Chapter 1

Introduction

"Freedom is not worth having if it does not include the freedom to make mistakes."
Mahatma Gandhi

This book is in the field of autonomous agents and multiagent systems. In this chapter, we give a motivational overview of the field and introduce the research objectives. There are two main research objectives addressed in this book (Section 1.1). The first objective is centered around the ideal of making software agents more autonomous by making them more flexible and adaptive. This looks for reasoning formalisms that incorporate uncertainty and dynamism in the world model without loosing the type of formal qualities that make BDI-like architectures so attractive for testability and reliability reasons. The second research objective addresses application of these autonomous agents to normative multiagent systems, an important motivation for this book. It focuses on ways to make autonomous agents social, capable of reasoning and deliberating about norms and forming sustainable agent societies. In the second section, we highlight the important contributions of this book. The first main contribution is the proposal of coherence-driven agents based on the cognitive theory of coherence as proposed by Paul Thagard [Thagard, 2002]. This includes a coherence framework with a formalisation of *deductive coherence* and a coherence-based architecture with a reasoning algorithm for coherence-driven agents. The second contribution is to model such agents as normative agents that are capable of reasoning about norms and modelling consensus on norm adoption. Finally we outline the organisation of the rest of the book in Section 1.3.

1.1 Motivations

Multi-agent systems (MAS) are a well-acknowledged methodology to model complex software systems and simulate intelligent behaviour mainly through interactions between autonomous entities having different information and/or conflicting interests. Research on Agents and MAS has matured during the last decade and many effective applications of this technology are now being deployed. Distributed healthcare management, e-commerce and e-governance, digital ecosystems, and entertainment and gaming are some of the emerging areas where autonomous agents and MAS are the natural technology of choice.

Some of the characteristic features that are shared by the above mentioned applications are:

1. they are composed of loosely coupled autonomous complex systems
2. they are realised in terms of heterogeneous components and legacy systems
3. they dynamically manage data and resources
4. they are often accessed by remote users and/or in collaboration

For example, a MAS for assisted cognition for elderly patients co-ordinate among various services such as monitoring, providing decision making and warning or reminder services. In its simplest form, such a system would be made up of a series of agents, like monitors and mobile robots capable of reminding, alerting and advising the assisted person. All the actors in the system would clearly be capable of carrying out individual reasoning, but would also need to collectively reason about the situations which can occur [Cesta et al., 2003].

However, the use of MAS at the deployment level is more for providing infrastructures to interoperate between different data formats, integrate different types of services, and unify information gathered from different sources. There is still a lack of technology readiness when it comes to applying MAS consisting of autonomous agents taking independent and autonomous decisions. For example, until recently agents modelling NPCs (Non Player Characters) in virtual worlds and online games [Aranda et al., 2008] have been painstakingly hardcoded by prethinking every potential encounter they might have in the course of the interaction. Fortunately, the situation is changing today and virtual worlds are seen as one of the most potential developing environment for introducing real intelligence in artificial agents. Their "relatively unsophisticated environment" makes it more practical to control and test the autonomous behaviour of artificial agents.

The increasing complexity of such systems and applications not only require that autonomous single agents become more and more intelligent and real, but groups of such agents most likely heterogeneous, interact and share information to achieve their individual goals, while also contributing to the collective goals of the system. For example, agents in mobile health management (providing health services to patients on the move) may need to share information and patient data, health care policy, and information on previous health history. They may also need to take into account rapidly changing national and international laws and regulations concerning the privacy of medical data and the security policies concerning transactions, may need to set up operational norms, and may even need to negotiate on some of the terms based on the specific needs and available services.

Hence in this book we explore two dimensions of agency, a cognitive dimension attempting to accomplish a more flexible and adaptive reasoning capability and a social dimension exploring normative reasoning and interactions in a regulated environment. In particular we try to identify and understand those characteristics that make autonomous agents and MAS suitable for the kind of applications mentioned above. These research problems are formulated in the next subsections.

1.1.1 Autonomous Agents

The use of agents making decisions and performing actions in real time while considering the effects of their actions and adapting to dynamic changes in the environment has increased significantly in the context of typical applications of MAS as discussed above. Such agents are alternatively called rational as in Wooldridge et al. [Wooldridge, 2000], autonomous as in [Maes, 1991] or intelligent as in [Russell and Norvig, 2003]. In this book, we use the term autonomous to represent such agents because we concentrate on the capability of the agent to make their decisions and actions without external intervention.

The BDI family of agent models originated from Rao and Georgiff are arguebly some of the most important existing models for designing such agents [Rao and Georgeff, 1995]. A BDI based reasoning process consists of a deliberative cycle in which an agent decides what state of affairs it wants to achieve from among all those desirable states of affairs [Dastani et al., 2003, Shoham, 1993, Rao and Georgeff, 1995]. A main aspect of BDI theory is that it helps selecting what action to perform at each moment. The model focuses on the role of intentions as they constrain the reasoning an agent is required to do in order to perform an action. Once a set of intentions are created and their associated preconditions (in the form of a set of beliefs) are met, then it is immediate that these intentions are realised. BDI models try to reduce the attention problem of an agent by providing an intention to focus on.

However, a key challenge for the BDI family of architectures in general is the need to formalise defeasible (non-monotonic) reasoning, and associated conflict resolution mechanisms. The BOID [Broersen et al., 2002] extension is designed specially for conflict resolution arising between some cognitive elements of an agent and its obligations. The BOID architecture characterises generated candidate goal sets as extensions of a prioritised default logic theory in which rules for inferring goals are modelled as defaults [Reiter, 1987], and a prioritisation of these defaults resolves conflicts between mental attitudes. However, in a BOID architecture, prioritisation on cognitive elements of agents to resolve conflicts is due to different agent types which are identified beforehand. For example, a selfish agent would always prefer goals generated from private desires than those from obligations. And a duty-bound agent would prefer the opposite. This, in our opinion, is not an efficient conflict resolution mechanism because such a mechanism should ideally take into account dynamic changes in a situation and possibly changes in cognitive elements of the agents.

Another way of resolving conflicts or choosing from competing cognitive elements is by introducing preferences as in the graded BDI model (henceforth referred to as *g-BDI*) proposed in [Casali et al., 2005]. The motivation in this work stems from an assumption that an agent's model of the world is incomplete and uncertain. Introduction of degrees is an attempt to capture and represent this uncertainty better in an agent's model. Using a g-BDI model reduces the ambiguity in selecting among the intentions since the degree of an intention is interpreted as its preference or priority and a higher degree implies a higher priority. However, one of the main problems of the BDI family of models is that they follow a linear reasoning structure. That is, an agent choses one or more desires to satisfy and then looks for intentions or plans to realise these desires, thus failing to evaluate desires and intentions in the context of other cognitive elements put together.

Another growing body of work in this context is the literature on argumentative agents that attempts to introduce defeasible reasoning models [Atkinson, 2005a, Amgoud et al., 2000, Modgil, 2008]. An argumentative agent does not reason with basic cognitive elements such as beliefs, desires or intentions, but with arguments computed from these cognitive elements. An action or an intention is selected from a set of intentions based on arguments that support the action. Hence, an action that is supported by the winning argument is chosen as the next action to pursue. Argumentative agents overcome some of the limitations of the BDI family of agents since arguments are generated considering the entire knowledge base of an agent and moreover they are defeasible and hence conflicts among cognitive elements are discovered in the process of constructing arguments that attack or defeat existing arguments.

Most argumentation systems instantiate the general framework of Dung that starts with a set of arguments and binary defeat relations and then determines the set of arguments that can be accepted together [Dung, 1995]. In some of them, tree structured instrumental arguments are composed by chaining the propositional rules with the top of the tree as the high level goal and leaf nodes as primitive actions. A set of instrumental arguments are chosen from sets of conflict-free instrumental arguments that maximise the set of agent goals realised. And some of them further include a preference relation among instrumental arguments based on the value or utility which roughly characterises the worth of the goal and its cost of realisation. A given ordering on values advanced by arguments then determine defeats among arguments [Atkinson, 2005a]. Some of these proposals also include a formal construction of the arguments in an underlying BDI type logic.

One important limitation of argument-based systems is that they tend to be very brittle by demanding conflict-free sets of arguments to be accepted as support for a goal or an action. Whereas in reality, it may only be possible to reduce conflicts but not eliminate them all together. Further, most realisations of argumentation logics only have a binary form of attack relations and are not suitable for modelling uncertainty, though this trend is changing in recent systems. Another limitation that argument-based systems share with BDI-based approaches is that their reasoning progresses in a linear fashion starting from selecting a goal or a set of goals to realise and then choosing instrumental arguments that support the goals. Alternatively, to resolve conflicts, and more importantly to select among the set of goals, beliefs and intentions, we believe an agent should look at all the relevant information it possess and then should evaluate which subset is more conflict-free from a global perspective.

To summarise, the main limitations of the above approaches are:

- There is a lack of clear cut methods by which some desires are promoted to the level of intentions.
- Even when methods exist, they do not take care of any potential conflicts that exists among desires or among other cognitive elements.
- Most discussed methods are not dynamic in readjusting to new or changed information.
- While the argument-based systems are the most dynamic since they depend on arguments which are constructed on the fly, the values which they use to resolve conflicts are decided a priori.
- None of the methods discussed above select cognitive elements that are most conflict-free from a a global perspective.
- All methods discussed follow a linear reasoning structure starting from a set of beliefs to chose among a set of desires and finally arriving at a set of intentions that realise the set of desires.

Given that, the current state of the art does not fully address the issues we have raised here, we put forward the following research objectives:

> To establish a suitable framework to model autonomous reasoning in agents that can incorporate uncertainty and dynamism in the agent's world model and is capable of resolving conflicts while not loosing the type of formal qualities such as testability and reliability.

This objective may be decomposed into sub-objectives. The first sub-objective is to find a formalism to design an autonomous agent. The idea is to look along the lines of BDI and argumentation logic while overcoming those limitations discussed previously. For example, unlike the intention-driven philosophy in a BDI logic, we need a formalism which would dynamically select intentions based on a global constraint maximisation. The second sub-objective is to define an agent architecture based on the defined formalism. This should further include a reasoning procedure for agents modeled with this formalism. The third sub-objective is to prove that the proposed formalism and architecture when implemented models an autonomous agent with the discussed properties. Concisely, the three sub-objectives are the following:

1. to find a formalism to model autonomous agents that are capable of resolving conflicts under dynamic and uncertain scenarios.

2. to define an agent architecture based on the defined formalism along with an agent reasoning algorithm.

3. to show that the defined architecture models autonomous agents with the specified properties.

1.1.2 Autonomous Normative Agents and Normative MAS

An interesting mechanism to co-ordinate the interaction of autonomous agents within a MAS is by making use of norms. Norms while prescribing the accepted behaviour of agents also respect agent autonomy on norm compliance. There is an increasing interest in norm regulated MAS in the computer science community, due to the observation in the AgentLink Roadmap [Luck et al., 2005]—a consensus document on the future of multiagent systems research—that norms must be introduced in agent technology in the medium term for infrastructure for open communities, reasoning in open environments and for trust and reputation. Since then an active community of researchers evolved focusing on norms and normative aspects of MAS. Based on a series of workshops, a consensus evolved as to what can be considered as a norm regulated MAS (referred to as a normative MAS). We quote here one of the definitions most aligned with the perspectives of this book.

> A normative multiagent system is a multiagent system organized by means of mechanisms to represent, communicate, distribute, detect, create, modify, and enforce norms, and mechanisms to deliberate about norms and detect norm violation and fulfilment.

It was remarked in Section 1.1.1 that autonomous agents should be equipped with an effective conflict resolution strategy. This is particularly relevant for autonomous agents situated in a normative MAS (hence forth will be referred to as *autonomous normative agents*) where conflicts among intentions motivated by private goals and those motivated by norm compliance are prevalent. There have been many attempts in the recent past to design agents that could handle such conflicts effectively [Moses and Tennenholtz, 1995, Conte et al., 1999, Boella et al., 2006, Pasquier et al., 2006, López et al., 2002, Kollingbaum and Norman, 2003, Noriega, 1997]. Many of these efforts are focused towards extending the cognitive agent theory (for instance BDI theory) with explicit representation of norms (BOID [Broersen et al., 2002], EMIL [Conte et al., 1999], and NoA [Kollingbaum and Norman, 2003]). However, the kind of conflict resolution strategies employed in most of the current literature limits to prioritising statically among norms and private goals of an agent. That is, a norm priority agent will always prefer norm compliance over satisfaction of private goals when there is a conflict. Hence, it is necessary to extend the features discussed for autonomous agents to autonomous normative agents.

In addition, an autonomous normative agent may need to participate in the set-up or adaptation of norms. This means an agent may need to generate norm proposals, reason about norm proposals of others, and deliberate to reach consensus on norms. In the literature, norm generation and normative agreement are fairly new areas of research and there are no prominent methods so far. However, norm generation is similar to intention generation by an agent that reasons about how to achieve its goals, while normative agreement is similar to reaching agreement on a course of action to solve a problem. For both phenomena logic-based argumentation models have been proposed [Bench-Capon and Prakken, 2006, Amgoud and Prade, 2009] most of which instantiate the general framework of Dung [Dung, 1995]. As discussed in Section 1.1.1, argumentation systems based on Dung's abstract argumentation framework do not take into account uncertainty in the world model of agents and cannot accommodate inconsistency in an accepted set of arguments. Since generating arguments and support for arguments are at the core of a deliberation process to agree on norm proposals, the argumentation system needs to be flexible and expressive.

Hence, the second part of the book deals with autonomous agents and their interaction in a normative MAS. In particular, we care about designing agents that can interact autonomously in a normative MAS by means of an argumentative process deliberate about norms. By this, we emphasize the fact that we not only are concerned with making autonomous normative agents, but are looking at ways to make a normative MAS sustain and adapt over changing situations. As discussed earlier, such agents and systems that adapt are necessary to most MAS applications. Hence, the research objective in the context of autonomous normative agents and normative MAS is the following:

> To design autonomous normative agents and to design a mechanism for such agents to interact and together form sustainable normative MAS.

This can be decomposed into two sub-objectives as follows:

1. To design normative autonomous agents that can
    - reason about norms autonomously,
    - generate norm proposals, and
    - reason about norm proposals of other agents.
2. To design a mechanism for autonomous agents to deliberate about norm change in a normative MAS.

1.2 Contributions

The two main contributions of this book are a proposal of coherence-driven agents based on the cognitive theory of coherence [Thagard, 2002] and a coherence-driven argumentation system for such agents to deliberate about norm adoption. In this section we briefly go over the arguments that make coherence an interesting and suitable theory for the kind of agents and MAS discussed in this book.

1.2.1 Autonomous Agents

Some of the properties we would like to have in autonomous agents are the ability to reason taking into account global constraints and the ability to adapt to situational changes (Section 1.1.1). One of the primary factors that facilitate this is a suitable representation of the cognitive elements. In a BDI architecture, cognitive elements are represented in isolation with no explicit reference to possible constraints among them. Such a representation makes it hard to keep track of constraints and more important incorporate them into a reasoning process. Ideally, a goal should be pursued that is not only most desired, but also that has least conflicts with other goals (that may be already in pursuit), other plans and beliefs. A similar process should be followed to incorporate new perceptions and pursue plans. Coherence-driven agents facilitate such a reasoning process with the architecture representing not just cognitive elements but any positive or negative constraints that exist between pairs of elements. Hence, with this architecture, it is possible to maximise satisfaction of constraints at a global level by a process of *maximisation of coherence*. In addition, the effects of dynamic changes in situation are understood by simply updating the cognitive elements in the agent's theory and re-computing satisfaction of constraints.

The coherence-based architecture we propose in this book is inspired by the *theory of coherence*. According to this theory, there are coherence and incoherence relations between *pieces of information* depending on whether they support each other (yielding a positive constraint) or contradict each other (yielding a negative constraint). If two pieces of information are not related, then, there is no coherence (constraint) between them. Based on the characterisation of Thagard, we propose a coherence framework consisting of a *coherence graph* and certain computable functions operating on the graph. A coherence graph consists of nodes to represent the pieces of information and weighted edges to represent constraints between them. Given such a coherence graph, Thagard defines a mechanism to compute the overall coherence of the graph based on maximising constraint satisfaction between pairs of nodes. Certain principles are also defined to characterise and differentiate various types of coherence relations that might exist between pairs of pieces of information. Using the principles of deductive coherence, we define a *deductive coherence function* to compute deductive coherence between pairs of pieces of information of a coherence graph.

We then propose a coherence-based architecture based on the coherence framework. For this, we extend the popular BDI agent architecture with the notion of coherence. By so doing we move away from the intention-driven philosophy of the BDI architecture while retaining the logical properties of the cognitions. Coherence is introduced as the central motivational drive for agents and intentions in a coherence-driven agent are chosen based on the coherence maximisation of the agent's cognitive elements. Finally, we have evaluated the feasibility of our proposal with empirical analysis and compared it to performance of humans and near optimal algorithms in a restricted setting.

Thus, the main contributions in this book in the field of autonomous agents are the following:

1. Formalisation of a coherence framework based on Thagard's theory of coherence.

2. Definition of a coherence-based agent architecture for autonomous agents consisting of an algorithm for coherence-driven agent reasoning.

3. Empirical evaluation of coherence-driven agents.

1.2.2 Autonomous Normative Agents and Normative MAS

The very arguments for coherence to be used in modelling autonomous agents may be extended to the case of autonomous agents with normative capabilities. As argued in Section 1.1.2, conflicts among cognitions are more likely when goals due to norms conflict with personal goals. Due to its representation and global maximisation of constraints, a coherence-based framework lends itself naturally to discovering conflicts. Hence we extend the coherence-based architecture to autonomous normative agents by introducing cognitive elements corresponding to norms in addition to those corresponding to beliefs, desires and intentions. For deliberation on norm adoption, we build upon an argumentation system. We choose argumentation technology since it has emerged as one of the most promising processes for multi agent deliberation with minimal assumptions on the initial positions of the agents, the common knowledge they share, the type of dialogue they engage in, or their motivations [Rahwan et al., 2003b]. In the proposed argumentation system, the notion of an argument consists of a *claim* and its *support* where support is defined in terms of coherence. Since agents are motivated by coherence, it is natural to compute a coherence-driven support. Each agent in a deliberation also evaluates an argument based on a coherence maximisation incorporating the argument into its coherence graph. Unlike traditional argumentation systems, such an argument incorporates degrees of support, and resulting argumentation systems are more tolerant to inconsistencies among arguments.

Thus the main contributions in this book in the field of autonomous normative agents and normative MAS is the following:

1. Definition of a coherence-based architecture for autonomous normative agents

2. Definition of an argumentation system based on coherence for deliberation on norm adoption.

1.2.3 Other Contributions

There are two contributions not directly intended nevertheless important in the field of artificial intelligence, cognitive science and economics. The first is the logical formalisation of the cognitive theory of coherence. We have analysed coherence formally, studied its logical properties and proposed a precise computable function to build a coherence graph. This is useful not only to build coherence-driven agents, but also for experiments in physiology and cognitive science and thereby making accessible the use of coherence to a wider audience.

The second contribution is our analysis of coherence in the context of other rationality theories. The game-theoretic concept of Nash equilibrium is one of the better known performance criteria to analyse strategic interactions amongst decision makers [Fudenberg and Tirole, 1991]. However, a number of assumptions make the concept of Nash equilibrium less useful in the context of autonomous agents. Firstly, it is defined for interactions among rational agents where rationality is often interpreted in the neo-classical economic sense of strict utility maximisation. However, strict utility maximisers are just one type of agents and, we need to be able to model different types of agents. Secondly, the concept of Nash equilibrium is developed only for situations where agents have perfect information and common knowledge about the utilities of outcomes of all agents involved. In most cases, these two assumptions do not hold for autonomous agents. A third assumption is that utility maximisation assumes a given ordering of preferences and most often also assumes that this ordering remains static during the interaction. However, a preference ordering of outcomes is a result of maximisation of satisfaction of multiple constraints that exists among an agents cognitive elements. Consequently, a preference ordering should ideally reflect the changes in the knowledge base of an agent, that, unfortunately, can neither be assumed nor remain static.

We in this book prove that coherence maximisation can emulate the properties of a utility maximising function, while getting rid of the strong assumptions that makes utility maximisation less useful. This we see as the first step in having theory of rationality that is more general than the economic notion of strict utility maximisation.

1.3 Organisation of the Thesis

This book is organised in four parts discussing each of the four components of the book. Below we give the organisation of these parts into chapters and briefly introduce their contents.

Part I contains two chapters including the present chapter which introduces the motivation for this book. Chapter 2 discusses those theories and research findings that serve as the base for the work on this book. Emphasis is given to introducing Thagard's theory of coherence which helps the reader to understand the basic notions of coherence and how it differs from other related theories. It also compares and contrasts the theory of coherence with some of the important related advances in the field.

Part II is organised in four chapters and addresses the first research objective of this book. Two chapters (Chapter 3 and Chapter 4) focus on *finding a formalism to model autonomous agents that are capable of resolving conflicts among cognitions and norms under dynamic and uncertain conditions.* In Chapter 3, we introduce a generic coherence framework, which can be used to create coherence-driven agents. We discuss in this framework how pieces of information can be organised in the form of a graph, along with the necessary computable functions to evaluate and maximise the coherence of such a graph. We then specialise the formulation for a particular type of coherence, namely deductive coherence. We derive a deductive coherence function based on the deduction relation of a logic, however the function we define is independent of the underlying logic. In Chapter 4, we introduce a proof-theoretic characterisation of coherence focusing on deductive coherence. We discuss the formal properties of coherence, and illustrate how these properties help us to derive coherence values between pieces of information.

Chapter 5 focuses on defining an agent architecture based on the coherence framework defined in previous chapters. In particular, we define a coherence-driven agent as a cognitive agent whose utility maximisation is achieved by coherence maximisation. For this purpose we define certain specific graphs corresponding to a cognitive agent. We adapt concepts from multi-context systems so that a coherence-driven agent can reason with its cognitions. We later sketch a procedure an agent may follow in the context of an action selection problem.

In Chapter 6, we prove experimentally the feasibility of a coherence-driven agent and analyse its performance. In particular, we prove the hypothesis that *the performance of a coherence-driven agent is indistinguishable or comparable to the performance of humans and near-optimal algorithms tuned for a specific application.*

Versions of Chapters 3, 4, and 5 have been published in [Joseph et al., 2008b, Joseph et al., 2009b, Joseph et al., 2008a]. Preliminary ideas on Chapter 6 has been published in [Joseph et al., 2010].

Part III is organised in three chapters and discusses the second research objective of this book. In Chapter 7, we define an *autonomous normative agent* and discuss an extension of coherence-based architecture to autonomous normative agents. We also focus on norm generation and evaluation aspects of these agents. Chapter 8 proposes an argumentation system for norm deliberation among coherence-driven agents. We focus on a deliberation protocol and the conditions under which coherence-driven agents reach consensus on norms.

Chapter 9 takes a step back to analyse the kind of agents and applications for which a coherence-based model is interesting. We place coherence in the context of other rationality theories and argue in favour of coherence to play a key role in the design of rational agents. In particular, we prove that coherence maximisation can emulate the functionality of other utility maximising functions.

Chapters 7 and 8 have been published in [Joseph and Prakken, 2009]. Chapter 9 has been published in [Joseph et al., 2009a].

Part IV concludes the book by discussing the main contributions. We also point to some of the more relevant future work which advances the research initiated in this book. We conclude the book by providing certain insights into the type of applications for which we would like to use coherence-driven agents. We do so by analysing the reasoning of coherence-driven agents in a real-world scenario where a few southern regions of India deliberate on sharing water.