law, i.e. the Principle of Rationality, is that it has meaning for a single agent and cannot be applied to a system of multiple agents. Therefore, Jennings and Campos (1997) extended the definition with a new principle that includes society; the Principle of Social Rationality: "If a member of a responsible society can perform an action whose joint benefit is greater than its joint loss, then it may select that action." (p. 16). This very strong notion not only assumes that the individual is self-interested, but also socially responsible for its actions.

In many Multi-Agent Systems agents are relatively weak, e.g. the well known predator-prey simulations based on the mathematical model of Lotka

and Volterra (1926). MAS agents are often constructed at a level that shows no internal mechanisms and therefore can only describe behaviour based at that (the intentional) level.

On the other hand, cognitive science tries to explore the functional level—see SOAR (Newell, 1990) and ACT-R (Anderson & Lebiere, 1998)—and even tries to touch the physical stance (connectionism). However, the weak spot of cognitive science (cognitive psychology) is its focus on the detailed individual organisation of the agent.

Therefore, the combination of cognitive science and MAS can create new opportunities for explaining cognitive and social behaviour varying from the individual level (functional and rational) to the social level. The strong notion of agency is necessary to explain behaviours at the intentional level (with help of the functional level) in a cognitive plausible way, i.e. an agent should be designed with the right grain size in order to have the predictive and explanatory power necessary to explain the phenomena studied. Apart from describing individual mental attitudes, social (and possible cultural) attitudes (Jennings & Campos, 1997) should be cognitively embedded as well in order to explain why an agent sometimes is obliged or prohibited to take actions that incur personal loss, but societal benefit.

2.2 Agent components and architecture

The notion of agency implies that there are different types of agents or architectures that describe how an agent (architecture) could be designed. The most common method used to discern agents is to define components and requirements that assign functions. The construction of an agent architecture based on these components determines the type of agent (cf. Conte & Castelfranchi, 1995b; Russell & Norvig, 2003; Van den Broek, 2001; Verhagen, 2000; Wooldridge, 2002).

Three main (cognitive) components for constructing the mind of the agent are often distinguished: perception, action and cognition. The peripheral components, i.e. perception and action, and the interaction between those components are discussed first.

2.2.1 Perception and attention

Perception is one of the components necessary to receive input signals from an environment, especially when the agent depends on the environment and the environment is constantly changing. Ferber (1999) distinguishes two types of perception, *passive* and *active* perception. With passive perception, the input signals are registered and classified accordingly. Active perception is a complex mechanism that combines and compares sensory data with data based on expectations, goals and experiences that are stored in the cognitive system. Active perception also can give specific instructions, based on classification and affordance, for tracking changes in the environment, e.g. which objects are able to move. In a complex environment, the agent should have some mechanisms that filter out information making it possible for the agent to give attention only to

senses that can be assigned as relevant input. Such an attention mechanism can be hard-coded in the design of an agent, by for instance giving it limitations for frequencies for hearing, or a certain range of colours in the spectrum of light.

2.2.2 Action

Actions are elements or means that enable the agent to change states. States can be situated and designed in the agent self, the environment or perceived and attributed to other agents. The number of actions is restricted by the agents limited number of known actions—cognitive limitation—and the allowance for these actions by the environment—physical limitation; e.g. the agent decides to throw a stone, but no stone is available. Van den Broek (2001) acknowledges this and states that the environment provides the agent a context in which actions have meaning and can be performed or not. Action, therefore, is a relative notion—known as *situated action* or *situated cognition*.

2.2.3 Interaction between perception and action

Susan Hurley (2001) discusses several approaches to the interaction of perception and action: the traditional view, the behavioural view, the ecological view, and her own two-level interdependence view. She distinguishes the subconscious sub-personal level from the conscious personal level. At the sub-personal level, there are sensory inputs and motor outputs. At the personal level, there are perceptual content (perceptions) and intentional content (intentions).

She explains that in the *traditional view* of the relation between perception and action, there is a vertical separation between perception, cognition and action. The mind passively receives sensory input from its environment, structures that input into perceptual content, processes this content in the cognitive system creating intentional content, and sends this intentional content to the motor system giving motor output. This is more or less a *linear* flow from environment through sensory system, cognition, and motor system and finally out to the environment again. Furthermore, sensory system and motor system are independent units that can have effects on each other through causal chains: the sensory system has effect on the motor system via the link perceptual units—cognitive processing—intentional units, while the motor system causes the position of the body or the environment to change, an effect that can be picked up by the perceptual system. One could say that this is an *instrumental* connection: "perception is a means to action and action is a means to perception" (Hurley, 2001, p. 12).

This traditional view on perception and action has received critique from two directions. First, instead of assuming a linear view, perception and action are seen to be more closely linked by dynamic feedback loops. Secondly, instead of a mere instrumental connection between perception and action, dynamical system approaches to the mind aim to show how cognition emerges as a function of an underlying dynamical system in which action and perception are constitutive parts. Hurley sees *behaviourism* as an approach that accepts the

linear view and rejects the instrumental view. It sees action and perception as constitutive parts of one functional unit.

Hurley places the *ecological approach* to perception and action as one that rejects the linear view but accepts the instrumental view. In this respect, it could be said that she does not understand Gibson's theory and wrongly places him in the corner of advocates of an instrumental link between perception and action:

> To place ecological views of the relations between perception and action into a box designating those relations as merely instrumental belies a fundamental misconstrual of Gibsonian psychology ...Most importantly, however, it is clear that ecological psychologists do not regard the relation between perception and action as merely instrumental: the very things they propose animals perceive are *affordances*, or opportunities for action. Affordances are neither features of the external world nor internal representations; they are neither objects nor mental phenomena. Instead, they are the relational properties between the available resources in an environment and an organism's abilities to utilize those resources. Therefore, if it is affordances that an organism perceives, as Gibsonians claim, then perception inherently involves action because the [...] very kinds of things perceived are relative to the actions an organism can perform. We stress that this is, moreover, a constitutive (i.e. non-instrumental) link between perception and action: an organism's perceptions are in part constituted by its behavioral repertoire. (Chemero & Cordeiro, 2006)

In contrast to the traditional view, Hurley explains her two-level interdependence view. She states that the relationship between perception and action is not linear, but consists out of numerous dynamic feedback loops. Furthermore, she argues that there is not a mere instrumental link between perception and action, but that perception and action are *constitutively linked*, that is, they are inseparable parts of a larger dynamic system: "The contents of perceptions and intentions are each constituted by processes involving both inputs and outputs. Both are functions (albeit different ones) of the same underlying relations between inputs and outputs" (Frankish, 2006).

Change in intentions can lead to a change in perceptions. Hurley argues that, therefore, there must be a constitutive link between perception and action and not a mere instrumental link. So, perceptual content is constitutively linked to sensory input processes and motor output processes, as intentional content is. Perceptual content and intentional content are thus also constitutively linked as parts of a larger dynamical system. A picture emerges of a sub-personal level and a personal level that are interdependent parts of a larger dynamical system: the *two-level interdependence view* (Hurley, 2001).

Constitutive links also exists between processes of sensory input and processes of motor output, leading to horizontally layered modules. These modules are directly coupled to the actor's environment:

> One way to think of these is in terms of layer upon layer of

> content-specific networks. Each layer or horizontal module is dynamic, extending from input through output and back to input in various feedback loops. Layers are dedicated to particular kinds of tasks. One network, for example, may govern spatial perception and the orientation of action (the so-called 'where' system). Another may govern food recognition and acquisition-type behavior (part of the so-called 'what' system). Another may govern predator recognition and fleeing-type behavior (another part of the 'what' system). Another may govern some of the variety of imitative responses to the observed behavior of others, and so on. Evolution and/or development can be seen as selecting for each layer. Each subpersonal layer is a complete input-output-input loop, essentially continuous and dynamic, involving external as well as internal feedback. Thus, not only are sensory and motor processes coupled, but the neural network is directly coupled to the creature's environment; horizontal modules are essentially 'situated'. (Hurley, 2001, p. 6)

With respect to this coupling of organism and environment, Hurley can be seen as an ultra-externalist by suggesting that perception content and intention content are constitutively linked to subpersonal processes that extend into the environment. (Frankish, 2006).

A critique on Hurley (2001)'s work states that she uses dynamical systems theory in a wrong way:

> Drawing on this dynamical cognitive science, Hurley says that a sub-person is a 'dynamic singularity'. These ideas are faulty because if sub-persons are well understood in terms of dynamical systems theory, then the ideas of input and output ... don't apply to them. Dynamic singularities include more than just the organism. There are often loops that go out into the external environment (as well as internal loops). These loops are part of the (sub-personal) dynamic singularity. But if the singularity is the whole dynamic system, the traditional ideas of inputs and outputs don't make sense. Inputs from what? Outputs to what? (Chemero & Cordeiro, 2006)

Hurley (2001) has made an interesting contribution with her two-level interdependence view. Especially the description of horizontally layered modules that consist of sensory input processes and motor output processes, that are based on dynamic feedback loops and that extend into the environment are interesting. These horizontally layered modules are constitutively linked to perceptual content and intentional content. We see that Hurley (2001)'s theory is more compatible with Gibson's theory and organisational semiotics than she thinks: her horizontally layered modules could correspond to the Gibson's physical affordances, while larger modules functionally dependent on these physical affordances could be the social constructs of organisational semiotics (see chapter 3). Furthermore, the extension of the horizontally layered modules into the environment can be seen as compatible with the theory of the semiotic Umwelt in semiotics (see chapter 3).

Although interaction between perception and action is interesting enough on its own, the focus of this dissertation is not on the complexity of perception or action but more on the social situatedness of the cognition. In this dissertation, a one-way circular system such as the traditional view can be applied for simplicity, i.e. complex perception-action systems go beyond the aim of this dissertation.

2.2.4 Cognition

Perception and action are enough when a system only has to control and adapt to the current situation. However, when an agent has the desire to improve over time with help of experience, then the agent requires intelligence or a cognitive system and a memory of representations and experiences from the past in order to be able to learn from previous actions. For now, we will postpone a detailed discussion about cognition and come back to it in chapter 4. In this section, we will describe the functional components comprising a cognitive system that requires at least memory, learning and goal-directed behaviour.

2.2.4.1 Memory

Assuming that agents have feedback loops for comparing actions with past actions and perceptions with past perceptions, then the agent should be able to maintain state or have a memory with experiences of past states. For instance, many economic agents that work with utility functions have a very short-term memory, only containing the outcome of the previous state, whereas deliberative agents have an underlying memory structure and management in which inferences of a longer period of time can be stored. Hence, a deliberative agent can store experiences and facts about the world around himself and with help of inferences and knowledge about "how the world works", a world-model can be constructed (Russell & Norvig, 2003).

Cognitive architectures, such as SOAR (Lehman et al., 2006) and ACT-R (Anderson & Lebiere, 1998) support memory models of long-term memory (LTM) and short-term memory (STM) (cf. Atkinson & Shiffrin, 1968). A taxonomy of memory types (see figure 2.2) distinguishes different types of long and short term memories (Anderson, 1976; Tulving, 1985; Lehman et al., 2006; Miyashita, 2004; Wilson, 1975)

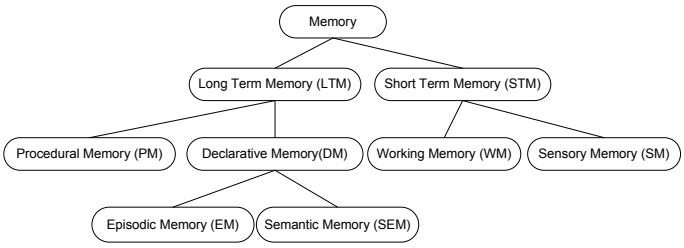

Figure 2.2: Taxonomy of memory types.

Long term memory (LTM), compared to STM, is able to store knowledge for a long time and in a cognitive architecture this memory can be subdivided into procedural memory (PM)—sometimes referred to as a rule-base—and declarative memory (STM), the storage of beliefs or facts. Procedural memory contains procedures and skills about how and when to act, whereas declarative memory contains facts and inferences[5] about the world. In declarative memory, a distinction can be made between semantic (SEM) and episodic memory (EM). Semantic memory is used to store meanings, understandings and factual knowledge about the world. The episodic memory (Tulving, 1972) is used for specific episodes with a reference to an event that happened at a certain time in the past and is often associated with emotions that strengthen its presence in memory.

Short term memory can be divided into sensory (SM) and working memory (WM). Sensory memory (Sperling, 1960) is a very short lived memory in which traces of input are stored for a while. Sensory memory can be detected in practise when looking at a fast turning wheel with spokes; the wheel seems to turn anti-clockwise because there are still past traces in memory.

Working memory (Baddely, 1986) is a storage unit for temporary information storage and manipulation. The capacity is limited (Miller, 1956), e.g. think about the use of paper when trying to calculate a complex equation.

In this dissertation, the procedural memory, the declarative memory and the working memory get the most attention. A further explanation of all other memory aspects is beyond the scope of this dissertation.

In current cognitive architectures like SOAR and ACT-R, the focus is on procedural, declarative (semantic) and working memory. In both architectures, the working memory functions as a placeholder for knowledge that is relevant for the current situation. Knowledge inside working memory is directly available, while knowledge in long-term memory first needs to be retrieved in order to be used. SOAR and ACT-R both describe procedural memory as the place where skills and rules (if $\rightarrow$ then) are stored that enable the actor to act when certain conditions take place. The function of the declarative memory is to supply missing information that a procedure needs to solve the problem. Scientists in the field of memory research, are especially interested in the way people recognise, recall and re-learn certain items in memory (Anderson & Lebiere, 1998). ACT-R allows to model these phenomena with help of sub-symbolic[6] activation in order to make memory degrade over time and simulate how people forget and learn, and how associative memory can increase latency times when more elements are added to memory (cf. the fan effect: Anderson, 1974). The properties of memory have an effect on the performance of remembering and learning.

2.2.4.2 Learning

Closely connected to theories of memory are theories of learning. Remembering, recognition and recall are processes that are inherent to processes of learning. Memory alone supports in collecting and maintaining knowledge, however there is a need to classify over time, based on criteria, what experiences were

[5]For example: All men are mortal, X is a man $\rightarrow$ X is mortal.

[6](cf. Smolensky, 1988)

successful or not. Learning defines these criteria that can be based on, for example, external or on internal goals, drives or beliefs, often with the help of utility functions. Following, Rumelhart and Norman (1978), three kinds of learning[7] can be distinguished:

> (a) *accretion*, or the encoding of new information in terms of existing schemata[8]; (b) *restructuring or schema creation*, or the process whereby new schemata are created; and (c) *tuning or schema evolution*, or the slow modification and refinement of a schema as a result of using it in different situations. (Shuell, 1986, p. 421)

Accretion and restructuring (also called symbolic learning) involve the creation of new knowledge (or symbols). With accretion, new knowledge is added to memory without changing the structure of memory. With restructuring, inferences or structures in memory are changed without acquiring new knowledge from the outside world. However, new knowledge can be generated or inferred from prior knowledge already present in the mind. Tuning of information is the (evolutionary) change of generalising or constraining the applicability of knowledge.

In a similar way, machine learning also offers different types of learning, e.g. supervised, unsupervised, and reinforcement learning (Russell & Norvig, 2003). Supervised learning concerns the construction of a function (induction) with the help of training examples delivered by the outside world. For instance, the supervised learner is able to classify objects that were previously unknown to the learner and the learner's reference is based on training examples delivered to him by the outside world. Unsupervised learning is the learning of associations or classifying about the outside world without having feedback from the outside world. For example, in machine learning, the Hebbian learning rule (Hebb, 1949) states that the connection between two neurons is strengthened if they fire at the same time.

While supervised learning assumes that there is a clear feedback teacher signal, reinforcement learning is learning what to do, i.e. how to map situations to action, in order to maximise a numerical reward signal. Reinforcement learning takes into account the trade-off between *exploration* and *exploitation*:

> A central concern of studies of adaptive processes is the relation between the exploration of new possibilities and the exploitation of old certainties (Schumpeter, 1934; Holland, 1975; Kuran, 1988)...Adaptive systems that engage in exploration to the exclusion of exploitation are likely to find that they suffer the costs of experimentation without gaining many of its benefits. They exhibit too many undeveloped new ideas and too little distinctive competence. Conversely, systems that engage in exploitation to the exclusion of

[7] These three kinds of learning are similar to the concepts 'assimilation', 'accommodation' and 'equilibration' of Piaget (1985).

[8] Schema(ta) represent knowledge of all kinds from simple to complex.

> exploration are likely to find themselves trapped in suboptimal stable equilibria. As a result, maintaining an appropriate balance between exploration and exploitation is a primary factor in system survival and prosperity. (March, 1991, p. 71)

Another type of learning, evolutionary learning or better evolutionary algorithm, supports the creation of new generations of rules with help of cross-over or mutation. Cross-over "causes the characteristics of the parents to appear in new combinations in the offspring. It is the recombination of *sets* of alleles[9] that is most interesting from the point of view of rule discovery..." (Holland, 1995, p. 66) and mutation is "a process whereby individual alleles are randomly modified, yielding a different allele for the gene" (ibid., p. 70). With help of cross-over or mutation, it is possible to explore and discover new rules. These new rules will compete with the parent or previous rules in a selection process that determines if new rules are better adapted to the current situation. Such a selection process is often defined as a utility-based mechanism that tries to determine the best fit with the current environment.

Utility functions or utility-based mechanisms are often used as a performance measure—relative to a goal—for learning and can be distinguished into two types: (1) the agent has a model of the environment, which is used as an estimator for expected utility, e.g. cellular automata, and (2) the utility function is independent of the environment and choices are based on the utility of actions alone, e.g. Q-learning (Watkins, 1989).

Learning defines mechanisms that support the agent in the decision about what is wrong or what is right to do. However, the criteria for making the right decision are delivered by separate normative processes. These internal or external normative processes, i.e. internal drives or norms of a society, determine what goals the agent has to achieve and what behaviour is desired.

2.2.4.3 Goal-directed behaviour, meta-cognition and emotion

Goal-directed behaviour in combination with a performance system[10] allows the agent to learn and use the desired goal as a comparison to the current state of the agent. In SOAR, for example, every task of attaining a goal is formulated as finding a desired state in a *problem space*—a space with a set of operators that apply to a current state to yield a new state (Laird, Newell, & Rosenbloom, 1987). With the help of a goal, the agent can start to divide and conquer the problem[11] by introducing sub-goals—called universal subgoaling in SOAR—that are easier to fulfil than the main goal.

[9] Allele: one of two or more alternative forms of a gene that arise by mutation and are found at the same place on a chromosome (Pearsall, 2002).

[10] A performances system is: ... a system capable of certain performances [which is transformed by learning] into a system capable of additional [performances] (usually better ones; usually accomplished without losing much of the preexisting performance capability; and usually integrated with existing capability so they can be evoked on appropriate occasions). (Newell & Simon, 1972, p. 7)

[11] Among some of the methods used for solving problems are hill climbing, means-ends analysis and alpha-beta search.

The solving of problems requires goal generators[12] and goal comparators that are necessary to resolve conflicts (Sloman, 1987). These goal generators and comparators are driven by motives and constraints. Motives and constraints are often designed in a subsystem, e.g. a motivational, meta-cognitive, or emotional subsystem.

The motivational subsystem monitors drives (such as primitive drives, hunger, need for social interaction, etc.) and interactions that have impact on the selection of the goal and thereby the actions that satisfy the selected goal (Breazeal & Brooks, 2004; Sun, 2003). The meta-cognitive subsystem monitors and tries to improve the cognitive performance (Sun, Zhang, & Mathews, 2006).

Compared to the motivational system, meta-cognition is self-reflective; it can alter learning mechanisms or even switch to other tactics to improve performance.

The emotional subsystem perceives internal and external affordances. It regulates other cognitive systems, promotes appropriate decision-making and expresses the internal states to the outside world, e.g. emotional facial expressions (Breazeal & Brooks, 2004).

This outside world of an agent determines what is necessary to function in such a world. For instance, when the environment is populated with other agents, there can be a requirement to communicate in a language that allows agents to behave socially.

2.2.4.4 Environment, communication, language and social ability

The previous components were clearly fixed components of an agent. The elements mentioned here (environment, communication, language and social ability) are not explicit components of the agent, but rather influence the design of the components of the agent. For instance, agents that are not situated and not dependent on other agents and can solve their problems like chess computers, do not need language or social ability; they can solve problem spaces for a known pre-programmed accessible environment without 'really' perceiving the outside world. However, agents that live in an in-accessible environment surrounded with social (unpredictable) agents that require language to communicate with each other, depend on more than only a set of rules, i.e. they continuously have to adapt to new situations in a changing environment.

The definition of the complete agent with all its necessary components is always under debate. Many components will be discovered, removed or replaced in designing the agent, i.e. it evolves under the hands of an agent model designer. A nice bottom-up example of how an architecture can be constructed, see figure 2.3, is given in Autonomous Agents as Embodied AI (Franklin, 1997).

In figure 2.3, we can clearly see the complexity that occurs when we build a model out of just a couple of components. Therefore, researchers often put their efforts in certain parts of the architecture, e.g. perception has grown towards a specialised topic in cognitive science. Also there have been many discussions about whether a complicated agent, such as a cognitive architecture,

[12] Without goals the system is heading to an unknown direction and is not pro-active.

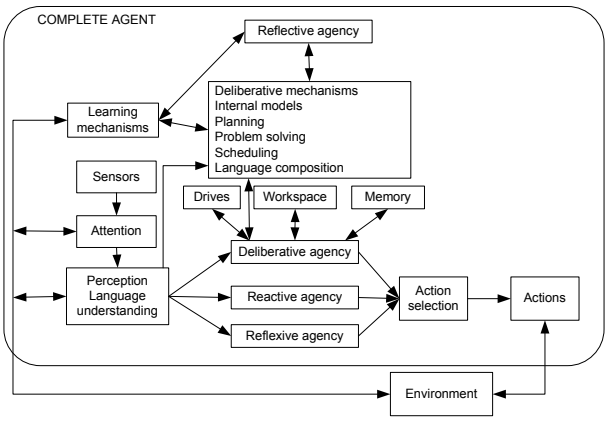

Figure 2.3: Agent with components (adapted from Franklin, 1997).

is really necessary, and feasible to implement in practical situations. These varying points of view have delivered many agent types that can be put together/constructed from a collection of components; even the way they are implemented varies. Among those variations we are interested in two types of agents: the *cognitive agent*, focusing on the internal components of the agent, and the *social agent* that focuses on the (social) interaction—cooperation and coordination—with other agents.

2.3 The cognitive and the social agent

With help of the presence (or not) of certain components, agents can be classified in different types. Many scientists give examples of classifications of agents (Genesereth & Nilsson, 1987; Jorna, 2002; Russell & Norvig, 2003; Wooldridge, 2002; Franklin & Graesser, 1996). The agents are often constructed of functional components, with each component fulfilling a specific function, e.g. vision, memory, motor control etc. In this section, we apply a classification commonly used in cognitive science and AI that builds up an agent from simple mechanisms towards complex physical and social situated mechanisms. Davis (2001) and Sloman (1993, 2001) have created such a classification of agent types: the reflexive/reactive, the deliberative and the reflective agent, comparable to the layers of agency shown in figure 2.3. This classification forms a basis under which all other cognitive approaches can easily find a home.

Two types of agent classification / agents are discussed in this section. The first is the cognitive approach (section 2.3.1 until 2.3.3) that varies from reflexive towards reflective. The second approach (section 2.3.4) will be the social agent. The cognitive agent is mainly concerned with the internal mechanisms of the agent, while the social agent is concerned with the influence of the environment affecting its behaviour and discusses aspects like autonomy, interaction with other agents, and normative and social behaviour.

2.3.1 The reflexive and reactive agent

Reflexive agents respond to the environment by an immediate action, whose functionality is often hardwired in the brain. Deliberation processes do not take place and the design of the model is often pure behaviour based, i.e. only a restrict set of known stimulus-response behaviours is implemented. The reactive agent is slightly more complicated. It gives structure to these behaviours and provides a bit more flexibility and control with help of mechanisms in order to adjust itself towards more different types of tasks. These agents do not have mechanisms that involve explicit deliberation or making inferences (a symbol system). They lack the ability to represent, evaluate and compare possible actions, or possible future consequences of actions (Sloman, 2001). For the reflexive agent, if necessary, a (static) goal can be set that states that the goal/behaviour, e.g. 'searching for food', becomes active as soon as the amount of energy in the body is at a too low level.

In case of the reactive agent: the subsumption architecture[13] of Brooks (1986) is an example of a reactive agent. The agent is physically situated and responds only to the current situation it gets involved in.

2.3.2 The reasoning agent

The reasoning or deliberative agent adds deliberative processing to the mechanisms used by the reflexive and reactive agent. The deliberative agent has many components in common with classical cognitive architectures (GOFAI[14]) and includes a representation of the environment, a memory, a workspace, a planning unit, management of goals and many other components that make process deliberation possible. According to Sloman (2001), a deliberative agent has a set of context-dependent and flexible processes, e.g. plan and goal generation, comparison, evaluation and execution providing the basis for a number of components, that cannot be handled by a purely reactive architecture. The interaction and the feedback that can be compared with the past representations stored in the memory give the deliberative agent opportunities to build up expectations of what effect certain actions will have on its environment and its own well-being. The ability to build up representations about its environment gives the agent the possibility to adapt and survive based on an experienced repertoire of the past. We can distinguish three types of agents that are commonly associated with the reasoning agent: the deductive reasoning agent, the practical reasoning agent (Wooldridge, 2002) and the cognitive plausible agent.

The deductive and practical reasoning agents often have a system that maintains a symbolic representation (logics) of its desired behaviour and a system that can manipulate this representation. Whereas the deductive reasoning agent works with deduction and theorem-proving, the practical reasoning agent is specialised in reasoning that is directed towards the future in which selection between conflicting considerations are provided by the agent's desires/values/cares and what the agent believes (Bratman, 1990). A well known

[13] We will give an explanation of the subsumption architecture in section 4.1.2

[14] Good Old Fashioned AI

implementation of this agent is the beliefs/desires/intentions (BDI) architecture or Procedural Reasoning System (PRS) (Wooldridge, 2000). The weakness of these agents is that although they form representations, such as beliefs, desires and intentions, they only operate at the intentional level (Dennett, 1987), and therefore developers of these agents are not concerned with how these agents could be symbolically grounded. Or as Wooldridge states: "... I will not be concerned with *how* beliefs and the like are represented... the assumption that beliefs, desires, and intentions are symbolically represented is by no means necessary for [the modelling of BDI agents]" (Wooldridge, 2000, p. 69).

In contrast to the deductive and the practical reasoning agents, the cognitive plausible agent operates at the intentional *and* functional level, and is based on more than ascribing mental attitudes to agents alone. The agent is said to be cognitive plausible when it has a cognitive architecture not only based on a physical symbol system (Newell, 1980), cognitive mechanisms, goal-directed behaviour and learning capabilities, but is empirically tested as well (cf. Newell, 1990; Anderson, 1990; Van den Broek, 2001), i.e. the definition of a cognitive plausible agent is:

> [A *cognitive plausible agent*] is a goal-directed decision-maker who perceives, learns, communicates, and takes action in pursuit of its goals, based upon [a physical symbol system] that implements theories of cognition [and is supported by empirical evidence]. (Van den Broek, 2001, p. 88)

The introduction of production systems (Newell, 1973) as systems that can implement theories of cognition are adopted by many researchers in cognitive science. A production system is a physical symbol system that exists out of a set of productions (condition-action patterns) and a set of data-structures. A set of goals in combination with means-end reasoning allows the agent to explore problem spaces and exhibit intelligent action (Newell & Simon, 1972). SOAR (Newell, 1990) and ACT-R (Anderson & Lebiere, 1998) are well known examples of agents with a cognitive architecture. These systems can store and manipulate representations and contain subsystems and mechanisms that enable the actor to adapt (e.g. sub-symbolic learning in ACT-R) and learn (e.g. chunking in SOAR). For more details, we refer to chapter 4 that discusses cognitive theories and gives an elaborate description of a cognitive plausible agent.

2.3.3 The reflective agent

The reflective agent builds on top of the previous agent. It is equipped with a meta-management module that observes and monitors its own cognitive processes in order to reach better performance that serves the goal or objective the agent has set its mind to. For instance, the reflective agent can for instance use the following operations for monitoring its own cognitive processes (Sloman, 2001): (1) the ability to think about and answer questions about one's own thoughts and experiences, (2) the ability to notice and report circularity in one's thinking, and (3) the ability to notice opportunities for changing one's thinking.

Sun et al. (2006) term influencing the cognitive processes as meta-cognition and implement this in the cognitive architecture CLARION (Sun, 2003, 2002). They give examples of mechanisms that aim for—to name a few— (1) behaviour monitoring, e.g. setting of goals and reinforcement functions, and (2) information filtering, acquisition and utilisation of different learning methods and reasoning methods. Research at this layer has just started and the addition of reflection creates more complexity, especially when an agent is equipped with many components that can be reflected upon.

Another example of reflection is the introduction of emotional aspects that can drive the agent in changing its behaviour radically. Such changes can be caused by properties that say something about the current status of the system, e.g. the agent feels itself bored, angry or afraid.

The focus of this dissertation is not to include the reflective mechanisms of the individual, but is more directed at reflection by interaction with the outside world and other actors. In the next chapter we will introduce the social construct that operates at the normative level of an agent and can influence the production system in its (performance) outcomes (e.g. allowing or prohibiting certain outcomes).

2.3.4 The social agent

Another approach of looking at agent typology is to incorporate the social and organisational aspects of the agents. For instance, the social and cognitive agent of Conte and Castelfranchi (1995b) and the Computational & Mathematical Organization Theory (cf. Carley & Gasser, 1999) highlight the importance of agents that create social goals and are physically, socially and culturally situated as well. Verhagen (2000) defines a typology (cf. Conte & Castelfranchi, 1995b) of autonomous agents that have their own goals, in the pursuit of which they might refuse to do what they are asked to do.

1. Reactive agents
   Comparable with the cognitive typology: a reactive agent has no means of influencing the environment in a preconceived way. The autonomy resides completely in the combination of environmental cues and properties of the system.

2. Plan autonomous agents
   Plan autonomous agents have more autonomy than reactive agents in the sense that they may choose how to achieve a certain state of the world. They still do not have the notion of goal, i.e. their goals are determined by a control system from the outside. A given goal can be connected to a repertoire of actions that they can aggregate into a sequence resulting in a plan that will achieve the objective, i.e. the agent has a plan library that is matched with a given goal.

3. Goal autonomous agents
   Goals are not just created by requests but must also be linked to goals the

agent already has. A goal autonomous agent has the autonomy to determine what its "prevailing interest" is, considering its goals. It will judge which states of the world are its interests by evaluating what it provides in terms of degree of goal satisfaction and goal priority.

4. Norm/social autonomous agents
   These agents choose which goals are legitimate to pursue, based on a given system of norms. The agent has the autonomy of generating its own goals and is able to choose which goal it is going to pursue. Besides that, the agent is equipped with the capability to judge the legitimacy of its own and other agents' goals. When a goal conflict arises, the agent may change its norm system thereby changing priorities of goals, abandoning a goal, generating a new goal, etc. Norm autonomous agents generate norms they can use to evaluate states of the world in terms of whether or not they could be legitimate interests. Legitimacy is a social notion and is in the end determined by the norms of the agent with respect to the agent society it is part of.

Right now, we do not want to explain exactly what is necessary for constructing a social agent, but only show a typology of agents. However, we will return to social capabilities in section 2.5 that is concerned with coordination and in chapter 3 that elaborates about the characteristics of the social actor.

2.3.5 The complete cognitive and social agent

In this section we want to reflect on the previous typologies of the cognitive and social agent and try to map their characteristics or capabilities next to each other and thereby see what differences there are.

Carley and Newell (1994) made an attempt to determine the characteristics of a cognitive and social agent. They have created the so-called Model Social Agent (MSA) which is in our opinion a complete social and cognitive agent. Their description of the Model Social Agent targets quite well the type of agent we are looking for in our research.

> The Model Social Agent has information-processing capabilities and knowledge. Agents' information-processing capabilities are goal oriented. They control the agent's ability to handle information. Agents exist within an environment which is external to the agent's processing capabilities. The agent's knowledge is to an extent dictated by the external environment in which it is situated. The Model Social Agent exists in a particular situation (both physical and social). This situation is the environment perceived by the agent, but how the agent encodes it, and how much of the environment is encoded by the agent is an open issue[15]. The agent has a goal. The agent enters a situation with prior knowledge. The agent may have

[15] In this dissertation we will try to fill in this open issue with the introduction of the social construct in the next chapter.

> an internal mental model of the situation that differs from that held by other agents in the same situation. Throughout, we take the agent as having the typical human sensory and motor devices to sense the environment and affect the situation. We are concerned with the nature of the inner system that makes interaction social... (Carley & Newell, 1994, p. 223)

Carley and Newell describe the agent along two dimensions: increasing limited *processing capabilities*—limitations that enable it to be social, and increasing rich situations—*richness* of the agent's *perceived environment* that evokes and supports social behaviour, i.e. how the agent is situated in the environment. The dimensions give the possibility to create a 1x1 classification matrix (Carley & Newell, 1994, p. 243) along the lines of limited processing capabilities and the limitation in richness of the perceived environment.

Along the line of limited processing capabilities, five types of agents can be discerned:

1. The Omnipotent Agent (OA): the agent "knows all there is to know about the task environment in which it is embedded" (ibid., p. 227).

2. The Rational Agent (RA): "An agent has its own body of knowledge and it behaves rationally with respect to that knowledge by taking those actions that its knowledge indicates will lead to attaining its goals" (ibid., pp. 227–228).

3. The Boundedly Rational Agent (BRA): The boundedly rational agent has ecological and cognitive limitations of both (environmental) knowledge and processing capability (Simon, 1945), i.e. the agent "has limited attention and therefore cannot process all the knowledge available in its task environment" (ibid., p. 228).

4. The Cognitive Agent (CA): "The cognitive agent is the boundedly rational agent with a fully specified architecture... Where the boundedly rational agent is a set of general claims, the cognitive agent is a set of specific operational details. Moving from general principles to specific architecture further limits the agent. Where the boundedly rational agent may have some type of knowledge structure called a chunk, in the cognitive agent we know the exact form of these chunks and the procedure for creating them" (ibid., p. 230).

5. The Emotional Agent (EA): "The emotional agent, while more human than the cognitive agent, also is more constrained in its ability to attain its goals. We could add emotions to each of the other limited agents, the rational, boundedly rational and cognitive agents. These all would reflect the same notions of how emotions limit goal attainment" (ibid., p. 235).

Carley and Newell also made a distinction along the other dimension that varies in the richness of the situation the agent perceives, i.e. the types of knowledge available to the agent determines the types of behaviours the agent can

engage. They have distinguished five[16] types of situations the agent can be situated in.

1. Non-social Task Situation (NTS): The non-social task situation is a situation in which an agent only deals with a task environment that is not affected by social factors from other agents, i.e. in the extreme case, we can take the chess computer that only deals with the outcome of the chess-game. The chess agent deals with a fixed physical environment and deals with the moves of a player, but does not take into account any social factors of the player that plays against him.

2. Multi Agent Situation (MAS): In this situation, the agent deals with other agents and treats them as having their own goals, but the goals are not social goals that maintain social relations.

3. Real-time Interactive Situations (RIS): The situation is similar to the previous situation, but the agent has to balance the processing of interpreting information versus acting with its environment as well as responding to its environment in a timely fashion.

4. Social Goal Structure Situation (SGSS): "This situation is characterized by the presence of groups and group related knowledge such as rules for membership, maintenance, and dissolution... The existence of social structure permits the separation of task and self; i.e. the agent may now be faced with goals that are demanded of it by an external entity, such as a group or organization. The agent, situated in a social structure, may now be faced with multiple goals[:]... (1) task-related goals, (2) self-maintenance and enhancement goals, and (3) social-maintenance goals... that may compete and conflict with each other, as well as multiple goals at the same level" (ibid., pp. 239–240).

5. Cultural-Historical Situations (CHS):

   > The agent exists at a particular point in time and history, in a particular society, at a particular place, with particular institutions, and so on. The situation is affected by, and has resulted from, a historical process that led to the current state and moment. The agent has available a body of social practice that has been shaped by a specific culture in the ways that may be idiosyncratic to that culture. The agent has "developed" within this situation and so is socialized to a particular way of doing things, to specific beliefs, norms, and values. (ibid., p. 239)

Carley and Newell created the "Model Social Agent" as a combination of the Cognitive Agent (CA) and the characteristics of the emotional agent (EA)

[16] Instead of the six types that Carley and Newell distinguish, we have merged two types or situations into one, i.e. we have merged Social Goals Situation (SGS) and Social Structure Situation (SSS) into Social Goals Structure Situation (SGSS), because they are, according to their description, almost similar.

that deals with other agents in a cultural-historical situation (CHS). With help of the dimensions, they have tried to create a better understanding of the social nature of human beings and applied it as a framework in order to determine how far current research is and what the aim of that research should be to reach the "Model Social Agent".

We adopt both dimensions to determine what type of agent and environment, as suggested by Carley and Newell, fits the characteristics of the cognitive and social agent in the previous sections. Table 2.1 gives an overview of the characteristics of the cognitive and social agent and is an estimation of the capabilities of the agents[17].

Referring back to research question 1 (particular questions 1.1. and 1.2) and after examining table 2.1, we can state that in order to create an agent that is both cognitive and social, it requires a combination of the characteristics of the cognitive plausible agent and the norm/social autonomous agent. Therefore, we will address the social agent in chapter 3, and the cognitive agent in chapter 4 in order to create a combined—social and cognitive—agent that conforms as much as possible to the Model Social Agent.

The discussion in this section has elaborated that the notion of a social agent can only exist in an environment with other agents and moreover that bounded rationality creates at the same time the need for that interaction. This need can *only* be fulfilled if the environment is possible to deliver such interactions. Therefore, a well designed model of an agent should take into account the properties of the environment in order to make sure that coherence between environment and agent is optimal.

[17]The table is an estimation, in the sense that some agents can have more or less characteristics than shown in the table, i.e. the classification of the agents is not as strict as shown in this table, but is more a gliding scale or approximation.

	Level of description	*Production or plan oriented*	*Goal / behaviour oriented*	*Neural plausible*	*Meta-cognition / motivation*	*Mutual Modelling*[a]	*Physically situated*	*Communication (language)*	*Socially situated*	*Processing capability*	*Richness environment*
Cognitive agent											
Reflexive / reactive	B			-	-	-	+	-	-	OA	NTS
Deductive Reasoning	R	Pl		-	-	-	-	-	-	OA	NTS
Practical Reasoning	R	Pl	G	-	-	+/-	+/-	+/-	-	RA	MAS, RIS
Cognitive plausible	R,F	Pr	G	+	+/-	+/-	+/-	+/-	-	BRA, CA, EA	MAS, RIS
Social agent											
Reactive	B			-	-	-	+	-	-	OA	NTS
Plan autonomous	R	Pl		-	-	-	-	-	-	OA	NTS
Goal autonomous	R	Pl	G	-	-	+	+/-	+/-	-	RA	MAS, RIS
Norm / social autonomous	R,S	Pl	(S)G	-	-	+	+/-	+	+	RA	MAS, RIS, SGSS
B = Behaviour R = Rational F = Functional S = Social	+ = yes +/- = possible - = no	Pl = Plan Pr = Production G = Goal								OA = Omnipotent Agent RA = Rational Agent BRA = Boundedly Rational Agent CA = Cognitive Agent EA = Emotional Agent	NTS = Nonsocial Task Situation MAS = Multi Agent Situation RIS = Realtime Interactive Situation SGSS = Social Goal Structure Situation	

Table 2.1: Characteristics of the social and cognitive agent.

[a]Mutual modelling is the ability of an agent to build up a representation of the representations of other actors, e.g. I believe that he believes. We refer to section 2.5 for more details about mutual modelling.

2.4 The environment of the agent

The properties of the environment determine partly the constraints and the properties of the model of the agent. Russell and Norvig (2003) define a list of dimensions along which environments can be categorised.

- Fully observable vs. partially observable
  A task environment is fully observable if the sensors of the (omnipotent) agent detect all aspects that are relevant to the choice of action. On the other hand, in a partially observable environment, the agent has no access to all information that is hidden in the environment or other actors.
- Deterministic vs. stochastic
  When the agent can influence and create a next state in the way it wants, the agent is dealing with a deterministic environment, i.e. during the time the agent is deliberating, the environment does not change. An environment that consists of more agents and is partially observable is defined as being stochastic. The agent lives in an environment that continuously changes; behaviours of other agents become more or less unpredictable.
- Episodic vs. sequential
  The time line in an episodic environment is divided into different episodes that do not influence each other. An example of an agent that acts in an episodic environment is an agent at an assembly line in a factory that repeatedly mounts a part on a car in which the current decision does not affect future decisions, i.e. in the case of the assembly line, the environment returns to its begin state. Therefore, the learning capabilities in these types of environments are small. On the other hand, in sequential environments, it is possible for the agent to learn, explore and reason about future states.
- Static vs. dynamic
  The environment is dynamic when it develops while the agent is deliberating about the environment. A static environment, such as a chess game, does not change during the time the agent is calculating the next move.
- Discrete vs. continuous
  The discrete-event simulation and finite state machines are examples of discrete environments and the time in these environments evolves step by step. Whereas in a continuous environment the agent has a continuous-state and continuous-time problem, e.g. a real-time system[18] in an aircraft that controls the plane.
- Single agent vs. multi-agent
  A Multi-Agent System compared to a single agent is complex, caused by interaction that occurs between agents. Interaction in the form of communication can lead to organisational and social behaviour, conflicts or cooperation resulting in complex behaviour.

[18] Neglecting the fact that the processors of most machines are time-slicing and do divide any simulation in time steps

The environment has impact on the autonomy of an agent. Odell, Van Dyke, Fleischer, and Breuckner (2002) define an environment[19] as follows: "An environment provides the conditions under which an entity (agent or object) exists" (p. 2). They define three types of environment:

The physical environment: provides the principles and processes that govern and support a population of entities.

The communication environment: provides those principles, processes, and structures that enable an infrastructure for agents to convey information.

The social environment: a communication environment in which agents interact in a coordinated manner.

The communication environment cannot exist without a (virtual) physical environment and the social environment cannot exist without the communication environment, see figure 2.4.

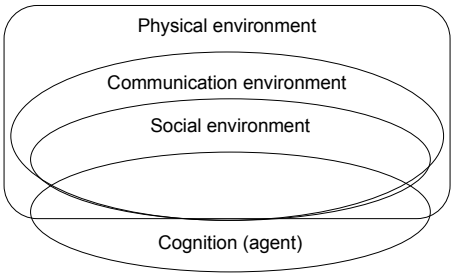

Figure 2.4: Nesting of environments (adapted from Odell et al., 2002).

The physical environment is often *simulated*, because simulated environments give the ability to easily create, change or test many different situations in which the agent may be involved. On the other hand, robotics is a field that tests agents in a real physical world, e.g. the laboratory in which the circumstances are still under a certain control. The simulated physical environment has to do the work that the real physical world does in the case of real-time systems: enforce the consistency in time, in space, and with respect to physical laws, i.e. it has to maintain synchronisation between agents and environment. In this respect, a simulation has to do more work (not less) than a system functioning in the real world. The same can apply for the communication and social environment. Communication and interaction in a MAS simulation are still under control of the experimenter, while HCI (Human-Computer-Interaction) tries to connect to the real interactive and social world by creating a mixed environment of humans and computers. Another important environment is the task environment. The task environment can be seen as an environment that represents tasks that are possible for an agent under restriction of the physical, communication and social environment.

[19] In the next chapter, we explain the 'semiotic Umwelt'; an environment that consists of signs and enables the agent to interact, i.e. the agent perceives, interprets and acts upon those signs.

Task Environment

The way tasks can be performed and what kind of tasks can be performed is constrained by the physical, communication and social environment. These environments construct a set of possibilities for the agent to respond to, i.e. the outer task environment. However, the agent has attention for only a subset of tasks in the outer environment[20] and this specific set is an unique task environment to the agent himself, i.e. it is the combination of the inner and outer environment that is relevant for an agent to do a task. Simon also addresses the distinction between an inner and an outer environment:

> The advantage of dividing outer from inner environment in studying an adaptive or artificial system is that we can often predict behavior from knowledge of the system's goals and its outer environment, with only minimal assumptions about the inner environment.... In one way or another the designer insulates the inner system from the environment, so that an invariant relation is maintained between inner system and goal, independent of variations over a wide range in most parameters that characterize the outer environment. (Simon, 1996, p. 8)

The difference between inner and outer environment is evident when we compare the reactive agent (embodied cognition) with the cognitive agent (classical approach). The reactive agent its task environment is mainly situated in the outer environment and is (re)constructed the moment the environment changes. On the other hand, the classical approach represents the majority of tasks as problem spaces in the mind of the agent self. Embodied cognition or situated cognition have stressed that there is clearly an interdependent relationship between the internal represented task environment and the physical, communication and social environment, i.e. the model of the agent can constrain for instance actions of other agents (the social environment), and the physical environment can constrain the options (tasks) of the agent.

The tasks that an agent has to accomplish in its environment create (partly) the need for a situated model of the agent that has to fulfil the requirements of the environment. For instance, the environment of a chess computer can be the chess computer itself, getting input from the position of the pieces on the board and its internal rule-base. Whereas an agent that guides a plane to land is more situated in the environment and has to respond quickly to changes in the environment, e.g. side-winds etc. The same aircraft agent does not only have to be communicative, but social (and economic) as well when it has to interact with other planes to negotiate for a time-slot to land the plane.

In our research, we are interested in the aspects of actors that explain social behaviour. Therefore we need a complex environment (see chapter 5), i.e. a physical, communication and social environment. This environment is a specific environment suited for the traffic experiments conducted in chapter 6. The environment is a relatively rich environment, comparable to the SGSS (Social Goal Structure Situation) that consists of other agents, a communication environment

[20] The bounded rationality argument (Simon, 1945).

and actors that have the capability to create a social environment. The purpose of the environment is to create the possibility for simulating (social) interaction and coordination (mechanisms) between actors that have to solve (shared) problems. The next section will discuss conflicts, shared problem solving between interacting agents and coordination mechanisms that try to reduce the conflicts between agents.

2.5 Agent and Coordination

Partly due to the inaccessibility of the environment, the life of a human being and its (knowledge) resources are scarce, therefore agents are boundedly rational (Simon, 1945). Hence, the need for cooperation and conflict in order to survive is inevitable. In the first part, we will elaborate when coordination occurs and what the motives[21] are to cooperate and solve conflicts. In the second part of this section, we will discuss several coordination mechanisms that try to avoid conflicts and coordinate problem solving.

2.5.1 Cooperation and Conflict

Conflict or cooperation suggests that there are relationships between activities that, according to the coordination typology (see figure 2.5 of Von Martial (1990), either are negative relationships (conflict) or positive relationships (cooperation).

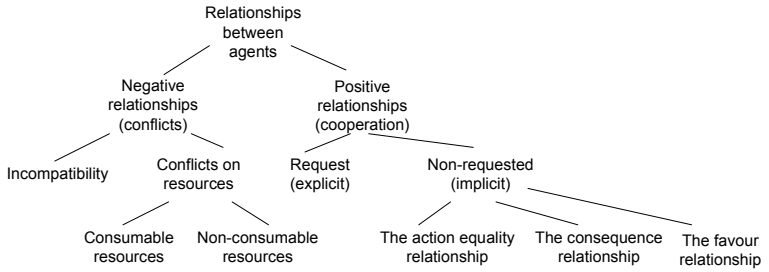

Figure 2.5: Typology of coordination relationships (Von Martial, 1990).

The negative relationship occurs when agents have an incompatibility between goals. For example when two agents execute actions that lead to a goal, but the outcome of the success (the goal) conflicts with the other, e.g. in a car race, only one car can win. The same can happen with resources that are shared

[21] Economic motives are important for estimating if cooperation is fruitful. Coordination costs are not explicitly discussed in this section, but for example, they can exist of "(Jorna, Gazendam, Heesen, & Van Wezel, 1996, p. 29): (a) organization, including establishing contracts with organizations that perform work that is contracted out, (b) improvisation, (c) slack capacity of actors, means of production and materials in stock, (d) making plans, (e) replanning and adjusting plans, (f) storing and communication of plans, including the costs of maintaining an information and communication system, and (g) accepting plans by the actors that have to perform the plans as well as by other stakeholders" (as cited by: Gazendam, 2006, pp. 167–168).

by agents, e.g. occupying a parking spot. The positive relationship is one in which there is either an explicit request (e.g. speech) for cooperation or there is a non-requested relationship that turns out to be favourably for one of the agents and is not harming the other.

Von Martial (1990, p. 112) distinguishes three forms of non-requested relationships:

1. The action equality relationship. We both plan to perform an identical action, and by recognising this, one of us can perform the action alone and so save the other effort.

2. The consequence relationship. The actions in my plan have the side-effect of achieving one of your goals, thus relieving you of the need to explicitly achieve it.

3. The favour relationship. Some part of my plan has the side effect of contributing to the achievement of one of your goals, perhaps by making it easier (e.g. by achieving a precondition of one of the actions in it).

In a similar vein, following Schmidt (1991), Gazendam and Homburg (1996) distinguish four forms of cooperation.

1. Augmentative cooperation. When single agents are limited by their physiological and mechanical capabilities, and physical or economic resources, cooperation allows agents to achieve a task which would otherwise be impossible.

2. Integrative cooperation. The task requires specialization from agents in order to achieve it, because specialization decreases throughput time and reduces a complex problem to smaller and simpler problems that can be solved by individual agents.

3. Debative cooperation. Agents bring in a variety of interests and values, and aim to come to a compromise in which every agent takes a position and is a process in which agents want to come to a mutual agreement.

4. Conflict handling. The handling of conflicts can be controlled by an authoritative organ or 'rules of the game' that prevent agents from ending up in a deadlock or destructive conflict.

The first two types of cooperation can clearly be nested under the positive relationships and fall in a similar category as the requested or non-requested relationships of Von Martial. The last two types are actually conflict and cooperation situations at the same time. The agents start of with conflicting interests, however debative cooperation transforms the initial conflict situation into a cooperation situation. The same applies for conflict handling; the conflict situation is oppressed by authority, norms and rules causing a potential conflict situation transform into a—possible (in)formally forced—cooperation situation.

Cooperation and prevention of conflicts urges the agent to communicate with others, be it verbal or non-verbal. Communication that is expressed in an

(mutually) agreed language gives the possibility to express and transfer meaning of complicated beliefs, goals and so on, to other agents. The possibility of understanding each other's needs gives the drive to social behaviour and the possibility for coordination. In society, coordination is often linked with conflicts about scarcity, i.e. many situations can occur in which scarcity is a reason for conflict.

Ferber (1999) assigns three criteria that determine the type of conflict or cooperation situation that can take place. In the first place, when agents are in an environment, their goals can be compatible or incompatible, i.e. cooperation vs. antagonism. Secondly, goals[22] are often connected to resources: when two agents want access to the same resource, a conflict may arise and coordination may be necessary to solve the problem. Thirdly, to achieve goals, they need to be linked to procedures or skills. Some agents do not have the required skills and need skills of others to achieve their goal. Based on these three criteria, a typology of different situations can be made explicit, see table 2.2.

Goals	**Resources**	**Skills**	**Types of situation**	**Cooperation type**
Compatible	Sufficient	Sufficient	Independence	
Compatible	Sufficient	Insufficient	Simple collaboration	Integrative
Compatible	Insufficient	Sufficient	Obstruction	Augmentative
Compatible	Insufficient	Insufficient	Coordinated collaboration	Augmentative Integrative
Incompatible	Sufficient	Sufficient	Pure individual competition	Debative Conflict handling
Incompatible	Sufficient	Insufficient	Pure collective competition	Integrative Debative Conflict handling
Incompatible	Insufficient	Sufficient	Individual conflicts over resources	Augmentative Debative Conflict handling
Incompatible	Insufficient	Insufficient	Collective conflicts over resources	Augmentative Integrative Debative Conflict handling

Table 2.2: Classification of interaction scenarios (adapted: Ferber, 1999)[23].

[22] Goals sometimes are seen as resources, (e.g. 'That is my goal!') however there are many routes that end up in Rome, in which resources (roads, cars) are connected to the goal. Skills are also often seen as a resource (e.g. an agent is a resource), however here we distinguish it as a mean-to-an-end.

From the typology, it becomes clear that interaction situations become more complicated when goals start to conflict, resources become limited and skills are not adequate. In the optimal situation (Independence) there is indifference about what any agent does, i.e. there is no scarcity. Situations, for which the agents have shortage of skills, compensation in the form of Integrative Cooperation is necessary. In the case of insufficient resources, agents are dependent on each other. Augmentative Cooperation can reduce the lack of physiological, mechanical, physical or economic resources.

Although in situations where goals are compatible *and* there are either insufficient resources or skills in which case augmentative or integrative cooperation can solve coordination problems, negotiations will always remain necessary to make sure that there is (mutual) agreement between agents.

In situations where goals are incompatible, a deliberative form of cooperation like debative cooperation in combination with conflict handling, is necessary. Incompatibility can occur when there are different interests in an organisation. For instance, the tension between the marketing department and the production floor. The marketing department tries to sell as much as possible, while the production floor tries to make sure that the marketing department does not create goals that are unable to fulfil.

Thus, solving conflicts and cooperation requires coordination between agents in order to achieve goals by negotiation, allocating resources and skills. Coordination mechanisms facilitate coordination and are discussed in the next section.

2.5.2 Coordination mechanisms

Coordination mechanisms are mechanisms that stabilise and better control issues concerning scarcity and aim at achieving coordinated action. Economic science (the economic perspective, see next section) has grown out to be an important science in studying scarcity issues. They attempt to regulate our economy by the value of capital and market mechanisms. These mechanisms are expected to control the way individuals should behave, thereby coordinating and regulating the actions of individuals in a structured way. However, economists study the population at a relative high level of abstraction and when we look at the economic situation today, market mechanisms as coordination mechanisms do not achieve what they are expected to do. An alternative for understanding coordination is given by Distributed Artificial Intelligence (DAI) (section 2.5.2.2).

[23] **Incompatibility of goals:** In the first place, incompatibility of goals can occur. It creates a situation in which agents conflict with each other in order to achieve a successful outcome. As mentioned before in a car race, only one can win.
Insufficient resources: The agents are not directly dependent on each other, i.e. the actions and goals are indirectly connected by the resources each agent wants to use. Two types of resources can be distinguished, resources that are used for temporary use and resources that are consumed and change ownership. In the first case, resources can be allocated with help of planning and division of tasks, i.e. first one agent uses the resource and then the other. In the second case, the resource is consumed and can only be allocated to one agent.
Insufficient skills: When an agent has not the capability of achieving a goal it can solve this shortage by outsourcing a task to another agent—when available—or receive training from other agents and thereby has built up experience when the problem reoccurs.

DAI, in cooperation with for instance economics, psychology and sociology, applies sophisticated methods and models to understand how individuals behave.

2.5.2.1 The economic perspective

Economists often take the market or bidding systems as coordination mechanisms that economise and disclose information from bidders and sellers. The actors in this approach are often companies and the market functions as a price system and prices are a way to inform the 'rational' actor about the equilibrium in the market. Prices or utility mechanisms give the ability to 'play games' and try to find an optimal (minimax) strategy to maximize a safe outcome. To describe 'gameplay', game theory is widely adopted in economics and also has found its way into MAS. Game theory (Axelrod, 1984; Morgenstern & Von Neumann, 1947) is applied in studying interaction—conflict and cooperation—and enables the researcher to analyse and understand strategic games or scenarios. Game theory applies a payoff matrix in which the agent has a self-interest that can be expressed with the help of utilities. Table 2.3 displays such a matrix: the numbers reflect the utility both players assign to the outcome if they choose for a certain strategy.

Player I / ***Player II***	Cooperates	Defects
Cooperates	3 3	5 0
Defects	0 5	2 2

Table 2.3: Strategic form of game theory; the prisoner's dilemma.

The matrix presents a popular example of game theory, i.e. the prisoner's dilemma (Axelrod, 1984). The scenario can be described as follows. If one of the prisoners confesses (defects) and the other not (cooperates), then the confessor will receive a reward by giving him less punishment and giving the other additional punishment compared to the standard. However, if both confess (defect), then they get both the standard punishment or if both don't confess (cooperate) they don't get any punishment at all.

Game theory assumes a varying surplus of cooperation and scenarios are often a dichotomy of actions such as in favour of, or against. It is very useful and shows the possibilities of agents in an easy to understand payoff matrix and is applicable in MAS as a *descriptive* theory, i.e. it is a high-level theory (rational level) that neglects underlying mechanisms.

There are however some counter arguments for the game theory (Wooldridge, 2002). First of all, people are not always self-interested, and also show

forms of altruism and spontaneity. Secondly, human-beings are not supposed to be totally rational and, hence, have not perfect information of the environment they live in. The third problem is that the game theoretical approach is often played with two players and does not take into account past experiences, i.e. the so called one-shot games. The Iterated Prisoner's Dilemma (IPD) of Axelrod (1984) introduced the concept of a Multi-Agent System in which every agent has prior knowledge about the step its opponent did before and selected 'cooperate' or 'defect' based on this step. The introduction of iteration and many other techniques makes game theory worthwhile to apply in Multi-Agent Systems for estimating and describing how a system might stabilise at its equilibrium points (cf. Nash, 1950).

Still, in game theory, the agent does not account for social action aimed at modifying the behaviour and mental states of the other agents. Game theory gives no explanation for why agents engage in social action and what internal structures make social interaction possible. Also no explanation is given of how the behaviour at the macro-social level influences the internal mechanisms of the agents (Conte & Castelfranchi, 1995b). Compared to the economic perspective, the DAI perspective tries to incorporate the needs of the individual and the society as well.

2.5.2.2 The DAI perspective

The DAI perspective[24] is a more detailed approach compared to the economic perspective, i.e. coordination is achieved by decomposing the organisational entity into manageable units, such as tasks, goals and actions. In order to control coordination, the process of managing interdependent activities is crucial (Malone & Crowston, 1991). DAI is concerned with coordination behaviours or mechanisms that take care of managing conflicts and cooperation.

> Coordination behaviors [or mechanisms] can be divided roughly into specification behaviors (creating shared goals), planning behaviors (expressing potential sets of tasks or strategies to accomplish goals), and scheduling behaviors (assigning tasks to groups or individuals, creating shared plans and schedules, allocating resources, etc.). Coordination behaviors in turn rest on more basic agent behaviors such as following rules [norms and social laws], creating organizations, communicating information, and negotiation or other conflict resolution mechanisms. (Decker, 1995, p. 9)

In this section, we will shed light on coordination mechanisms that try to solve problems of cooperation and conflict.

[24] As mentioned before, Multi-Agent Systems and Distributed Problem Solving are part of DAI. There is considerable overlap between MAS and DPS, therefore in this section there is not made a clear distinction between them. However, sections 2.5.2.2.1 until 2.5.2.2.3 can be seen as falling in the category of DPS, while the remaining sections have more characteristics from MAS.

2.5.2.2.1 Cooperative Distributive Problem Solving

Cooperative Distributive Problem Solving (CDPS) tries to coordinate the relationships between agents in order to cooperate and solve problems. In CDPS, the following assumptions are made (Durfee, 1999). There is already a certain group coherence, i.e. agents have been designed as a group and agents need to know how to work together. Besides that, the agents are assumed to be benevolent. Based on these assumptions, a joint plan can be constructed in which task allocation can take place.

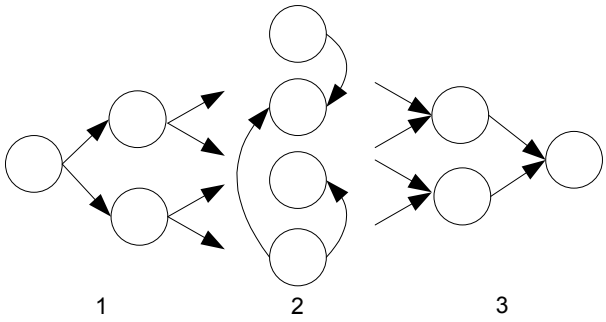

Figure 2.6: The three stages of CDPS (Smith & Davis, 1980; Wooldridge, 2002).

Task sharing in CDPS exists out of several stages (Durfee, 1999), see figure 2.6. A complex task can often be subdivided (1) into several tasks that (2) can be distributed among other agents. In the stage of solving the sub problems, agents solve their own problem and if necessary share knowledge with others to solve the problem. When the sub-problem is solved, the agent sends its solution to a manager agent (3) who oversees the whole problem and solves the main problem. Although simple problems can be solved efficiently and effectively in this way, in heterogeneous systems, where agents are trained as specialists, solving problems and scheduling jobs with agents becomes rather complex. A solution for this problem is a communication protocol, such as Contract-Net (Smith & Davis, 1980). With help of an Agent Communication Language it is possible with CNET to set up an (economic) market principle of bidding and selling in which managers try to find contractors (who can try to find sub-contractors). The process involves the announcement of tasks to others. The others can give a bid on the task and after some time contractor(s) are selected by the manager. When agents are confronted more than once with a problem to solve, then agents can try to improve their performance by sharing the results of sub-problems. Sharing provides for error checking, and increased confidence and precision, a better overview of the overall problem (completeness) and solving the overall problem faster (Durfee, 1999). However, coordination problems can be structured and solutions can be achieved faster and more properly with the help of planning and scheduling.

2.5.2.2.2 Planning

When in Multi-Agent Systems tasks of an agent depend on other agents' uncertainty increases and coordination becomes necessary. Planning and scheduling can release this tension and supports the increase of performance—efficiency and effectiveness—and possibly reduces uncertainty and gives more self-knowledge of the task-structure.

In a Multi-Agent System, planning is characterised by four aspects (Gazendam, 2006, p. 162):

1. Multi-agent plans are distributed, which means that they generally are represented at many places: in the minds of agents and in the documents that agents use.
2. The planning process aims at creating or maintaining coordinated behavior or social order by creating social constructs (see chapter 3) in the form of plans.
3. The planning process is characterized by problem solving intertwined with negotiation. In this process, social constructs (plan elements) emerge out of solving conflicts.
4. In the planning process, actors determine the acceptability of a proposed course of action by estimating their share in the costs and benefits of the resulting cooperative activity.

In DAI, a similar approach, i.e. Partial Global Planning (PGP) (Durfee & Lesser, 1991) is taken by addressing the resources (including goals and skills) as nodes.

> [PGP is] a flexible approach to coordination that does not assume any particular distribution of subproblems, expertise, or other resources, but instead lets nodes coordinate in response to the current situation. Each node can represent and reason about the actions and interactions of groups of nodes and how they affect local activities. These representations are called **partial global plans** (PGPs) because they specify how different parts of the network plan to achieve more global goals. Each node can maintain its own set of PGPs that it may use independently and asynchronously to coordinate its activities.
>
> - PGP avoids redundant work among nodes by interacting among the different local plans.
> - It schedules the generation of partial results so that they are transmitted to other nodes and assist them at the correct time.
> - It allocates excess tasks from overloaded nodes to idle nodes.
> - It assumes that a goal is more likely to be correct if it is compatible with goals at other nodes.
>
> (Decker, 1995, pp. 33–34)

Thus, the first step is that each agent—able to explicitly represent its goals and actions—constructs its own local plan. The agent has no knowledge of other agents' plans and first needs to create its own plan and uses scheduling techniques in order to clarify its actions—when, what, where, etc.—to others. The next step is to communicate and exchange information between agents, followed by a reordering of the local plan and storing of (intermediate) agreements or results in the partial global plan. This process can be iterated until the joint goal is solved or desired situation is reached. The explicit modelling of teamwork is another approach for creating a joint goal that needs to be solved and is discussed in the next section.

2.5.2.2.3 Teamwork models

Teams exist out of a (temporary) collection of agents that have interdependencies. These interdependencies occur when local decisions of agents influence decisions of other agents in the team or there is a possibility of conflicting interests between team members. In Distributed AI, problem solving agents work together in order to solve their local goals and the goals of the community as a whole (Jennings, 1995). Each individual agent can be assigned its private beliefs, desires and intentions and can communicate with other agents. Assume a purely self-interested agent that acts in isolation and tries to optimise its own local performance with no concern about the overall outcome of the process, e.g. a sales-agent sells and promises more than a production agent can ever produce. To prevent such situations, there is a need for coordination. Coordination is necessary when there are dependencies between agents' actions, when there is a need to meet global constraints and when an individual does not have sufficient competence, resources or information to solve the entire problem (Jennings, 1996).

Cohen and Levesque (1991) discuss teamwork and coordination, i.e. what is involved in doing something together. They extend the model of the agent with a belief-goal-commitment model of the mental states of individuals (Cohen & Levesque, 1990)). The extended model specifies intentions as *internal commitments* to perform an action. A team is considered as a collection of agents to which the mental state of joint intention can be assigned: a joint intention is a joint commitment to perform a collective action while in a certain shared mental state, as the glue that binds team members together (Cohen & Levesque, 1991). The formation of a joint commitment is established when all members have mutual belief of a joint persistent goal[25].

Next, being part of a team implies a certain responsibility towards other members, i.e. when one of the members drops the individual commitment—it thinks the goal is finished, or the goal will never be satisfied, or the motivation for the goal is no longer present—then the agent has the persistent goal to notify other agents of this fact. Jennings (1995) states that not only agents must have a

[25]The model of Cohen and Levesque is at the rational level, which assumes that the collection of actors has a joint intention and does not take into account the cognitive mechanisms at lower levels. In the next chapter, we will approach the organisation as being a social construct and as a set of agents that have each their own representation of the organisation.

joint goal and the wish to achieve that goal, but also a common recipe for attaining the joint goal, see figure 2.7. The difference is that when the commitment to the goal is dropped, the joint action is over. However, when the commitment to the recipe of actions is dropped, the team still has the ability to come up with a new recipe of actions that could lead towards the joint goal. The *Joint Responsibility Model* is introduced as an explicit model that requires: "[...] all the team members to have a joint persistent goal to attain their joint goal X, that they all execute the common recipe Y according to the principles of joint recipe commitment, and moreover, that all the agents mutually know what they are doing whilst they are doing it" (Jennings, 1995, p. 209).

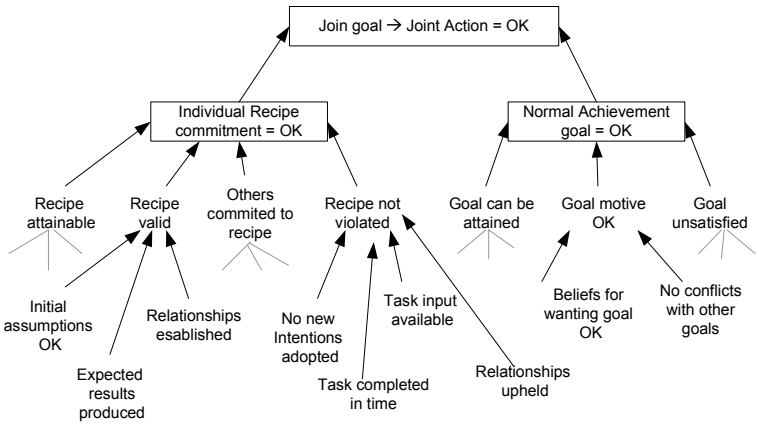

Figure 2.7: Causal network for commitment to joint goal (adapted from Jennings, 1995).

The Joint Responsibility Model is implemented in an architecture, the ARCHON (Architecture for Cooperative Heterogeneous ON-line systems) framework (Jennings & Pople, 1993; Wittig, 1992). This model is applied in industrial control systems such as GRATE* (Jennings, 1995), see figure 2.8.

The ARCHON architecture is a high-level architecture and functions as a cooperation module that can be placed on top of (in software terms: wraps) software legacy systems. The architecture distinguishes two separate layers, the *cooperation and control layer* and the *domain level system*. Within the cooperation layer, there are three main problem solving modules.

> The *control module* is the interface to the domain level system and is responsible for managing all interactions with it. The *situation assessment module* makes decisions which affect both of the other two modules. It decides which activities should be performed locally and which should be delegated, which requests for cooperation should be honoured, how requests should be realised, what actions should be taken as a result of freshly arriving information, and so on... The cooperation module is responsible for managing the agent's social activities, the need being detected by the situation assessment module. Three primary objectives related to the agent's social role are

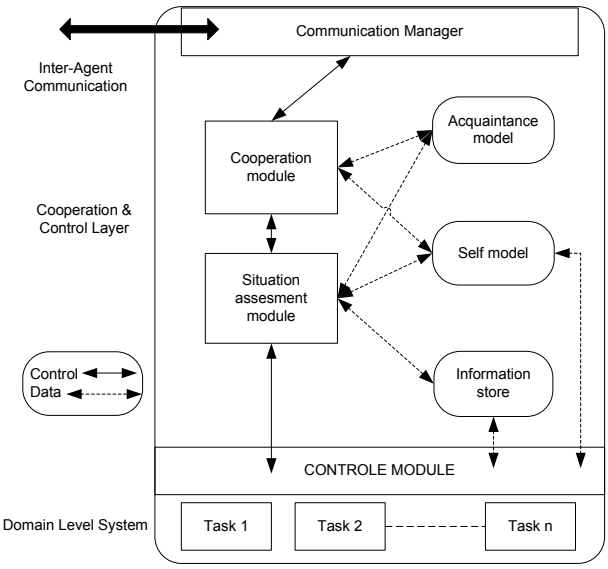

Figure 2.8: ARCHON (GRATE*) agent architecture (Jennings, 1995).

> supported. Firstly, new social interactions have to be established (e.g. find an agent capable of supplying a desired piece of information). Secondly, ongoing cooperative activity must be tracked (using the Responsibility convention). Finally, cooperative initiations from other agents must be responded to. (Jennings, 1995, p. 212)

The situation assessment module retrieves situation-dependent information from (1) the *acquaintance model*—skills, interests, currents status of others, (2) the *self model*—skills, interest of the agent self, (3) the information store that keeps up the results of the underlying domain level system, and (4) the results of interaction with others. The acquaintance model is a (mutual) modelling approach (see next section) that enables an actor to anticipate on actions of other agents and achieve quicker and better performance in coordination problems.

2.5.2.2.4 Coordination by mutual modelling[26]

Agents that have to survive in an environment must be equipped with perception, action and memory components that give the ability to learn and anticipate on new situations. Mutual modelling creates explicit models of other agents in the memory of the agent. For example, a kid constructs an intentional model of its parents in memory. Based on experience, it knows that its mother is more giving in to complaints than dad. So whenever it needs consolation or something to be done, the kid will pester its mum. On the other hand, dad has

[26]The reader probably notices that we are shifting from a top-down approach (DPS) towards the bottom-up approach (MAS). Although there is a distinction between both, they are often interweaved with each other.

a model of (the interaction of) the kid and his wife and knows that in order to prevent mum from giving in too easily, he will enforce his wife to be more strict with their kid. The father puts himself in the position of his wife allowing him to (indirectly) coordinate the actions of the child more effectively, i.e. he has an internal model of his wife—the beliefs, desires, intentions, and so on—to be able to cooperate and create predictions about the way his wife and his child behaves. Mutual modelling creates a better mutual understanding between agents. Jennings (1992) states that mutual modelling plays an important role in:

- *Focusing Activity and Reducing Communication*
  Agents with knowledge about other agents' intentions and capabilities are more effective and efficient in coordinating their actions with other agents. The amount of knowledge gathered about other agents creates the possibility to predict how other agents behave. The gathered information reduces the need for communication between agents and gives better directions to select the right actions.

- *Coordination and Coherence*
  Mutual modelling allows agents to maintain models of each other and prevent disturbing each other goals while attempting to achieve the joint goal, i.e. it lowers the probability of conflicts and improves coordination between agents. It also increases coherence or effectiveness with which the primitive actions in a complex action conspire to bring about a particular joint goal (Wooldridge, 1994).

- *Self Reasoning*
  When the agent is able to compare local capabilities with capabilities of others it can reason what tasks can be solved by himself and what tasks require social interaction, what information must be available to initiate a task, what results will be produced and when a particular activity will be finished.

The first application that started to test mutual modelling and MAS was MACE (Gasser, Bragamza, & Herman, 1987). MACE constructs a database of acquaintances and adapts the knowledge based on changes in the environment. The agent has a model in which it stores:

- Class: groups of agents are identified with a class name.
- Identification of the modelled agent by name, connected to a class.
- Roles: the function the modelled agent has in that class.
- Skills: skills that are connected to that role.
- Goals: the goals that the modelled agent wants to achieve.
- Plans: the means of the modelled agent to achieve the goal.

Whereas some of the knowledge stored of other agents concerns physical characteristics, much of the knowledge is referring to socially constructed artefacts created by human beings themselves. Social psychologists, such as Mead (1934) and Blumer (1969) (see chapter 3) indeed state that creating a mental representation of others is necessary in interaction with others. The coordination by norms and laws is closely linked to mutual modelling.

2.5.2.2.5 Coordination by norms and social laws

Coordination by norms and social laws is possible when agents can imitate and maintain models of other agents. Wooldridge (2002) states that a norm simply is an established, expected pattern of behaviour and that social laws exist out of norms to which authority and roles are assigned. Conventions or social constructs (see chapter 3) provide stable patterns of behaviour in otherwise unstable and uncertain situations. When applied in an agent society, two approaches can be distinguished. The first approach is the implementation of a fixed normative system that is designed and installed in the agent *off-line* (Shoham & Tennenholtz, 1995). The second approach (Shoham & Tennenholtz, 1992b) is the emergence of a convention between agents. Experiences with and about others are remembered and with help of a strategy or utility-function, the agent determines what type of behaviour will probably be most successful. Such behaviour becomes a habit of action when it is reinforced by all agents (mutual reinforcement) that perceive the behaviour as being successful. Mutual reinforcement of behaviour can eventually lead to the emergence of norms that prescribe what behaviour is desired in certain situations. In the establishment or spreading of norms, there can be made a distinction between reactivation of the same representations under the effects of a given perception (social-facilitation) and behaviour spread by social cognitive processes mediated by the agents' social goals and beliefs (imitation), i.e. exchange of social constructs (Conte & Paolucci, 2001).

In chapter 6, based on a traffic law system (cf. Kittock, 1993; Epstein, 2001; Shoham & Tennenholtz, 1995), two experiments demonstrate the spreading of norms. The first experiment demonstrates social facilitation or implicit change of cognitive preferences for certain rules, influenced by perception and reward. The second experiment demonstrates the explicit transfer of the mental representation of a norm by an authoritative agent to an other agent. In the next chapter, we will explain social constructs that serve as coordination mechanisms and hold knowledge about mental attitudes of others and social structures (e.g. authority and roles).

Although the field of MAS provides much more as can be shown in this chapter, we will end this section. Still, the reader is entitled to have some insights into applications that have emerged, especially with the growth and importance of the Internet, the first generation of applications has entered the market.

2.6 Multi-Agent System Applications

A large variety of Multi-Agent Systems are available on the commercial market and as open-source on the Internet. Agent technologies are not restricted to academia, companies and research labs anymore, but they have entered gradually into our computer systems and homes. In solving problems that concern interactions between people, the application of Multi-Agent Systems has proven that it is a good candidate, i.e.

> Multi-agent systems are ideally suited to representing problems that have multiple problem solving methods, multiple perspectives and/or multiple problem solving entities. Such systems have the traditional advantages of distributed and concurrent problem solving, but have the additional advantage of sophisticated patterns of interactions. Examples of common types of interactions include: cooperation (working together towards a common aim); coordination (organising problem solving activity so that harmful interactions are avoided or beneficial interactions are exploited); and negotiation (coming to an agreement which is acceptable to all the parties involved). It is the flexibility and high-level nature of these interactions which distinguishes multi-agent systems from other forms of software and which provides the underlying power of the paradigm. (Jennings et al., 1998, p. 9)

First, as mentioned before, Multi-Agent Systems consist of agents that have incomplete information, no centralised control, data is decentralised and computation is asynchronous. Secondly, the weak notion of agency states that agents are flexible and in general have the properties of autonomy, social ability, reactivity and pro-activeness. And thirdly, we have distinguished different types of agents, i.e. the cognitive and the social agent. Creating a classification of Multi-Agent Systems is a tedious task and we could take an approach that is based on the types of agents applied and their interaction with the environment resulting in a classification of reactive towards social pro-active systems. However, the difference in opinions and definitions about what agents are, automatically results in a problematic classification of MASs.

An approach by Wooldridge (2002) is to first draw the distinction between distributed systems—multiple agents are nodes in a distributed system—and agents that have the (pro-active) function of assisting users working with applications (Maes, 1994). Next, Luck, McBurney, and Preist (2004) divide MAS into (real-time) Multi-Agent decision systems—agents participate in a system have to make joint decisions, and Multi-Agent simulation systems, where MAS is applied as a model to simulate some real-world domain. The important distinction between both is that with decision systems, there is often a direct outcome ready for application in the real world, while simulation systems require first an interpretation of the researcher before results can be applied in the real world[27].

[27] We are aware of the fact that games or entertainment software are somehow in the middle, but most of the time, games have no direct impact on the world (e.g. someone shot in a movie does not create a real murder case).

Nevertheless, we follow the approach of Luck et al. by classifying systems in two groups and show in which domains MAS applications are developed. The domains can be structured varying from hardware, such as robotics and manufacturing towards software, e.g. simulation systems and entertainment. In the next sections, 2.6.1 and 2.6.2, we respectively mention examples of multi-agent decision systems and multi-agent simulation systems.

2.6.1 Multi-Agent decision systems

The *Industrial Application* is one of the domains where MAS needs to interact with the physical world, be it humans or other agents. Industrial manufacturing and (real-time) control systems in the form of MASs are introduced to solve coordination problems in factories, air-traffic control, telecommunication networks and transportation systems. The general goal of industrial applications is to manage processes efficiently. Examples of industrial applications are the manufacturing system YAMS (Parunak, 1995) and the process control system ARCHON (Jennings & Pople, 1993). Closely linked to manufacturing systems (primary processes) are applications for workflow and business process management (secondary processes). An example is the ADEPT project that views the business process "[]...as a collection of autonomous problem solving entities that negotiate with one another and come to mutually acceptable agreements that coordinate their interdependent sub-activities" (Jennings, Faratin, Norman, O'Brien, & Odgers, 2000, p. 2).

Industrial applications are mostly custom-made solutions for specific problem areas in the Business to Business market, while *commercial applications* are more directed towards the consumer market. Coordination of interdependent sub-activities, similar to ADEPT, can also be applied in information retrieval and sharing systems. A 'web' of agents that are experts in a certain area of interest form a connected network of information sharing nodes. Other agents function as brokers and have knowledge about where to go for what. In P2P applications, e.g. Gnutella, Bit Torrent and Skype, this functionality is already delivering music, video and voice over IP. Following up the information agent network, e-commerce is the next step in which (mobile) agents take over tasks from the user and 'walk' over the Internet to compare prices and find the most suitable deals with supplier agents. For instance, Chavez and Maes (1996), define a web-based multi-agent auction system (Kasbah) in which humans are represented by bidder and seller agents.

MAS also entered the entertainment and leisure industry, where agents play a part in computer games, e.g. the Creatures (Grand & Cliff, 1997), movies (Lord of the Rings, Massive Limited[28]) and virtual reality applications. Commercial applications of Multi-Agent decision systems probably will surround us in the near future and offer services that today can be spotted in Science Fiction movies, e.g. a car that is equipped with a diagnosis agent will automatically offer its service to the driver when the car needs maintenance; it can arrange an appointment with the garage that fits the schedule of the driver. Such a service

[28] http://www.massivesoftware.com

is established by multi-agent interaction between many agents in order to get the best quality combined with the lowest possible price.

The entertainment industry and application service providers have created a surrounding in which the physical world and the virtual world become more and more connected. We see the connection between the virtual and physical world also in the area of multi-agent simulation systems.

2.6.2 Multi-Agent simulation systems

Agent based simulation systems are used to describe, prescribe and predict real-world scenarios and are applied in many areas, e.g. traffic systems, crowd behaviour and riot control, combat scenarios and many alike. In this dissertation, the (cognitive) agent-based social simulation (Gilbert & Troitzsch, 1999; Sun, 2006a) is applied as a modelling tool for studying the behaviour of agents that represent individual people. In the field of agent-based social simulation, two broad approaches can be distinguished (Luck et al., 2004). One approach—the emphasis in this dissertation—defines systems that underlie social interaction, and the other focuses on observing social processes and modelling those. The combination of both approaches can be an iterative process, by first analysing the field, next creating a multi-agent simulation system and finally observing and validating the model.

An example of an application is given by Wooldridge (2002) who mentions a social system, the EOS project (Doran & Palmer, 1995) that consists of agents making it possible to describe a number of social phenomena, such as 'overcrowding'—when too many agents attempt to obtain resources in some locale or 'clobbering'—when agents accidentally interfere with each other's goals. Another well known example of a simulation system is Swarm (Minar, Burkhart, Langton, & Askenazi, 1996). Swarm is a generic discrete event model used for simulating a collection of independent agents of any kind. The areas in which Swarm already is applied is diverse, e.g. chemistry, economics, physics, anthropology, ecology, and political science.

Many simulation toolkits have evolved and the search for adopting a toolkit that is suitable for our current research has been difficult. Therefore, in the discussion of this chapter, we explain why we invest in constructing a new toolkit.

2.7 Discussion

The aim of this dissertation is the creation of multi-agent based social simulations[29] whose agents are grounded with help of a plausible cognitive architecture and are equipped with the capability of social construction[30] in order to express social or normative behaviour.

In this discussion, first we want to explain the type of agent that is suitable for our research purpose and in the second place, we want to discuss what agent

[29] Chapter 1, research question 1.1.

[30] Chapter 1, research question 1.2; see also next chapter.

toolkit would be appropriate for modelling a cognitive agent-based computational social simulation model that can explain interactive social behaviour.

In order for an agent to exhibit intelligent social behaviour, the agent should be equipped with a cognitive architecture[31], i.e. a physical symbol system that makes it cognitive plausible and allows it to exhibit intelligent action. Apart from that, the agent should also be able to socially construct its world and show social or normative behaviour, i.e. it should be physically and socially situated. The physical and social situatedness requires the capabilities of a reactive agent (embodied/situated cognition) that responds to changes in its environment and of a social agent that takes needs of other agents into account. Therefore, the type of agent applied in this dissertation should be a hybrid agent, existing out of the capabilities of a cognitive, a reactive and a social agent.

The first step in this research was to find a suitable cognitive agent or a suitable agent toolkit. This requirement restricted our options dramatically, because the amount of plausible cognitive architectures is low. The most well known cognitive architectures are SOAR (Newell, 1990) and ACT-R (Anderson & Lebiere, 1998). In chapter 4, we will explain what type of architecture we have adopted for constructing a cognitive plausible agent.

The next step was to see how to create physical and social situatedness in our cognitive actor. As discussed in section 2.4, three types of environments need to be supported by our simulation toolkit, i.e. the physical, communication and social environment. The hybrid agent that responds to physical and social changes in a simulated environment, requires (the development of) a simulated physical, communication and social environment that allows for the exchange of social constructs and norms. In the last step we added a social or normative layer to the cognitive agent (cf. ARCHON: Wittig, 1992) and applied subsumption (Brooks, 1986) that allows the agent to beware of changes in the social environment. These three steps resulted in a cognitive agent-based social simulation. The model and design of the complete (social and cognitive) agent, i.e. RBot (cf. Roest, 2004) and the simulated environment (Multi-RBot System/MRS) will be discussed in chapter 5.

In chapter 3 and chapter 4, we will study the individual agent, respectively the social and the cognitive actor, and will discuss what is required for an actor to exhibit *cognitive plausible social behaviour*. Chapter 5 will discuss the implementation of these requirements in a multi-agent model—RBot and MRS—that will provide us with a multi-agent platform for conducting demonstrative experiments in chapter 6.

[31] Chapter 1, research question 1.3; see also chapter 4.