Computational Intelligence, Volume 17, Number 4, 2001

SOCIAL MENTAL SHAPING: MODELLING THE IMPACT OF SOCIALITY ON THE MENTAL STATES OF AUTONOMOUS AGENTS

PIETRO PANZARASA AND NICHOLAS R. JENNINGS
Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK

TIMOTHY J. NORMAN
Department of Computing Science, University of Aberdeen, Aberdeen AB24 3UE, UK

This paper presents a framework that captures how the social nature of agents that are situated in a multi-agent environment impacts upon their individual mental states. Roles and social relationships provide an abstraction upon which we develop the notion of *social mental shaping*. This allows us to extend the standard Belief-Desire-Intention model to account for how common social phenomena (e.g. cooperation, collaborative problem-solving and negotiation) can be integrated into a unified theoretical perspective that reflects a fully explicated model of the autonomous agent's mental state.

Key words: multi-agent systems, agent interactions, BDI models, social influence.

1. INTRODUCTION

Agent-based computing is rapidly gaining acceptance as a pervasive and powerful model for analysing, designing and implementing a wide range of software systems (Jennings 2000). The paradigm is based upon the notion of an agent as an autonomous, internally-motivated entity that is situated within a dynamic and not entirely predictable environment from which it receives perceptual inputs and to which it effects changes by performing actions (Franklin and Graesser 1997). The fact that agents are autonomous means they have a high degree of self-determination: they decide for themselves when and under what conditions their actions should be performed. Despite this self-determination, however, agents are often required to attain goals that are only possible, made easier or satisfied more completely by interacting with other, similarly autonomous, agents. In this context, "interaction" is used as a generic term for a wide range of inter-agent social behaviour and joint activities, such as cooperation (working together to achieve a common objective), collaborative problem-solving (going together through the search space of a problem in order to find a joint solution) and negotiation (coming to a mutually acceptable agreement on some matter). All of these forms of interaction are united by the fact that they are inherently intertwined with processes of social influence that agents try to exert upon each other's decision-making and behaviour. For example, agents may influence their acquaintances by trying to make them endorse some new beliefs or goals, or perform some new individual or inter-agent activity or abandon/modify some existing activity. In short, it is the existence of an influence process that lies at the heart of sociality in agent-based systems. Whereas the behaviour of an asocial agent is entirely determined by its internal drivers and their interplay with the physical environment as seen through its percepts, the behaviour of a social agent can additionally be influenced by the social interactions in which it is, or it could be, engaged.

Although the process of influencing represents the very essence of social interaction, it has received surprisingly little attention to date. Particular forms of influence have

Address correspondence to Pietro Panzarasa at the Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, United Kingdom. e-mail: *pp@ecs.soton.ac.uk.*

been studied in the context of specific types of interaction (e.g. cooperation (Grosz and Kraus 1996; Jennings 1995; Tambe 1997; Werner 1989), coordination (Barbuceanu 1997; Durfee 1999), collaborative decision-making (Panzarasa et al. 2001a) and negotiation (Kraus et al. 1998; Faratin et al. 1998)), but, to date, there has been no systematic analysis of the generic process *per se*. This is a serious shortcoming and one that this paper seeks to rectify. To this end, a conceptual framework is presented that indicates how the process of influencing works for a broad class of agent models. Specifically, consideration is given to deliberative agent architectures (Wooldridge and Jennings 1995) whose decision-making can be viewed in terms of beliefs (what the agent knows about its world), desires (what the agent wants) and intentions (what the agent is actually committed to doing). Canonical examples of such Belief-Desire-Intention (BDI) architectures include IRMA (Bratman et al. 1988) and PRS (Georgeff and Lansky 1987). For such architectures, the mechanisms and structures that enable the social nature of agenthood to impact upon individual decision-making behaviour are identified and characterised. Since the focus is exclusively on cognitive agents that are conceptualised using an intentional stance (Dennett 1987), the term *social mental shaping* (Panzarasa et al. 1999) is used to more accurately denote the process of influence in this case.

In providing a clear conceptual model for how the social nature of agents impacts upon their individual decision-making behaviour, this work extends the state of the art in two ways. Firstly, it identifies the fundamental structures and mechanisms that underpin a wide variety of social interactions in multi-agent systems. Thus it is shown how the notion of social mental shaping can be used to characterise common forms of joint behavioural processes such as cooperation, collaborative problem-solving and negotiation. By identifying the common ground, this paper provides a unifying framework for expressing and analysing the disparate work on formal models of social agency that has taken place to date. Secondly, by interfacing models of individual and social behaviour, social mental shaping represents a step towards the goal of producing a comprehensive account of collaborative decision-making in multi-agent systems (Panzarasa et al. 2001*a*). In this respect, social mental shaping can be regarded as the fundamental socio-cognitive process that underpins the relationship between micro and macro practical reasoning and decision-making within a multi-agent environment.

The remainder of the paper is organised as follows. Section 2 shows the inadequacy of the standard BDI model when it comes to sociality. Section 3 informally introduces the notion of social mental shaping and the key concepts of roles and social relationships that are central to our model. Section 4 presents the formal framework in which the model of social mental shaping is expressed (a complete specification of the underlying formal language is given in the Appendix) and Section 5 gives a detailed formalisation of the model. Section 6 introduces a number of axioms that ensure a degree of consistency in an agent's mental state in the face of competing social influences. Section 7 shows the use of social mental shaping in a number of key types of social relationship. Section 8 describes an experiment in which this notion is applied to a real-world economic domain. Finally, Section 9 situates this work within the related literature, and Section 10 concludes and indicates avenues for future work.

2. THE SOLIPSISTIC NATURE OF EXISTING BDI MODELS

The BDI model has been widely advocated as a conceptual framework for building intelligent agents (e.g. Bratman et al. 1988; Georgeff and Lansky 1987; Mueller 1996) and many researchers have attempted to capture its key properties in formal models

(e.g. Cohen and Levesque 1990; Rao and Georgeff 1991; Singh 1995; Wooldridge 2000). However, all of this work concentrates on an individual agent's decision-making processes; there are no explicit structures for modelling social activities. The main reason for this limitation is that the mental states of BDI agents are usually viewed as constituted only by *internally motivated* attitudes.[1] To overcome this, and provide a more complete model of an agent's behaviour, BDI models need to be extended. In particular, the mechanisms that capture the social nature of agents and their ability to interact with one another need to be defined (Carley and Newell 1994). To this end, the first issue to address is how the social nature of agents can impact upon their individual mental states. Particularly, the following questions must be answered:

- How is an agent's mental state influenced by the multi-agent system in which it is situated?
- What is the relationship between the mental states of interacting agents?
- What are the conceptual mechanisms that enable these relationships to be effected?

Recently, a number of theoretical models have been proposed for extending the BDI framework to deal with some facets of the questions raised above. In particular, Castelfranchi (1995) introduces the notion of social commitment, defined as one agent's commitment to another to performing a certain action. Castelfranchi then goes on to argue that this notion is not reducible to the internal commitments of individual agents or groups of agents, and is thus able to capture the dependence relationships between agents. Cavedon et al. (1997), on the other hand, use the notion of a social attitude to reflect commitments, persistent over time, to performing a certain action. Such attitudes are defined as ternary dependence relations between two agents (or teams),[2] with respect to some state of affairs. For example, in the logical framework of Cavedon et al., $SCOM_I(a_i, a_j, \phi)$ means that agent a_i is Intention-committed to agent a_j with respect to ϕ (there are also analogues for beliefs and desires). These social attitudes capture interactions between agents: specifically, they convey the idea that an agent adopts a mental attitude because it is committed to another agent.

While social commitments and social attitudes capture some aspects of the relationships between agents (or teams), they suffer from a number of shortcomings. Firstly, the social nature of agents can impact upon their individual mental states even without any form of social commitment between agents to performing actions. For example, an agent may be persuaded to adopt a mental attitude of another agent's (Kraus et al. 1998; Parsons et al. 1998). Persuasion is a mode of social influence clearly grounded upon some form of social interaction, but the agent who adopts a mental attitude as a result of being persuaded by another was not necessarily committed *a priori* to that other agent. Indeed, the former may decide to change its mental state merely on the basis of its confidence in the latter.

Secondly, neither concept says anything about how an agent's mental state may be influenced whenever it takes on a role in a multi-agent system outside of any particular social relationship with another agent. Sociality, in its basic forms, is reflected not only by actual interactions between agents, but also by potential ones. These potential interactions between agents may be described and operationalised through the web of interconnections that exist between roles within the role structure of a multi-agent system

[1]Internally motivated mental attitudes are those that an agent would generate for itself if it were completely unaffected by its social setting. This contrasts with *socially motivated attitudes* that an agent is led to adopt or to keep by the very fact that it is situated in a system that contains other agents.

[2]Cavedon et al. (1997) describe teams as single (higher-order) agents endowed with *primitive* social attitudes.

(see Section 3). A role, by its very nature, is fundamentally a social construct: its enactment implies the enactment of other roles (because all roles are related to at least one other). Adopting a role may have an impact upon an agent's mental state simply because the role implies a set of expectations of the behaviour of the agent. No actual social relationship between two or more particular agents is necessarily involved in such role-based sociality. Still, the agent's mental state is influenced by its role-based attitude to eventually establish social interactions with others in related roles. For example, an agent that has adopted the role of secretary may be expected to adopt the goal of fulfilling the boss's goals whenever required. This can be the case even when there is no specific agent occupying the role of boss (i.e., the role is uninstantiated). Thus, when no actual social relationship is involved, the secretary's adoption of the new goal of achieving the boss's goals when required can be regarded as simply driven by the adopted secretary role, regardless of the particular agent that will eventually take on the boss role.

Finally, the social nature of agents can impact upon their mental states even outside of any actual or potential role-based social relationship. To the extent that agents have social capabilities and can reason about and represent other agents in intentional terms (Dennett 1987), their mental apparatus can be changed and/or complemented simply by internalising others' mental attitudes. This may happen outside of any actual or potential social relationship between the agents involved. For example, an agent may be induced to adopt another agent's goal, without the latter's being aware of any social interaction with the former, and regardless of the specific role the former occupies.

In the light of the above observations, neither social commitment, nor social attitudes capture the whole impact that the social nature of agents has on their mental states and behaviour. Given this, an attempt will be made in this paper to overcome these limitations by providing a unified framework in which, at a more general level, the influence that sociality may have on autonomous agents' mental states can be investigated. There are two components to this framework. Firstly, a representation and formalisation of the agent's decision-making apparatus through a relatively standard BDI model (Cohen and Levesque 1990; Panzarasa et al. 1999, 2001a; Parsons et al. 1998; Rao and Georgeff 1991; Singh 1995; Wooldridge 2000). Secondly, our framework is enriched by a characterisation of the structures, mechanisms and principles that interface to the agent's decision-making apparatus, and that modify it in the light of social interactions. The notions of roles and types of relationships between roles are used as the key abstractions in this framework. Informally, a role can be viewed as a kind of task description (Handy 1993; Huczynski and Buchanan 1991) and a relationship type as a link between two (or more) roles (see Section 3). For example, Prime Minister, professor and secretary can all be viewed as roles and the links between them as relationship types or abstractions. The fact that the Prime Minister and professor roles will typically involve the exercise of some control over a secretary role gives rise to two different relationship types. These types can then be variously instantiated by different agents, bringing about differing social relationships. We choose these notions as our key abstractions because they provide a high-level and neutral means for modelling sociality. Furthermore, roles and their relations can be easily expressed in a logical language, and this represents a first step towards an in-depth formalisation of the interplay between sociality and agent intentionality (Dennett 1987). The next section will informally introduce the notion of social mental shaping and concentrate on defining roles and social relationships more fully. Sections 4 and 5 will bring these concepts together with the individual decision-making apparatus to provide a unified picture of a social agent.

3. INTRODUCING SOCIAL MENTAL SHAPING

Social mental shaping is concerned with the conceptual mechanisms that describe *how* an agent's beliefs, desires, goals and intentions are influenced by its social nature. It is not concerned with the issue of *why* and *when* agents are influenced to adopt a particular mental attitude. Given this, it is our view that an agent may be socially motivated to adopt or to keep a mental attitude either by the role(s) it takes on, or by other agents with which it has social relationships, or simply by other agents outside of any social relationship. In this respect, the social environment (i.e., roles and other agents within or outside social relationships) plays an active causal role in governing an agent's behaviour, in the same way that the agent's internally motivated mental attitudes usually do. An agent's mental state, therefore, needs to be broadened to account for the impact that roles and other agents may exert by providing either new socially motivated mental attitudes or new reasons for maintaining individually motivated attitudes. To do this, the mechanisms that enable such a broadening effect must be identified.

A central idea that aids the examination of the nature and functioning of social mental shaping is that of *multi-agent system structure*. This notion refers to the way in which the members of a multi-agent system relate to one another (Huczynski and Buchanan 1991) and represents one of the basic aspects of system development when it comes to designing and managing the generation of inter-agent interaction protocols (Wooldridge et al. 2000). When agents come together and interact, various relations are established between them. More specifically, the ways agents relate to one another can be studied and conceptualised in terms of a set of relatively stable patterns of interactions. System structure is the label we give to these patterns of interactions. There are as many types of system structures as there are dimensions along which a multi-agent system can be differentiated and patterns of interactions identified (e.g. status, role, power, leadership). However, for the purposes of this paper, we restrict our attention to the concept of role and, accordingly, to the *role structure* of a multi-agent system. A role structure can be defined as the stable pattern of relations that exist among the roles in a multi-agent system (Huczynski and Buchanan 1991). This stable pattern of relations between roles provides the foundations for a certain degree of stability in the ways agents interact with one another. In fact, to some extent, the behaviour of each agent in a multi-agent system can be regarded as constrained by a set of expectations regarding the agent's decision-making and its interactions with the other members of the system. In turn, these expectations represent the cognitive components of the notion of social role. Therefore, by providing the agents with a set of expectations about each other's behaviour, the role structure can be viewed as a source of relatively stable patterns of interactions among the agents.

More precisely, the role structure of a multi-agent system can have an impact upon an agent's mental state in two main ways. Firstly, adopting a role imposes responsibilities, influences and constraints upon the role-player since it provides others with a set of stable expectations of what it will do and how it will respond to them. Specifically, it provides the role-player with: (a) a set of expectations about the other role-players' mental attitudes and behaviour, and (b) a set of expectations of how the other role-players will respond to its own mental attitudes and behaviour. Secondly, the role structure of a multi-agent system may be instantiated by agents who can thus influence one another through the social relationships that flow from their interrelated roles. A social relationship involves at least two related roles occupied by two agents who are aware of the connection between them. Thus, a social relationship implies the instantiation by two agents of at least two related roles, while a relation between two roles may, in general,

give rise to any number of instantiated social relationships. Each of these perspectives will now be dealt with in turn.

First, consider the case of adopting a role in a multi-agent system. Social scientists differ in the way in which they use the term "role." Definitions depend on how roles are to be used. We may consider prescriptive, evaluative, descriptive and action definitions of the concept of role (Handy 1993; Huczynski and Buchanan 1991). A prescriptive definition is concerned with what an agent should do when it plays a specific role. An evaluative definition, on the other hand, assesses how a role is being performed. A descriptive definition of a role is based on the actual duties performed by the agent being studied. Finally, an action definition is based on the actions that are performed while pursuing the role duties. However, any role may be considered from any of these four perspectives. Further, all four views are interrelated and interdependent. For example, consider the role of secretary. A job description specifying the main duties (e.g. managing correspondence, taking minutes of meetings) represents an example of a prescriptive definition. Such a definition provides criteria or standards against which to make assessments of how well the role of secretary is being performed (evaluative definition). Further, by observing and noting in detail what an individual does (e.g. how the secretary spends his or her time), a descriptive definition of the role of secretary can be developed. Finally, by observing the actions that are performed in pursuing the role tasks (e.g. the secretary may speak with colleagues, ask questions, establish a rapport), an action definition of secretary can be developed.

For the purposes of this work, however, a new cognitive definition of the concept of role is required (see Section 4.3 for more details). Thus, a role is here conceived of as a system of mental attitudes that an agent may adopt by occupying that role. This perspective is necessary because roles are viewed as providing agents with much of the information and many of the goals and other attitudes that drive their behaviour. Roles inform them of the problems and decisions they face, and possibly how to address these. For example, a role may inform an agent of where to look for appropriate goals, informational resources and value premises, how to achieve goals, and how to react to changes in mental attitudes. Returning to the particular example of a secretary role. In a business unit, this role may be regarded as leading the role-player to adopt and be motivated by the goal to contribute to the success of the unit, as well as to adopt the more specific goal of drafting routine correspondence. The secretary role can also provide the agent with some of the information needed to become more familiar with the task. Likewise, taking on this role may influence the agent to commit itself to proceed forthwith to do what any official instruction may require of it.

Differences among agents can, however, substantially affect their behaviour in roles that are identical from an observer's standpoint. While we may conceive of an ideal type of role as a system of attached mental attitudes, the instantiation of roles invariably involves both role-based mental attitudes and internally motivated ones. That is, a role is a specification of some, but not all, of the mental attitudes that drive the role-player's behaviour. Many other mental attitudes also underpin the agent's behaviour, including idiosyncratic attitudes that reflect differences between agents. Adoption of role-based attitudes may, however, bring about a modification of idiosyncratic attitudes. For example, adopting a goal attached to a role may require the agent to drop an internally motivated goal which is inconsistent with the role-based one (see Section 6). Consider, for instance, the case of a manager who believes in and wants to adopt a relaxed and participative style of behaviour, but whose role expectations lead him or her to adopt a more formal and directive style of behaviour. In such a situation, compliance with one set of goals makes it difficult or impossible to comply with other goals

and expectations. The two sets of goals are in conflict and create inconsistency. This inconsistency between socially and internally motivated mental attitudes must be overcome in order for the agents to behave rationally[3] (Cohen and Levesque 1990; Rao and Georgeff 1991; Simon 1957).

Let us turn now to the influence that can be exerted upon an agent's mental state by other agents and by social relationships. More generally, there are a number of ways in which agents can influence one another's mental states. Some of the main modes of social influence that are found in multi-agent systems are:

- *Authority*. An agent may be influenced by another to adopt a mental attitude whenever the latter is entitled to guide the behaviour of the former (Barnard 1938). Authority may thus be defined as the power of an agent to guide the behaviour of another by affecting the latter's mental state.
- *Helping disposition*. An agent may be influenced by another to adopt a mental attitude simply because it intends to contribute to the welfare of the latter. An example of this is the assumption of benevolence that is present in many of the early distributed problem-solving systems (Erman et al. 1980; Lesser and Corkhill 1983).
- *Trust*. An agent may be influenced by another to adopt a mental attitude merely on the strength of its confidence in the latter, without any critical scrutiny or any review of the evidential basis (Griffiths and Luck 1999; Marsh 1994).
- *Persuasion*. An agent may be influenced to adopt another agent's mental attitude via a process of argumentation (Kraus et al. 1998; Parsons et al. 1998; Walton and Krabbe 1995). In such cases, the adoption of a mental attitude depends upon conviction. Suggestions are judged partly on their merits, but partly on the merits of the agent making them. This is true both because the agents acting upon the recommendations often do not have the expertise needed to judge them, and because time constraints require them to accept the recommendations of the agents they trust (see above). Furthermore, an agent may be persuaded to adopt another agent's mental attitude via a process of bargaining or negotiation (Handy 1993; Huczynski and Buchanan 1991). In this case, an agent may agree with another to adopt one (or more) of the latter's mental attitudes in return for desired behaviour.
- *Threat*. An agent may be threatened to adopt a mental attitude on the basis of future negative interference or denied help (Kraus et al. 1998; Sierra et al. 1998). In the first case, an agent may be influenced by another to adopt a mental attitude whenever the latter threatens to execute an action that will conflict with the former's desires, goals or intentions. In the second case, an agent is influenced by a threat that another will not act to help the former to fulfil its desires, goals or intentions.

This list of the main modes of influence found in multi-agent systems is meant to be neither exhaustive nor mutually exclusive. For example, when an agent decides upon a particular course of action, some of the mental attitudes upon which this decision is based may have been imposed by the exercise of someone else's authority, some may have been the result of helping disposition, and so forth. Moreover, not all modes of social influence can be suitably exercised in all circumstances. For example, in the boss-secretary relationship the main forms of influence are authority (from the boss to the secretary) and helping disposition (hopefully symmetric!). Finally, the effectiveness of different modes of social influence depends on whether they are exercised within or

[3]The actual process by which this resolution is achieved depends on the specifics of the agent architecture that is used. For this reason, it is beyond the scope of this paper. However, in Section 6 we will introduce some of the axioms for consistency that underpin a social agent's mental state.

outside social relationships. Whereas some modes of influence can be exploited mainly within pre-existing social relationships, others can be exercised also when no social relationship already exists and one of the agents is unaware of the existence of the other (see our definition of social relationship in Section 4.4). For example, benevolence and helping disposition often reflects forms of social influence that occurs within existing social relationships (e.g. friendship). Conversely, authority and persuasion can be exercised even in the absence of any social relationship between agents. For example, the Prime Minister may exercise authority over the citizens without being aware of the identity of each of them, and therefore outside of any social relationship. Or, alternatively, let us consider a consumer who might be persuaded simply by an advertisement to adopt the intention to buy a new product. Again, in this case persuasion is exercised outside of any social relationship, since the producer cannot be aware of all the consumers who are potentially interested in buying the product.

In our framework, the notion of social relationship builds on the concept of multi-agent system role structure and the instantiation of two (or more) related roles by two (or more) agents. Given a role structure, each role presumes the appropriate enactment of the other roles that are interrelated with it. For example, let us consider again the pair of related roles "boss" and "secretary" and their instantiation by two different agents. These roles are related because each of them entails a set of expectations concerning the mental state and behaviour of the agent who occupies the other. In fact, attached to the secretary role are beliefs about the beliefs, goals and intentions attached to the boss role, whereas attached to the boss role is a set of beliefs about the beliefs, goals and intentions attached to the secretary role. A specific instantiation of the related roles "boss" and "secretary" by two agents gives rise to a social relationship between those agents. This social relationship can be regarded as a major causal force that shapes the role-players' mental states and behaviour. For instance, by taking on the role of secretary, an agent adopts the role-based beliefs about the boss's goals and intentions, and his or her behaviour becomes constrained by the role-based expectations concerning the response that he or she will have to the boss's goals and intentions. In turn, by interacting with the agent who is in the boss role, the agent in the secretary role may be influenced to adopt some of the former's (idiosyncratic) beliefs, goals or intentions. Correspondingly, by taking on the role of boss, an agent adopts the role-based beliefs about the secretary's goals and intentions, and his or her behaviour will be influenced by a set of expectations concerning the principal short-run effect that a change in his or her own goals and intentions might have upon the secretary's day-to-day activities. Furthermore, by interacting with the agent in the secretary role, the agent in the boss role may be induced to rectify its expectations, update its beliefs, and modify some of its goals and intentions concerning the secretary's activity.

In summary, an agent's social nature can clearly be seen to impact upon its individual mental state and behaviour. This impact can be explained either in terms of role adoption and the mental attitudes that are associated with roles, or in terms of social relationships and other agents' mental states. Besides providing new reasons for keeping individually motivated attitudes, roles, agents and social relationships also offer mental attitudes that can be adopted to complement or merely to change individual mental states. This suggests a view of mental attitudes as increasingly decentralised; i.e., spread out in the social environment. It is then the social nature of agents that leads them to internalise socially situated mental attitudes. From this, an agent can be regarded as a kind of *associative* entity, engaged in an iterated series of social actions and interactions aimed at *completing* its mental state. The complex interplay between the agent and its

social environment turns out to be a process in which roles, other agents and social relationships play a key function: they complement and augment the agent's bare individual mental attitudes. It is this completing process that we term social mental shaping. With these intuitions in place, the next section will introduce a formal multi-modal language through which this process will be expressed.

4. BASIC DEFINITIONS

This section introduces the formal framework within which the process of social mental shaping can be explored in more detail (a complete formal definition is given in the Appendix). First, we give a brief description of the model of time that underpins our logic (Section 4.1). Second, we concentrate on the formalisation of the analytical tools and principles for reasoning about and representing the agent's individual decision-making apparatus (Section 4.2). Finally, we formalise the notions of roles (Section 4.3) and social relationships (Section 4.4). The formalism used is a many-sorted, first-order, multi-modal language L which both draws upon and extends the work of Bell and Huang (1997), Cohen and Levesque (1990), Rao and Georgeff (1991), and Wooldridge and Jennings (1999). L allows us to reason about agents and their mental attitudes, with explicit reference to time points and intervals.

Informally, the $=$ operator is the usual first-order equality. The operators $\neg$ (not) and $\vee$ (or) have classical semantics, as does the universal quantifier $\forall$. The remaining classical connectives and the existential quantifier are assumed to be introduced as abbreviations in the usual way. We also use the punctuation symbols ")", "(", "]", "[", and comma ",".

4.1. Time

In L we have terms that denote *time points*, and we use t_i, t_j, ... and so on as variables ranging over time points. Every occurrence of a formula ϕ is stamped with a time t_i, written $\phi(t_i)$, meaning that ϕ holds at time t_i. Time is taken to be composed of points and, for simplicity, is assumed to be discrete and linear. In addition to time points, we have terms that denote *temporal intervals*, and we use i_i, i_j, ... and so on as variables ranging over time intervals. Temporal intervals are defined as pairs of points. Intervals of the form (t_i, t_i) can equally be written as time points. For time point t_i, t_i+1 is the time point that increments t_i; that is, t_i+1 is the time point obtained by extending t_i by a time point. Similarly, for interval i_i, i_i+1 is the interval that increments i_i. For example, if i_i is (3, 8) then i_i+1 is (3, 9).[4]

It will be convenient to adopt the following abbreviations:

- Interval terms of the form (t_i, t_i) will usually be abbreviated simply to t_i.
- Multiple occurrences of the same interval term may be eliminated when the result is unambiguous. For example, $(\phi \wedge \psi)(i_i)$ abbreviates $(\phi)(i_i) \wedge (\psi)(i_i)$.
- In complex sentences the same temporal terms are often repeated. In what follows we will adopt the convention that a missing temporal term in a well-formed formula is the same as the closest temporal term to its right. For example, $Bel(a_i, Goal(a_j, \phi))(t_i)$ states that at time t_i agent a_i believes that at time t_i agent a_j has the goal to make ϕ true at time t_i.

[4]Note that such formulae as $(t_i < t_j)$ and $(t_i \leq t_j)$ will be given a special treatment in the Appendix.

4.2. Agents

We have terms that denote *agents*, and we use a_i, a_j, ... and so on as variables ranging over individual agents. Agents are typically required to perform several tasks, and have to make decisions about how to achieve them. There are a number of properties that characterise agents (Franklin and Graesser 1997; Jennings 2000; Wooldridge 2000; Wooldridge and Jennings 1995). First, agents are autonomous, that is, they have control over their tasks and resources and will take part in cooperative activities only if they choose to do so. Second, agents are reactive: they respond to any perceived change that takes place within their environment and that affects their mental states. Third, agents are proactive: they do not simply act in response to their environment, but they exhibit opportunistic behaviour, take the initiative where appropriate, exploit and create serendipity (Wooldridge 2000). Fourth, agents have social ability: they can initiate social relationships with one another and may be mutually supportive during the execution of their joint actions (Carley and Newell 1994).

Given this, agents are here conceptualised as cognitive agents endowed with mental attitudes representing the world and motivating action (Panzarasa et al. 2001a; Wooldridge 2000). Furthermore, not only does an agent have an intentional stance towards the world, but it also represents other agents as cognitive agents similarly endowed with mental attitudes for representational and motivational purposes (Dennett 1987). The cognitive characterisation of the agents' mental states and decision-making apparatus is here formalised building on a fairly standard BDI framework (Cavedon et al. 1997; Cohen and Levesque 1990; Rao and Georgeff 1991; Singh 1995; Wooldridge 2000). More specifically, agents' mental states are here seen as sets of interrelated mental attitudes, among which there are doxastic attitudes (beliefs) and motivational attitudes (desires, goals, intentions). In what follows, we will develop the technical apparatus for dealing with the semantics of individual agents' beliefs (Section 4.2.1), desires (Section 4.2.2), goals (Section 4.2.3), and intentions (Section 4.2.4). In Section 4.2.5, we will then introduce a joint doxastic mental attitude, namely mutual beliefs, that builds on and transcends individual agents' beliefs.

4.2.1. Beliefs. An agent's belief set includes beliefs concerning the world, beliefs concerning mental attitudes of other agents, and introspective beliefs (see discussion below). This belief set may be incomplete. An agent may update its beliefs by observing the world and by receiving messages from other agents.

To express an agent's beliefs, we introduce the modal operator $Bel(a_i, \phi)(t_i)$, which means that at time t_i agent a_i has a belief that ϕ holds. The formal semantics of this modal operator are a natural extension of the traditional Hintikka's possible-worlds semantics (Hintikka 1962, 1972). In a model M, for each world w, agent a_i and time point t_i, there is an associated possible-worlds frame $(W_{(Bel, a_i, t_i, w)}, R_{(Bel, a_i, t_i, w)})$ which is centred at w. In other words, $W_{(Bel, a_i, t_i, w)}$ is a set of possible worlds that contains w, and $R_{(Bel, a_i, t_i, w)}$ is a binary relation on $W_{(Bel, a_i, t_i, w)}$ such that $(w, w') \in R_{(Bel, a_i, t_i, w)}$ for every $w' \neq w$ in $W_{(Bel, a_i, t_i, w)}$. If $(w, w') \in R_{(Bel, a_i, t_i, w)}$, then w' is a doxastic alternative for a_i at t_i in w; that is, in w at t_i, a_i cannot distinguish w' from the actual world w.

For model M, world w in M and variable assignment g, the semantic clause for *Bel* sentences is as follows:

$$M, w, g \models Bel(a_i, \phi)(t_i) \text{ iff } M, w', g \models \phi \quad \text{for all } (w, w') \in R_{(Bel, a_i, t_i, w)}$$

For simplicity, we assume the usual Hintikka-style schemata for *Bel*, that is the KD45 axioms (corresponding to a "Weak S5 modal logic") and "necessitation" rule[5] (Chellas 1980):

$[K_B]$ $\models Bel(a_i, \phi)(t_i) \wedge Bel(a_i, (\phi \supset \psi))(t_i) \supset Bel(a_i, \psi)(t_i)$ (closure under logical consequence)

$[D_B]$ $\models Bel(a_i, \phi)(t_i) \supset \neg Bel(a_i, \neg\phi)(t_i)$ (consistency axiom)

$[4_B]$ $\models Bel(a_i, \phi)(t_i) \supset Bel(a_i, Bel(a_i, \phi))(t_i)$ (introspection axiom)

$[5_B]$ $\models \neg Bel(a_i, \phi)(t_i) \supset Bel(a_i, \neg Bel(a_i, \phi))(t_i)$ (negative introspection axiom)

$[N_B]$ $\models \phi(t_i) \supset \models Bel\,(a_i, \phi)(t_i)$ (inference rule of necessitation)

The following conditions are imposed on the Belief-accessibility relation (Cohen and Levesque 1990):

[B1] Each $R_{(Bel,\, a_i,\, t_i,\, w)}$ is serial.

[B2] Each $R_{(Bel,\, a_i,\, t_i,\, w)}$ is transitive.

[B3] Each $R_{(Bel,\, a_i,\, t_i,\, w)}$ is euclidean.

A Belief-accessibility relation that satisfies conditions B1 to B3 validates the $D_B4_B5_B$ axioms. Furthermore, since axiom K_B is valid, it will be a theorem of any complete axiomatisation of normal modal logic. Similarly, the necessitation rule N_B is a rule of inference in any axiomatisation of normal modal logic (Chellas 1980).

4.2.2. Desires. An agent's desires are here conceived as the set of states of the world that the agent wishes to bring about (Bell 1995; Bell and Huang 1997; Kraus et al. 1998). Desires may not always be consistent. For example, an agent may desire to be healthy, but also to smoke; the two desires may lead to a contradiction. Furthermore, an agent may have unrealisable desires; that is, desires that conflict with what it believes possible.

To express an agent's desires, we introduce the modal operator $Des(a_i, \phi)(t_i)$, which means that at time t_i agent a_i has a desire towards ϕ. We take desires to be either present- or future-directed; that is, $Des(a_i, \phi(t_j))(t_i)$ means that agent a_i has, at time t_i, the desire that ϕ holds at t_j, where $t_i \leq t_j$.

The semantic clause for *Des* is analogous to that for *Bel*. We take K_D as the basis of our logic of desires:

$[K_D]$ $\models Des(a_i, \phi)(t_i) \wedge Des(a_i, (\phi \supset \psi))(t_i) \supset Des(a_i, \psi)(t_i)$

Furthermore, we have a necessitation property (Chellas 1980):

$[N_D]$ $\models \phi(t_i) \supset \models Des(a_i, \phi)(t_i)$

Again, axiom K_D and the necessitation rule N_D are, respectively, a theorem and a rule of inference in any axiomatisation of normal modal logic (Chellas 1980).

[5]In all the following axiom schemas, we will assume that the unbound variables are universally quantified as follows: $\forall\, a_i \in D_{Ag}, \forall\, t_i \in D_T, \forall\, w \in W$, where D_{Ag}, D_T, and W are, respectively, non-empty sets of agents, time points, and possible worlds (see Appendix). In addition, in all the axiom schemas, we assume that ϕ and ψ can be replaced by any well-formed formulae in the language.

4.2.3. Goals. Goals can be defined as a set of consistent states of the world that an agent autonomously chooses as potential candidates for motivating and governing its current and future behaviour. Thus, goals represent an agent's agenda comprising the autonomously selected states that the agent might be expected to bring about (Bell and Huang 1997). Even though an agent may well choose some of its goals among its desires, nonetheless there may be goals that are not necessarily desires. On the one hand, the goals that are also desires represent those states of the world that the agent might be expected to bring about precisely because they reflect what the agent wishes to achieve. In this case, the agent's selection of goals among its desires is constrained by two conditions. First, since goals must be consistent and desires may be inconsistent (see Section 4.2.2), only the subsets of consistent desires can be the potential candidates for being moved up to goal-status, and also the selected subsets of consistent desires must be consistent with each other. Second, since desires may be unrealisable whereas goals must be consistent with beliefs (see below), only a set of feasible (and consistent) desires can be potentially transformed into goals. On the other hand, an agent may generate goals that are not desires; that is, an agent might be expected to bring about states that do not necessarily represent the states of the world that it wishes to achieve.[6] Typically, these are goals that are instrumental to the achievement of those goals that are also desires. For example, an agent may generate the goal to work hard simply because it believes this is an appropriate way to fulfil its goal/desire to become successful. In this case, even though the agent may not desire to work hard, nonetheless adopting this goal may be instrumental to the achievement of another goal that is also a desire.

To express an agent's goals, we introduce the modal operator $Goal(a_i, \phi)(t_i)$, which means that at time t_i agent a_i has a goal towards ϕ. Like desires, goals can only be present-directed or future-directed, that is, $Goal(a_i, \phi(t_j))(t_i)$ means that agent a_i has, at time t_i, the goal that ϕ holds at t_j, where $t_i \leq t_j$.

The following axioms K_G and D_G state that goals are assumed to be, respectively, closed under implication and consistent:

$[K_G] \quad \models Goal(a_i, \phi)(t_i) \wedge Goal(a_i, (\phi \supset \psi))(t_i) \supset Goal(a_i, \psi)(t_i)$

$[D_G] \quad \models Goal(a_i, \phi)(t_i) \supset \neg Goal(a_i, \neg\phi)(t_i)$

We also introduce a *weak realism* constraint for goals (Wooldridge 2000). Agents do not have goals towards propositions the negations of which are believed. That is, agents' goals do not contradict their beliefs. Formally, we have the following axiom:

$[G_1] \quad \models Goal(a_i, \phi)(t_i) \supset \neg Bel(a_i, \neg\phi)(t_i)$

The logic of *Goal* is therefore $K_G D_G G_1$. The following conditions are imposed on the Goal-accessibility relation:

$[G_1]$ Each $R_{(Goal, a_i, t_i, w)}$ is serial.

$[G_2]$ $\forall w \exists w'$ s.t. $(w, w') \in R_{(Goal, a_i, t_i, w)}$ iff $(w, w') \in R_{(Bel, a_i, t_i, w)}$ (or $R_{(Bel, a_i, t_i, w)} \cap R_{(Goal, a_i, t_i, w)} \neq \varnothing$).

A Goal-accessibility relation that satisfies conditions G1 and G2 validates axioms D_G, and G_1. Again, axiom K_G is valid, and we have a necessitation property

[6]This property contrasts with the framework of Kraus et al. (1998), in which every goal is also a desire. In contrast to them, in our framework an agent may have a goal towards ϕ, but may not desire ϕ. Furthermore, Cohen and Levesque (1990) assume that all the agent's beliefs are also its goals. We do not have such a property. Indeed, our framework is more flexible, because it allows an agent to believe ϕ, but not to adopt it as one of its goals at the same time.

(Chellas 1980):

$[N_G] \quad \models \phi(t_i) \supset \models Goal(a_i, \phi)(t_i)$

Finally, the semantic clause for *Goal* is analogous to that for *Bel* and *Des*.

4.2.4. Intentions. A fundamental characteristic of individual intentions is that they involve a special kind of "self-commitment" to acting (von Wright 1980). As long as an agent intends to achieve a state, it has committed itself to act accordingly, that is, to perform all those actions that it deems appropriate for achieving that state (Norman and Reed 2001). Fundamentally, we can distinguish between two different forms of intentions, *Intentions-to* and *Intentions-that*, depending on whether the argument is respectively an action or a proposition (Bell 1995; Grosz and Kraus 1996).[7] That is, intentions can be subdivided into: (a) action-directed intentions (Intentions-to), involving the performance of some action; and (b) state-directed intentions (Intentions-that), involving the achievement of some state of affairs by performing some action. As indicated by Grosz and Kraus (1996), both types of intention commit an agent not to adopt conflicting intentions (Werner 1989), and constrain replanning in case of failure (Bratman 1987). However, since our main concern in this paper is with the agent's decision-making apparatus, in what follows we will restrict our attention to Intentions-that because they represent the basic attitudes that commit the agent to practical reasoning and, therefore, are inherently intertwined with its decision-making.

An agent will not adopt all its goals as intentions. The intuition is that an agent will not, in general, be able to achieve all its goals simultaneously. It must therefore choose a subset of its goals and commit itself to act in such a way to fulfil this subset. In this case, the agent is committed to fulfilling intentions that are also goals. However, not every intention is also a goal. In fact, there are intentions that the agent *ought* to adopt and does not *autonomously* choose as potential motivators of its own behaviour. In this case, to the extent that they do not result from the agent's autonomous selection and adoption, these intentions cannot be regarded as goals that the agent is committed to achieving (see our definition of goals in Section 4.2.3). In our view, intentions that are not goals are typically adopted by the agent for two related reasons: (a) they are instrumental to the achievement of some of the agent's goals; and (b) the agent is *obliged* to adopt them as a result of its being subjected to others' authority and/or to the influence of norms, rules, and regulations. The intuition is that, once the agent has autonomously selected its goals and, among them, its intentions, it may be automatically compelled to adopt other intentions that are imposed from the social environment and that are instrumental to the achievement of (some of) the goals/intentions autonomously chosen. For example, an agent may be forced to adopt an intention by another agent who has the authority to control the former's behaviour (see Section 7.3). In this case, the agent, based on its own goals and intentions, autonomously decides whether or not to be subjected to another's authority; however, once the authority relation has been established, the agent may be obliged by the other to change its mental state, adopt new intentions and modify its behaviour correspondingly. Or, alternatively, an agent may be obliged to adopt an intention by the set of the rules, norms and regulations reflected by the role(s) it has taken on in a multi-agent system (see Section 4.3.1.3). In these circumstances, the agent autonomously decides whether or not to become a member of the multi-agent system. However, once the decision to enter the system has been made,

[7]Grosz and Kraus (1996) also identify *potential* intentions, that is "intentions that an agent would like to adopt, but to which it is not yet committed" (p. 281). Indeed, potential intentions, so conceived of, are quite similar to what we call goals, that is, action-drivers that are candidates for being moved up to intention-status (Section 4.2.3).

the agent ought to change its mental state by adopting the intentions imposed by the rules and norms governing that system. In both the examples above, the adoption of new intentions is imposed by external forces (e.g. others' authority, norms, rules) and, as such, reflects the deontic aspects of the agent's practical reasoning.

The modal operator $Int(a_i, \phi)(t_i)$ is used to represent agent a_i's intention that proposition ϕ holds at time t_i. In other words, $Int(a_i, \phi)(t_i)$ means that at time t_i agent a_i is committed to doing some action after which ϕ holds. Like desires and goals, intentions can only be present-directed or future-directed, i.e., $Int(a_i, \phi(t_j))(t_i)$ means that agent a_i has, at time t_i, the intention that ϕ holds at t_j, where $t_i \leq t_j$.

An axiomatisation for intentions can now be presented. Intentions are here taken to be closed under implication (K_I) and consistent (D_I):

$[\mathrm{K_I}] \quad \models Int(a_i, \phi)(t_i) \wedge Int(a_i, (\phi \supset \psi))(t_i) \supset Int(a_i, \psi)(t_i)$

$[\mathrm{D_I}] \quad \models Int(a_i, \phi)(t_i) \supset \neg Int(a_i, \neg\phi)(t_i)$

As with goals, we introduce a weak realism constraint. Agents do not intend propositions the negations of which are believed. This ensures that agents' intentions do not contradict their beliefs (Kraus et al. 1998; Wooldridge 2000). Formally, we have the following axiom:

$[\mathrm{I_I}] \quad \models Int(a_i, \phi)(t_i) \supset \neg\, Bel(a_i, \neg\phi)(t_i)$

Again, we have a necessitation property (Chellas 1980):

$[\mathrm{N_I}] \quad \models \phi(t_i) \supset \models Int(a_i, \phi)(t_i)$

The logic of *Int* is therefore $\mathrm{K_I D_I I_1}$. The following conditions are imposed on the Intention-accessibility relation:

[[I1]] Each $R_{(Int,\, a_i,\, t_i,\, w)}$ is serial.

[I2] $\forall w \exists w' \,\mathrm{s.t}\, (w, w') \in R_{(Int,\, a_i,\, t_i,\, w)}$ iff $(w, w') \in R_{(Bel,\, a_i,\, t_i,\, w)}$ (or $R_{(Bel,\, a_i,\, t_i,\, w)} \cap R_{(Int,\, a_i,\, t_i,\, w)} \neq \varnothing$).

An Intention-accessibility relation that satisfies conditions I1 and I2 validates axioms $\mathrm{K_I}$, $\mathrm{D_I}$ and $\mathrm{I_1}$. Finally, the semantic clause for intentions is analogous to that of beliefs, desires, and goals.

4.2.5. Mutual Beliefs. In this section, we will develop the formalisation of mutual beliefs, namely a form of joint doxastic higher-order mental attitude that is jointly maintained by a group of two or more individual agents. Crudely, a mutual belief can be defined as an infinite conjunction of an agent's belief about an agent's belief about an agent's belief and so forth, that a proposition holds (Cohen and Levesque 1991). In what follows, the focus will be only on mutual beliefs attributable to a *pair* of agents. The same considerations can be extended to groups of n agents.

For agents a_i and a_j and formula ϕ, our language L includes the modal operator $M\text{-}BEL(\{a_i, a_j\}, \phi)(t_i)$, which means that, at time t_i, agents a_i and a_j have a mutual belief that proposition ϕ holds. More formally, the semantics for this modal operator can be defined as follows. Firstly, we examine the semantics of each of two members of a group having a mental attitude towards a formula. Following Rao et al. (1992), we introduce the operator $E\text{-}BEL(\{a_i, a_j\}, \phi)(t_i)$, which means that, at time t_i, both agents a_i and a_j believe that ϕ holds. We have the following definition:

$$\forall\, a_i, a_j, \forall\, t_i \;\; E\text{-}BEL(\{a_i, a_j\}, \phi)(t_i) \equiv Bel(a_i, \phi)(t_i) \wedge Bel(a_j, \phi)(t_i).$$

Now we say that, at time t_i, agents a_i and a_j mutually believe that ϕ, $M\text{-}BEL(\{a_i, a_j\}, \phi)(t_i)$, iff at time t_i both a_i and a_j believe that ϕ and each of them believes that

each of them believes that ϕ and each of them believes that each of them believes that each them believes that ϕ and so on *ad infinitum* (Wooldridge and Jennings 1999; Wooldridge 2000). If $k \in \mathbb{N}$ such that $k > 0$, we define $E\text{-}BEL^k(\{a_i, a_j\}, \phi)(t_i)$ inductively in the following way. Let $E\text{-}BEL^k(\{a_i, a_j\}, \phi)(t_i)$ be an abbreviation for $E\text{-}BEL(\{a_i, a_j\}, \phi)(t_i)$ if $k = 1$, and for $E\text{-}BEL(\{a_i, a_j\}, E\text{-}BEL^{k-1}(\{a_i, a_j\}, \phi))(t_i)$ otherwise. The mutual belief of ϕ in group $\{a_i, a_j\}$ at time t_i is then defined as follows: $M\text{-}BEL(\{a_i, a_j\}, \phi)(t_i) \equiv \bigwedge_{k>0} E\text{-}BEL^k(\{a_i, a_j\}, \phi)(t_i)$ (see Wooldridge (2000) and Fagin et al. (1995) for details).

Unlike some of the previous work in this area (Cavedon et al. 1997), in our framework mutual beliefs are not required to be first-class entities. In fact, the formalisation has been expressed in terms of the beliefs of the individual agents that jointly maintain the mutual belief. As a result, in common with other work in this field (Kinny et al. 1994), our definition has the advantage of capturing the interplay between the agents' individual beliefs that are propagated upwards to a higher-order level at which the mutual belief can be regarded as transcending the individual agents' mental states.

4.3. Roles

Our logic L is enriched by terms that denote *roles*, and we use $r_i, r_j, \ldots$ and so on as variables ranging over roles. Drawing on the cognitive perspective we advocated earlier (Section 3), a role can be viewed as *a set of mental attitudes governing the behaviour of an agent occupying a particular position within the structure of a multi-agent system*. The mental attitudes that typically define roles are beliefs, goals and intentions. An agent, by occupying a role, can adopt these role-based attitudes, and such adoption will in turn impact upon the agent's mental state. Some of the agent's mental attitudes will be modified; some simply complemented with other attitudes. So conceived of, a role turns out to be a *sub-cognitive entity* (Cavedon and Sonenberg 1998; Werner 1989). That is, a role is an entity that is endowed with mental attitudes concerning the physical and social environment. In this view, connections among roles rest on role-based cognitive representations of other roles in terms of their attached mental attitudes.

Our conception of roles as sub-cognitive entities needs to be further clarified in two respects. First, being sub-cognitive means that, although endowed with mental attitudes, roles do not have the capacity to perform cognitive processes (e.g. reasoning, decision-making) through the transformation of their mental attitudes. For example, a role cannot transform a goal that is attached to it into an intention. Such a transformation can only be done by a (cognitive) agent who takes on the role, then adopts the attached goal, and finally commits itself to fulfil it (Rao and Georgeff 1991). Second, although endowed with mental attitudes, roles are just organisational constructs. As such, they cannot be regarded as (sub-cognitive) agents, because they do not have the capacity to perform actions and, therefore, they are unable to affect the external world in order to reduce the discrepancy between the world and their regulatory internal representations (Bratman 1987). For example, roles may be endowed with goals, but do not have the capacity to pursue them. Only an agent, by adopting a role-based goal, can bring about a state of the world in which the goal is attained.

Our attempt to develop a cognitive model of roles is not entirely new in the literature. Werner (1989) characterises a social role as a description of an abstract agent, R_r, that defines the state information, permissions, responsibilities, and values of that agent role. According to Werner, when an agent a takes on a role r, it changes its mental state R_a so that R_r becomes part of R_a. In the same vein, Cavedon and Sonenberg (1998)

also characterise roles in terms of mental attitudes. They associate goals with roles, using the modality *RoleGoal*. In their framework, $RoleGoal(r, \phi)$ asserts that adopting role r involves adopting goal ϕ. This provides them with a useful level of abstraction to express generic social commitment (see Section 2). Thus, according to Cavedon and Sonenberg's model, whenever an agent a adopts role r, a commitment to taking on goal ϕ arises. This social commitment is taken to be between a and the other agents that are involved in the corresponding relationships. Furthermore, Cavedon and Sonenberg extend their framework to capture the different influences and responsibilities that different roles may have on the same agent. To model this, they introduce the predicate $Influence(a, \langle r, rn\rangle, \langle r', rn'\rangle)$ to assert that role r in relationship rn is more influential to agent a than role r' in rn'.

The cognitive model of roles presented in this paper will extend the above-mentioned frameworks along the following lines. Roles are associated not only with goals and information, but more generally with the standard set of mental attitudes that are usually introduced to describe and reason about rational agency (however, here roles are not taken to be associated with desires; see Section 4.3.1 for details). This is important in order to investigate the *full* impact that a role in a multi-agent system can have on an agent's mental state. To this end, the logic proposed here is enriched by a predicate, *In*, and three primary modal operators—*RoleBel, RoleGoal*, and *RoleInt*.

First, the 2-place predicate $In(a_i, r_i)(t_i)$ expresses the fact that some agent is in a role. Specifically, $In(a_i, r_i)(t_i)$ asserts that agent a_i is in role r_i at time t_i. For example, if a_i is Linda and r_i is "secretary of the department," $In(a_i, r_i)(t_i)$ means that at time t_i Linda acts as the secretary of the department.

Second, three modal operators are introduced that allow the development of a cognitive characterisation of roles in terms of their attached mental attitudes (see Section 4.3.1 for details). Together these three operators express the mental attitudes that an agent can internalise by adopting the role to which they are attached. In Section 5 it is shown that this characterisation of roles in terms of their attached mental attitudes underpins our formalisation of the impact that the role structure of a multi-agent system has on the mental states of the system's members. Like agents' mental attitudes (see Sections 4.2.1–4.2.4), the semantics of role-based mental attitudes will be expressed via accessibility relations between possible worlds.

The cognitive characterisation of roles advocated here must be further specified. Attached to roles there are two main types of mental attitudes: *mandatory* attitudes and *optional* attitudes. On the one hand, role-based mandatory attitudes are constitutive and relevant to the role to which they are attached. These are the attitudes that the agent *must* adopt whenever it takes on the role. On the other hand, role-based optional attitudes are not intimately constitutive of the role to which they are attached. These are the attitudes that the role-player may *decide* whether or not to adopt. For example, attached to the role of secretary there might be the goal of supervising the boss's correspondence. This refers to a job specification and the role-player may be obliged to adopt such a goal. However, there might well be the attached goal of being friendly with the other people in the department, and the secretary is just expected but not obliged to behave this way with his or her colleagues. In Section 5, this distinction between mandatory and optional role-based mental attitudes is shown to be fundamental to the problem of *preventing automatic attitude-adoption* whenever an agent occupies a role in a multi-agent system. Indeed, it is both unnecessary and dangerously strong to always force an agent to adopt the mental attitudes associated with its role through the use of axiom schemas. Although there is no doubt that an agent is committed to adopting a subset of the mental attitudes attached to its role, there might well be circumstances in

which an autonomous agent can decide whether or not to adopt a role-based mental attitude. In particular, there might be attitudes attached to roles that an agent decides not to adopt, for its own reasons.

Note that the model described here is more flexible than Cavedon and Sonenberg's (1998) framework, according to which the role-player is expected/requested to adopt *all* the mental attitudes (goals) attached to the role. In the approach advocated in this paper, an agent is expected/required to adopt *only a subset* of the mental attitudes that are attached to the role it has taken on, namely the subset of mandatory attitudes. Also, this notion of mandatory role-based mental attitudes is consistent with the concept of *organisational commitment* (Handy 1993). When a member, a_i, is organisationally committed to its group, it is committed to adopting (some of) the mental attitudes that are attached to the role it has taken on within the group. Then, a_i's organisational commitment to the group implies that a_i is committed to acting in accordance with the responsibilities, expectations, requests, obligations relative to (some of) the mental attitudes attached to its role. Therefore, our characterisation of roles as sets of mental attitudes that include mandatory ones is consistent with the prescriptive account of roles as sets of behavioural obligations, rights and privileges based on the role-player's organisational commitment to the group.[8]

4.3.1. Role-based Mental Attitudes. The set of role-based mental attitudes includes beliefs, goals and intentions. Each of these attitudes, once adopted by the role-player, will impact upon its behaviour by modifying its mental state. First, an agent, by taking on a role, may complement and/or change the information it already maintains about the world with the information attached to the role. In this respect, adopting a role affects the role-player's decision-making behaviour, the way alternatives are searched for and evaluated, and ultimately how actions are selected and performed (Simon 1957). Second, attached to a role are goals that represent the set of states of the world that the role-player might be expected to bring about. Examples of role-based goals include general indications of conduct that might be expected to guide the role-player's behaviour, such as recommendations concerning support to co-members, attitudes towards superiors, and means of communication. By taking on a role and adopting the role-based goals, an agent therefore complements and/or modifies its own agenda that might motivate its current and future behaviour (Bell and Huang 1997). In our view, role-based goals are optional mental attitudes that the role-player may decide whether or not to adopt. Conversely, the third category of mental attitudes provided by roles, namely role-based intentions, are mandatory attitudes typically reflecting the role-player's duties and obligations.[9] Examples of role-based intentions are role prescriptions such as rules and regulations, standards, policy decisions, job descriptions, or directives from superiors. Thus, role-based intentions are mandatory mental attitudes that the role-player is compelled to adopt. By taking on a role and internalising its attached intentions, an agent will modify its self-commitments to acting and its behaviour will be affected accordingly (von Wright 1980). Finally, unlike individual agents' mental states, roles do not include

[8]Note that although our notion of roles is consistent with the account of roles as sets of behavioural obligations, rights and privileges, this distinction between mandatory and optional attitudes is not as rich as the variety of individual and collective normative positions proposed by authors such as Lindahl (1977) and Sergot (1998).

[9]Note that, even though role-based intentions reflect the rules, norms and regulations that govern a multi-agent system, nonetheless the role-player may still have the freedom to decide how to comply with (some of) these role-based intentions. That is, there may be different degrees of cogency in role-based intentions. For example, the role-player may have the opportunity to determine its own role expectations where role-based intentions are specified loosely or only in very general terms (Handy 1993).

desires. In fact, in our framework, desires are taken to typically reflect the cognitive freedom of those agents who are endowed with a fully explicated model of cognition and a decision-making apparatus. Therefore, desires cannot be regarded as attached to sub-cognitive entities such as roles.

In what follows, we will now give a brief description of the technical details for dealing with the semantics of role-based beliefs, goals and intentions. To this end, we will introduce three new modal operators, $RoleBel(r_i, \phi)(t_i)$, $RoleGoal(r_i, \phi)(t_i)$, and $RoleInt(r_i, \phi)(t_i)$, which mean that at time t_i, respectively, the belief that ϕ holds, the goal towards ϕ, and the intention towards ϕ are attached to role r_i.

4.3.1.1. Role-based Beliefs

The belief set attached to a role includes beliefs concerning the world and beliefs concerning the mental attitudes attached to other roles. This belief set may be incomplete. As with agents' beliefs, we use Hintikka's possible-worlds semantics to develop a formalisation for role-based beliefs (Hintikka 1962, 1972). In a model M, for each world w, role r_i and time point t_i, there is an associated possible-worlds frame $(W_{(RoleBel, r_i, t_i, w)}, R_{(RoleBel, r_i, t_i, w)})$ which is centred at w. If $(w, w') \in R_{(RoleBel, r_i, t_i, w)}$, then w' is a doxastic alternative for r_i at t_i in w; that is, in w at t_i, and for r_i, w' cannot be distinguished from the actual world w.

Given this, the semantic clause for *RoleBel* sentences is similar to the semantic clause for *Bel* sentences. Finally, the axiomatisation of *RoleBel* simply reflects the fact that role-based beliefs are here assumed to be closed under consequence (Chellas 1980). Furthermore, we have the usual inference rule of necessitation. Unlike individual agents' beliefs, role-based beliefs are not taken to be consistent. In fact, attached to roles may be different pieces of information that might contradict each other. This situation is typically referred to as *role ambiguity*, and may result from a lack of clarity of the information available for the adequate performance of the role (Handy 1993). Role ambiguity often relates to such matters as the methods of performing tasks, standards of work, and the evaluation and appraisal of performance. Finally, unlike agents' beliefs, role-based beliefs are not axiomatised by the introspection and the negative introspection axioms. In fact, these axioms reflect a degree of cognitive awareness that goes beyond the restricted sub-cognitive capabilities of roles. Given this, the Role-Belief accessibility relation is not taken to be further constrained, as closure under logical consequence and the necessitation rule are, respectively, a theorem and a rule of inference in any axiomatisation of normal modal logic (Chellas 1980).

4.3.1.2. Role-based Goals

Role-based goals reflect the set of behaviours that might be expected of a role-player by the occupants of other related roles. They provide guidelines for certain patterns of behaviour that, although not specified in detail, might nonetheless, once adopted, impact upon the role-player's mental state. Role-based goals can only be present-directed or future-directed, that is, $RoleGoal(r_i, \phi(t_j))(t_i)$ means that attached to role r_i, at time t_i, there is a goal that ϕ holds at t_j, where $t_i \leq t_j$.

Like beliefs, the semantics of role-based goals draw on Hintikka's possible worlds semantics, and therefore the semantic clause for *RoleGoal* is analogous to that for *RoleBel* (Hintikka 1962, 1972). Furthermore, the axiomatisation of role-based goals simply reflects closure under consequence and the necessitation property (Chellas 1980). Therefore, unlike agents' goals, role-based goals are not assumed to be consistent. In

fact, attached to roles, there might be goals reflecting contradictory expectations of behaviour. Inconsistency between role-based goals is typically referred to as *role incompatibility*, and arises when the role-player faces a situation in which simultaneous different or contradictory expectations create conflict (Handy 1993). In these circumstances, compliance with one set of role-based goals makes it difficult or impossible to comply with other goals. An example concerns an agent's acting as the managing director of a business unit, who might face opposing expectations from the finance director (who is mainly concerned with balancing outgoing and incoming cash-flows) and the marketing director (who is mainly concerned with the capability of the products and/or services sold of meeting the customers' needs and maximising the revenues raised). Another typical example of inconsistency between role-based goals is *role overload*. This occurs when the role-player is faced with too many expectations and is therefore unable to meet all of them satisfactorily (Handy 1993). In these circumstances, some role expectations must be neglected in order to satisfy others, and this leads to a conflict of priority (see Section 6).

Finally, unlike agents' goals, role-based goals are not taken to be constrained by a weak realism condition (Wooldridge 2000). This means that attached to roles there might be pieces of information that contradict the pattern of behaviours that might be expected of the role-player. This represents another case of role ambiguity that reflects uncertainty and a lack of clarity as to the precise requirements of the role (Handy 1993). In these situations, the role-player might be unable to adequately perform what is expected, and its own perceptions of the role-based goals might differ from the expectations of other agents.

4.3.1.3. Role-based Intentions

Role-based intentions provide the role-player with a self-commitment to acting (von Wright 1980). That is, by adopting a role-based intention, an agent commits itself to act accordingly. Since a role is not a cognitive agent with capabilities of performing behaviour, role-based intentions cannot be action-directed intentions (Intentions-to). In fact, these intentions involve the performance of some action by the same entity that holds them (Bell 1995; Grosz and Kraus 1996). Therefore, all role-based intentions are here taken to be state-directed intentions (Intentions-that), thus reflecting some state of affairs that a potential role-player will commit itself to bring about.

Like role-based goals, role-based intentions can only be present-directed or future-directed, that is, $\mathit{RoleInt}(r_i, \phi(t_j))(t_i)$ means that attached to role r_i, at time t_i, there is an intention that ϕ holds at t_j, where $t_i \leq t_j$. Furthermore, in accordance with our formalisation of the individual agent's cognitive make-up, we do not take role-based intentions to be a subset of role-based goals. In fact, as mentioned in Section 4.3.1, role-based intentions are mandatory attitudes that the role-player ought to adopt whenever it takes on the role, whereas role-based goals are optional attitudes that the role-player may autonomously decide whether or not to adopt.

Finally, like beliefs and goals, the semantics of role-based goals draw on Hintikka's possible worlds semantics, and the semantic clause for *RoleGoal* is therefore analogous to that for *RoleBel* and *RoleGoal* (Hintikka 1962, 1972). Given this, in common with role-based goals, the axiomatisation of role-based intentions simply reflects closure under logical consequence and the necessitation rule (Chellas 1980). Like role-based goals, role-based intentions are not constrained by a consistency axiom and a weak realism property. Therefore, attached to a role there might be intentions that contradict one another and that are inconsistent with some of the pieces of information attached

to the same role. The implications of this in terms of role incompatibility, overload, and ambiguity are similar to what has been said above about role-based goals that might be inconsistent with one another and with role-based beliefs.

4.4. Social Relationships

Attitude adoption via role-adoption is not the only way in which sociality can have an impact upon an agent's mental state. Indeed, an agent can also be socially influenced to adopt or to keep a mental attitude by its being in a social relationship with an acquaintance (as discussed in Section 3). Therefore, in order to analyse this form of social influence we need to introduce and formalise the notion of social relationships between agents.

In addition to roles, we have terms that denote *relationship types*, and we use (r_i, r_j), $(r_j, r_k), \ldots$ and so on as variables ranging over relationship types. A relationship type, (r_i, r_j), represents a relationship abstraction between a pair of roles. For example, the roles "boss" and "secretary" can be linked by a particular type of social relationship that empowers one to dictate certain aspects of the work agenda of the other. An instantiation of a relationship type gives rise to a *social relationship* between agents.[10] For example, the boss-secretary relationship type can give rise to a number of relationship instances—one involving agent a_i as the boss and agent a_j as secretary, another one involving the same agent a_i as the boss and agent a_k as secretary.

We introduce the operator $rel(a_i, a_j, (r_i, r_j))(t_i)$ to indicate that agents a_i and a_j are in a social relationship of type (r_i, r_j) at time t_i. Formally, we have the following definition:

Definition 4.1. At time t_i, agents a_i and a_j are in a *social relationship* of type (r_i, r_j) iff, at t_i: (a) a_i occupies role r_i and a_j occupies role r_j, or vice versa; and (b) a_i and a_j mutually believe that (a). Formally, we have:

$$\begin{aligned}
&\forall a_i, a_j, \forall (r_i, r_j), \forall t_i\, rel(a_i, a_j, (r_i, r_j))(t_i) \\
&\quad \equiv (In(a_i, r_i) \wedge In(a_j, r_j) \wedge M\text{-}BEL(\{a_i, a_j\}, (In(a_i, r_i) \wedge In(a_j, r_j))))(t_i) \\
&\qquad \vee (In(a_j, r_i) \wedge In(a_i, r_j) \wedge M\text{-}BEL(\{a_i, a_j\}, (In(a_j, r_i) \wedge In(a_i, r_j))))(t_i)
\end{aligned}$$

5. FORMS OF SOCIAL MENTAL SHAPING

Social mental shaping is a social process affecting an agent's mental state. More precisely, it refers to the phenomenon that the mere social nature of agents impacts upon their mental attitudes, and thereby motivates their behaviour. The process of social mental shaping may involve social roles, social relationships or simply other agents outside of any relationship (see discussion below). Its typical outcome is a modification of an agent's mental state. This may happen either when the agent is socially motivated to adopt a *new mental attitude*, or when the agent is socially motivated to have *new reasons* for keeping a mental attitude adopted at an earlier stage.

[10]The fact that in our framework social relationships are defined in terms of types of relationships between roles implies that a relationship between two agents exists as long as the agents occupy two related roles. For example, should two individuals, who are husband and wife, play, respectively, the role of boss in the marketing department and secretary in the human relations department in a business firm, and should these two roles not be connected with each other, then the two individuals are not in a social relationship with respect to that particular pair of roles. Conversely, they will be in a social relationship with respect to the related roles "husband" and "wife."

In what follows we will give the technical apparatus for dealing with the process of social mental shaping in all of its three basic forms: (a) social mental shaping based on social roles (Section 5.1); (b) social mental shaping occurring between agents outside of any social relationship (5.2.1); and (c) social mental shaping occurring within social relationships (5.2.2). To this end, we will introduce the modal operator *Infl* that will be used to formalise all the above forms of social mental shaping. For simplicity, we write $Att(a_i, \phi)(t_i)$ to indicate that agent a_i, at time t_i, has a mental attitude towards ϕ.

5.1. Roles and Social Mental Shaping

We formalise the influence of a role on an agent's mental state by expressing the modal operator *Infl* in terms of $Att(a_i, \phi)$ and r_i, where $Att(a_i, \phi)$ represents either a belief, a goal or an intention that agent a_i holds towards ϕ, whereas r_i is a role. We have:

$$\forall a_i, \forall r_i, \forall t_i Infl(Att(a_i, \phi), r_i)(t_i) \equiv RoleAtt(r_i, \phi)(t_i) \wedge Bel(a_i, RoleAtt(r_i, \phi))(t_i) \wedge In(a_i, r_i)(t_i) \wedge (In(a_i, r_i) \supset Att(a_i, \phi))(t_i)$$

Informally, the meaning of $Infl(Att(a_i, \phi), r_i)(t_i)$ is that at time t_i agent a_i is socially influenced by role r_i to have the attitude *Att* towards a state of the world ϕ iff at time t_i: (a) *Att* towards ϕ is an attitude attached to role r_i; (b) agent a_i believes that *Att* towards ϕ is attached to r_i; (c) a_i occupies r_i; and (d) a_i's taking on role r_i implies a_i's adopting or keeping the attitude *Att* towards ϕ.[11]

Given the above definition, it is trivial to note that if at time t_i agent a_i is socially influenced by role r_i with respect to a mental attitude *Att* towards ϕ, then at time t_i agent a_i will hold that attitude. In such situations, *Att* is a *socially motivated attitude*. Formally, we have the following social mental shaping rule (schema):

[S1] $\models \forall a_i, \forall r_i, \forall t_i, Infl(Att(a_i, \phi), r_i)(t_i) \supset Att(a_i, \phi)(t_i)$.

Note that the following schema is *not* valid:

$$\not\models \forall a_i, \forall r_i, \forall t_i (In(a_i, r_i)(t_i) \wedge Att(a_i, \phi)(t_i)) \supset Infl(Att(a_i, \phi), r_i)(t_i).$$

Thus, a role-player might occupy a role and hold a mental attitude without necessarily being influenced by that role to adopt or keep that attitude. For example, the attitude might not be attached to the role, and the agent might simply be internally motivated to adopt it.

Further, note that the following schema holds *only* for role-based mandatory attitudes, namely role-based intentions (Section 4.3.1.3):

[S2] $\models \forall a_i, \forall r_i, \forall t_i (RoleInt(r_i, \phi)(t_i) \wedge Bel(a_i, RoleInt(r_i, \phi))(t_i) \wedge In(a_i, r_i)(t_i)) \supset Infl(Int(a_i, \phi), r_i)(t_i)$.

That is, if there is a mandatory intention attached to a role, and the agent is aware of this, then the agent will *automatically* adopt such an intention by occupying the role.[12]

[11] Note that this definition does not imply that the agent is socially influenced by a role only when the role strictly motivates the agent to adopt a new attitude. In fact, the agent might already have an individually motivated attitude that is the same as the one attached to the role. In this case, however, occupying the role and believing that the attitude is attached to the role provide the agent with another reason to keep its attitude. An attitude that was only individually motivated when adopted originally by the agent, now becomes supported and socially motivated by the role that the agent occupies. The same considerations also apply to the other two forms of social mental shaping occurring between agents (Section 5.2).

[12] In this case, if the agent already has the role-based mandatory intention before occupying the role, then the role will simply provide the agent with new reasons for keeping the intention. Specifically, an intention that the agent autonomously chose at an earlier stage, may become a mandatory intention that the agent ought to keep.

In such a situation, the role-player is subjected to social mental shaping. Note that S2 does not hold with optional role-based attitudes: in fact, in these circumstances, the role-player may decide not to adopt the attitudes attached to the role, for whatever reason.

5.2. Social Mental Shaping Between Agents

Social mental shaping between agents may take place either within or outside social relationships. In either case, one agent is influenced by another to adopt or to keep a mental attitude. Furthermore, a social relationship may represent not only the input but also the output of a social mental shaping process. For example, when an agent wishes to get involved in a *new* relationship with another agent, it may well decide to persuade the latter via a process of argumentation aimed at having that new type of relationship established. Here, the new social relationship represents the outcome of the exercise of social influence. In what follows, we will formalise the basic form of social mental shaping between agents, namely that one occurring outside of any social relationship. In Section 5.2.2, social mental shaping within social relationships will be dealt with in more detail.

5.2.1. Social Mental Shaping Outside Social Relationships. To formalise social mental shaping occurring between a pair of agents and outside of any social relationship, we express the modal operator *Infl* in terms of $Att(a_i, \phi)$, where $Att(a_i, \phi)$ represents either a belief, a desire, a goal or an intention that agent a_i holds towards ϕ. We have:

$$\forall a_i, \forall t_i\, Infl(Att(a_i, \phi))(t_i) \equiv \exists a_j \text{ s.t. } Att(a_j, \phi)(t_i) \wedge Bel(a_i, Att(a_j, \phi))(t_i) \wedge (Bel(a_i, Att(a_j, \phi)) \supset Att(a_i, \phi))(t_i)$$

Informally, $Infl(Att(a_i, \phi))(t_i)$ means that, at time t_i, agent a_i is socially influenced to adopt or to keep the mental attitude *Att* towards ϕ iff at time t_i: (a) there is another agent a_j who holds *Att* towards ϕ; (b) a_i believes that a_j holds *Att*; and (c) a_i's believing that a_j holds *Att* implies a_i's adopting or keeping attitude *Att*. This form of social mental shaping covers several influence patterns among agents, from imitation of others' desires to spontaneous goal-adoption, from benevolent (not due) adhesion to others' intentions, to exploitation and adoption of others' beliefs (see Section 7). Note that in all these cases of social mental shaping the two agents *need not be in a social relationship* (e.g. one of them might not be aware of the other), as opposed to the form of social mental shaping detailed below.

5.2.2. Social Relationships and Social Mental Shaping. We now want to formalise how an agent's mental state can be influenced by its being within a *social relationship* with another agent. To this end, we now express our modality *Infl* in terms of $Att(a_i, \phi)$ and $rel(a_i, a_j, (r_i, r_j))$, where $Att(a_i, \phi)$ represents either a belief, a desire, a goal or an intention that agent a_i holds towards ϕ, whereas $rel(a_i, a_j, (r_i, r_j))$ represents a social relationship of type (r_i, r_j) between agents a_i and a_j. We have:

$$\begin{aligned}\forall a_i, a_j, \forall (r_i, r_j), \forall t_i\, &Infl(Att(a_i, \phi), rel(a_i, a_j, (r_i, r_j)))(t_i) \\ &\equiv rel(a_i, a_j, (r_i, r_j))(t_i) \wedge Att(a_j, \phi)(t_i) \wedge Bel(a_i, Att(a_j, \phi))(t_i) \\ &\quad \wedge (Bel(a_i, Att(a_j, \phi)) \supset Att(a_i, \phi))(t_i)\end{aligned}$$

Informally, we say that, at time t_i, agent a_i is influenced to adopt or to keep attitude *Att* towards ϕ by its being situated within a social relationship of type (r_i, r_j) with

another agent a_j iff at time t_i: (a) agents a_i and a_j are in social relationship of type (r_i, r_j); (b) a_j holds the mental attitude *Att* towards ϕ, (c) a_i believes that a_j holds *Att*; and (d) a_i's believing that a_j holds *Att* implies a_i's adopting or keeping the attitude *Att* towards ϕ. With the following schema S3, it is trivial to note that if at time t_i agent a_i's mental state is socially influenced by the social relationship that a_i has with a_j with respect to a mental attitude *Att* towards ϕ, then at time t_i agent a_i will hold that attitude. In such a situation, *Att* is a *socially motivated attitude*. Formally, we have:

$$[\text{S3}] \quad \models \forall a_i, a_j, \forall (r_i, r_j), \forall t_i \, Infl(Att(a_i, \phi), rel(a_i, a_j, (r_i, r_j)))(t_i) \supset Att(a_i, \phi)(t_i).$$

In general, this form of social mental shaping is based, and depends, on the agent's decision whether or not to adopt one of its acquaintance's mental attitudes. However, as with role-based social mental shaping, there are circumstances in which an agent involved in a social relationship with another is *required* to adopt or to keep one or more of its acquaintance's mental attitudes. In such cases, the agent might well *autonomously* decide whether or not to establish a relationship with another agent but, once established, the relationship may automatically impose a number of mental attitudes on the former's mental state. These are *relationship-based mandatory intentions* (see Section 7.3). For example, the boss is by right allowed to order other employees to perform particular activities. In this case, if a secretary decides to interact (i.e., establish a social relationship) with a boss, then he or she ought to change his or her mental state so as to adopt some of the intentions imposed by the boss. For such attitudes, we have the following schema:

$$[\text{S4}] \quad \models \forall a_i, a_j, \forall (r_i, r_j), \forall t_i \, (rel(a_i, a_j, (r_i, r_j))(t_i) \wedge Int_{M(a_i, t_i, (r_i, r_j))}(a_j, \phi)(t_i)$$
$$\wedge Bel(a_i, Int_{M(a_i, t_i, (r_i, r_j))}(a_j, \phi))(t_i)) \supset Infl(Int(a_i, \phi), rel(a_i, a_j, (r_i, r_j)))(t_i)$$

where $Int_{M(a_i, t_i, (r_i, r_j))}(a_j, \phi)(t_i)$ represents an intention held by agent a_j that a_j has the authority to impose upon agent a_i if, at time t_i, a_i relates to a_j within a relationship of type (r_i, r_j). Informally, if an agent establishes a social relationship with another agent who holds a mental attitude that is mandatory for the former with respect to that relationship, then the former will *automatically* adopt this attitude.[13] Note that S4 does not hold with non-mandatory relationship-based attitudes: in this case, the agent may decide whether or not to adopt its acquaintance's mental attitudes.

This form of social mental shaping, based on instantiations of role relationships, enables us to identify and formalise a particular type of relationship between agents that reflects the mental link that binds them together. We introduce the following definition:

Definition 5.1. At time t_i, agents a_i and a_j are in a *social cognitive relationship* of type (r_i, r_j) iff, at t_i: (a) they are in a social relationship of type (r_i, r_j); and (b) at least one of the two agents adopts or keeps a mental attitude simply because it believes that the other holds that attitude. Formally, we have:

$$\forall a_i, a_j, \forall (r_i, r_j), \forall t_i \; cogn(a_i, a_j, (r_i, r_j))(t_i)$$
$$\equiv Infl(Att(a_i, \phi), rel(a_i, a_j, (r_i, r_j)))(t_i) \vee Infl(Att(a_j, \phi), rel(a_i, a_j, (r_i, r_j)))(t_i).$$

[13] As noted with roles, the agent may already hold a relationship-based mandatory attitude before it enters the relationship. In this case, establishing the relationship gives the agent new reasons for keeping that attitude. Furthermore, note that an agent may also have the authority to impose upon another agent intentions that it does not hold. This case is not captured by our model since we define social mental shaping in terms of an agent's being influenced by another to adopt one of the latter's mental attitudes. We leave this topic for future investigation.

Informally, the operator $cogn(a_i, a_j, (r_i, r_j))(t_i)$ means that, at time t_i, agents a_i and a_j, are in a social cognitive relationship of type (r_i, r_j). This takes place iff at t_i at least one of the agents is subjected to the social mental shaping exercised by the other.

Not only can social mental shaping occur within a relationship between two agents; more generally, it can also arise within a group of two or more agents linked together by a set of social relationships. An obvious example is a project team where a number of agents work together and may influence one another to adopt such mental attitudes as beliefs, goals or intentions. In such situations, two or more agents are connected to each other via a web of social cognitive relationships that give rise to a *social cognitive structure* (Krackhardt 1987). The next step is, therefore, to formalise such structures in which an agent can be involved in multiple social cognitive relationships. To this end, we first need to introduce the notion of role structure; drawing on this, we will then define the concepts of social structure and social relationship structure.

As mentioned in Section 3, a role structure is a set of interrelated roles. Formally, we have:

Definition 5.2. A *role structure RS* is a tuple $\{R, T\}$, where:

- $R = \{r_i, r_j, \ldots, \} \neq \varnothing$ is a non-empty set of social roles; and
- $T = \{(r_i, r_j) \mid r_i, r_j \in R\} \neq \varnothing$ is a non-empty and weakly connected set of social relationship types.

We now define a social structure as a role structure instantiated by agents. Formally, we have the following definition:

Definition 5.3. A *social structure* SS_{t_i} is a tuple $\{RS, Ag_{t_i}\}$, where:

- $RS = \{R, T\}$ is a role structure; and
- $Ag_{t_i} = \{a_i, a_j, \ldots \mid \forall a_i \exists r_i \in R \text{ s.t. } In(a_i, r_i)(t_i)\} \neq \varnothing$ is a non-empty set of agents.

In a social structure, agents are weakly connected with one another through the relations that exist among the roles that they occupy. However, the agents of a social structure are not necessarily connected by social relationships (Section 4.4). In fact, they might be unaware of each other. Therefore, if we want to derive a social construct that describes the connections among the agents in terms of their social relationships, we need to enrich the above notion of social structure with a cognitive component that enables the agents to be aware of each other's instantiation of related roles. To this end, we now define the notion of social relationship structure as a social structure in which the agents are not only weakly connected as a result of the connections among the roles they occupy, but they are also aware of (some of) their connections. Formally, we have:

Definition 5.4. A *social relationship structure* SRS_{t_i} is a tuple $\{SS_{t_i}, Rel_{t_i}\}$, where:

- $SS_{t_i} = \{RS, Ag_{t_i}\}$ is a social structure; and
- $Rel_{t_i} = \{(a_i, a_j) \mid \exists (r_i, r_j) \in T \text{ s.t. } rel(a_i, a_j, (r_i, r_j))(t_i); a_i, a_j \in Ag\} \neq \varnothing$ is a non-empty and weakly connected set of pairs of agents that, at time t_i, are related to each other through social relationships.

In a social relationship structure, agents are weakly connected with one another, structurally through the relations that exist among the roles that they occupy, and cognitively through the mutual beliefs concerning each other's occupying related roles.

However, even though in a social relationship structure agents are cognitively related to one another, they do not necessarily impact upon one another's mental state. In fact, for example, they might well be able/willing/entitled to act in isolation and, therefore, they might not need to exercise any form of social influence upon others' mental states and behaviours. Or, alternatively, they might autonomously decide not to let their mental states and behaviours be affected by others' social influence. In these circumstances, the agents are not involved in social mental shaping processes and, therefore, they do not modify their mental states by adopting socially motivated attitudes. Thus, if we want to derive a social construct that reflects not only social relationships but also the inter-agent cognitive connections resulting from the social influence that agents exert upon one another, we need to couch the above notion of social relationship structure in terms of social cognitive relationships (Krackhardt 1987). This leads to the following definition of social cognitive structure:

Definition 5.5. A *social cognitive structure* SCS_{t_i} is a tuple $\{SRS_{t_i}, Cogn_{t_i}\}$, where:

- $SRS_{t_i} = \{SS_{t_i}, Rel_{t_i}\}$ is a social relationship structure; and
- $Cogn_{t_i} = \{(a_i, a_j) \mid \exists (r_i, r_j) \in T \text{ s.t. } cogn(a_i, a_j, (r_i, r_j))(t_i); a_i, a_j \in Rel_{t_i}\} \neq \varnothing$ is a non-empty and weakly connected set of pairs of agents that, at time t_i, are related to one another through social cognitive relationships.

Even though every social cognitive structure is a social relationship structure, nonetheless not every social relationship structure is a social cognitive structure. In fact, as is shown in the definition above, for a social relationship structure to be a cognitive structure, the agents must be weakly connected through a web of binary social cognitive relationships. In this view, a social cognitive structure may be seen as an extension of the concept of social cognitive relationship to those situations in which two or more agents are involved in complex forms of social mental shaping.

6. AXIOMS FOR CONSISTENCY

Thus far, we have described agents as autonomous cognitive entities, engaged in an iterated series of social actions and interactions aimed at completing their mental states. In a multi-agent system, roles and social relationships provide the agents with socially motivated mental attitudes that, when adopted, either replace or complement their individually motivated ones. Thus, the agent's mental state may be seen as constituted of two main types of mental attitudes: (a) individually motivated attitudes; and (b) socially motivated attitudes. In order for the agent to behave in a coherent and rational manner, these attitudes must be consistent with each other (Cohen and Levesque 1990; Rao and Georgeff 1991). In what follows, we shall briefly introduce the main axioms that underpin and govern consistency among socially motivated attitudes (Section 6.1) and between these and individually motivated ones (Section 6.2).

6.1. Consistency Among Socially Motivated Mental Attitudes

When an agent is socially influenced to modify its mental state by adopting a set of socially motivated mental attitudes, these attitudes must maintain a degree of consistency, in compliance with the axioms governing the agent's mental state (see Sections 4.2.1–4.2.4). As we mentioned in Sections 4.3.1.1–4.3.1.3, attached to roles may be mental attitudes that are inconsistent with one another. By adopting a role, an agent is

therefore confronted with such problems as role ambiguity, incompatibility and overload. Furthermore, an agent could be playing a number of roles that may impose different expectations, influences and responsibilities, some of which may well conflict. In order to behave rationally, the agent must overcome these inconsistencies by neglecting and/or modifying some role-based attitudes and maintaining others. To model this, we introduce a new set of axioms for consistency between socially motivated attitudes. These axioms also hold in the following two situations: (a) the agent is subjected to a number of social mental shaping processes outside of any social relationship and role (Section 5.2.1); and (b) the agent is subjected to different forms of influence exercised by another agent within a social relationship or, alternatively, it is involved in more than one social relationship and, therefore, is subjected to the influence of different acquaintances (Section 5.2.2). In either case, the agent may be induced by one or more agents to adopt a number of attitudes, some of which may well conflict. For example, an agent involved in two social relationships may be influenced by two other agents to adopt two conflicting goals. The agent must overcome such inconsistency in order to behave in a coherent and rational manner (see Section 4.2.3).

To model consistency between attitudes in the face of competing social influences, a number of axioms are introduced.[14] Note that these consistency axioms are similar to those introduced in Sections 4.2.1–4.2.4, and so their justification is similar. For simplicity, Att_{S1} and Att_{S2} are used to indicate two socially motivated mental attitudes that an agent may receive either from roles or from other agents (within or outside social relationships).

A1) Bel_{S1}/Bel_{S2}

An agent's socially motivated beliefs must not contradict each other.

[BB1] $\models Bel_{S1}(a_i, \phi)(t_i) \supset \neg Bel_{S2}(a_i, \neg\phi)(t_i)$.

Theorem 6.1. An agent's socially motivated beliefs must not contradict each other's implications:

$$\models Bel_{S1}(a_i, \phi)(t_i) \wedge Bel_{S1}(a_i, (\phi \supset \psi))(t_i) \supset \neg Bel_{S2}(a_i, \neg\psi)(t_i)$$

Proof. Since beliefs are closed under consequence (axiom K_B in Section 4.2.1), we have: $Bel_{S1}(a_i, \phi)(t_i) \wedge Bel_{S1}(a_i, (\phi \supset \psi))(t_i) \supset Bel_{S1}(a_i, \psi)(t_i)$. Thus, by applying axiom BB1, we have: $Bel_{S1}(a_i, \psi)(t_i) \supset \neg Bel_{S2}(a_i, \neg\psi)(t_i)$.

A2) Des_{S1}/Des_{S2}

According to the axiomatisation given in Section 4.2.2, an agent's desires need not be consistent with each other. They simply need to be closed under consequence. Therefore, like individually motivated desires, socially motivated ones may also contradict each other. For example, an agent may be influenced by another agent to adopt the desire to smoke, and at the same time emulate another agent by adopting its desire to lead a healthy life.

$$\not\models Des_{S1}(a_i, \phi)(t_i) \supset \neg Des_{S2}(a_i, \neg\phi)(t_i)$$

[14] In all the following axiom schemas, we will assume that the unbound variables are universally quantified as follows: $\forall a_i \in D_{Ag}, \forall t_i \in D_T, \forall w \in W$. In addition, in all the axiom schemas, we assume that ϕ and ψ can be replaced by any well-formed formulae in the language.

A3) $Goal_{S1}/Goal_{S2}$

An agent's socially motivated goals must not contradict each other.

[GG1] $\models Goal_{S1}(a_i, \phi)(t_i) \supset \neg Goal_{S2}(a_i, \neg\phi)(t_i)$.

Theorem 6.2. An agent's socially motivated goals must not contradict each other's implications:

$$\models Goal_{S1}(a_i, \phi)(t_i) \wedge Goal_{S1}(a_i, (\phi \supset \psi))(t_i) \supset \neg Goal_{S2}(a_i, \neg\psi)(t_i)$$

Proof. Since goals are closed under consequence (axiom K_G in Section 4.2.3), we have: $Goal_{S1}(a_i, \phi)(t_i) \wedge (Goal_{S1}(a_i, (\phi \supset \psi))(t_i) \supset Goal_{S1}(a_i, \psi)(t_i)$. Thus, by applying axiom GG1, we have: $Goal_{S1}(a_i, \psi)(t_i) \supset \neg Goal_{S2}(a_i, \neg\psi)(t_i)$.

A4) $Goal_{S1}/Bel_{S2}$

As goals are governed by the weak realism constraint (axiom G_1 in Section 4.2.3), an agent's socially motivated goals must not contradict its socially motivated beliefs. For example, if an agent adopts the role-based goal that ϕ will be eventually true, then the agent must not be socially influenced by another agent to adopt the belief that ϕ will be always false.

[GB1] $\models Goal_{S1}(a_i, \phi)(t_i) \supset \neg Bel_{S2}(a_i, \neg\phi)(t_i)$

A5) Int_{S1}/Int_{S2}

Socially motivated intentions must not contradict each other.

[II1] $\models Int_{S1}(a_i, \phi)(t_i) \supset \neg Int_{S2}(a_i, \neg\phi)(t_i)$

Theorem 6.3. An agent's socially motivated intentions must not contradict each other's implications:

$$\models Int_{S1}(a_i, \phi)(t_i) \wedge Int_{S1}(a_i, (\phi \supset \psi))(t_i) \supset \neg Int_{S2}(a_i, \neg\psi)(t_i)$$

Proof. Since intentions are closed under consequence (axiom K_I in Section 4.2.4), we have: $Int_{S1}(a_i, \phi)(t_i) \wedge Int_{S1}(a_i, (\phi \supset \psi))(t_i) \supset Int_{S1}(a_i, \psi)(t_i)$. Thus, by applying axiom II1, we have: $Int_{S1}(a_i, \psi)(t_i) \supset \neg Int_{S2}(a_i, \neg\psi)(t_i)$.

A6) Int_{S1}/Bel_{S2}

As with goals, socially motivated intentions must not contradict socially motivated beliefs (see axiom I_1 in Section 4.2.4).

[IB1] $\models Int_{S1}(a_i, \phi)(t_i) \supset \neg Bel_{S2}(a_i, \neg\phi)(t_i)$

6.2. Consistency Between Individually and Socially Motivated Attitudes

When an agent is subjected to a social mental shaping process, the mental attitudes it adopts must be consistent with the agent's individually motivated attitudes. These consistency relations are governed by the following group of axioms. For simplicity, with Att_I and Att_S we indicate, respectively, an agent's individually and socially motivated mental attitude.

B1) Bel_I/Bel_S

An agent's individually motivated beliefs must not contradict its socially motivated beliefs.

[BB2] $\models Bel_I(a_i, \phi)(t_i) \supset \neg Bel_S(a_i, \neg\phi)(t_i)$

[BB3] $\models Bel_S(a_i, \phi)(t_i) \supset \neg Bel_I(a_i, \neg\phi)(t_i)$

Theorem 6.4. An agent's individually and socially motivated beliefs must not contradict each other's implications:

$$\models Bel_I(a_i, \phi)(t_i) \wedge Bel_I(a_i, (\phi \supset \psi))(t_i) \supset \neg Bel_S(a_i, \neg\psi)(t_i)$$
$$\models Bel_S(a_i, \phi)(t_i) \wedge Bel_S(a_i, (\phi \supset \psi))(t_i) \supset \neg Bel_I(a_i, \neg\psi)(t_i)$$

Proof. It follows from axiom K_B (Section 4.2.1) and axioms BB2 and BB3.

B2) Des_I/Des_S

From the axiomatisation given in Section 4.2.2, it follows that an agent's individually and socially motivated desires need not be consistent with each other.

$$\not\models Des_I(a_i, \phi)(t_i) \supset \neg Des_S(a_i, \neg\phi)(t_i)$$
$$\not\models Des_S(a_i, \phi)(t_i) \supset \neg(Des_I(a_i, \neg\phi)(t_i)$$

B3) $Goal_I/Goal_S$

An agent's individually and socially motivated goals must not contradict each other.

[GG2] $\models Goal_I(a_i, \phi)(t_i) \supset \neg Goal_S(a_i, \neg\phi)(t_i)$

[GG3] $\models Goal_S(a_i, \phi)(t_i) \supset \neg Goal_I(a_i, \neg\phi)(t_i)$

Theorem 6.5. An agent's individually and socially motivated goals must not contradict each other's implications:

$$\models Goal_I(a_i, \phi)(t_i) \wedge Goal_I(a_i, (\phi \supset \psi))(t_i) \supset \neg Goal_S(a_i, \neg\psi)(t_i)$$
$$\models Goal_S(a_i, \phi)(t_i) \wedge Goal_S(a_i, (\phi \supset \psi))(t_i) \supset \neg Goal_I(a_i, \neg\psi)(t_i)$$

Proof. It follows from axiom K_G (Section 4.2.3) and axioms GG2 and GG3.

B4) $Bel_I/Goal_S$

As goals are governed by the weak realism constraint (axiom G_I in Section 4.2.3), an agent's socially motivated goals must not contradict its individually motivated beliefs. For example, if an agent has a socially motivated goal that ϕ will eventually be true, then it must not have an individually motivated belief that ϕ will always be false.

[GB2] $\models Goal_S(a_i, \phi)(t_i) \supset \neg Bel_I(a_i, \neg\phi)(t_i)$

B5) $Goal_I/Bel_S$

Similarly, an agent's individually motivated goals must not contradict its socially motivated beliefs.

[GB3] $\models Goal_I(a_i, \phi)(t_i) \supset \neg Bel_S(a_i, \neg\phi)(t_i)$

B6) Int_I/Int_S

An agent's individually and socially motivated intentions must not contradict each other.

[II2] $\models Int_I(a_i, \phi)(t_i) \supset \neg Int_S(a_i, \neg\phi)(t_i)$

[II3] $\models Int_S(a_i, \phi)(t_i) \supset \neg Int_I(a_i, \neg\phi)(t_i)$

Theorem 6.6. An agent's individually and socially motivated intentions must not contradict each other's implications:

$$\models Int_I(a_i, \phi)(t_i) \wedge Int_I(a_i, (\phi \supset \psi))(t_i) \supset \neg Int_S(a_i, \neg\psi)(t_i)$$
$$\models Int_S(a_i, \phi)(t_i) \wedge Int_S(a_i, (\phi \supset \psi))(t_i) \supset \neg Int_I(a_i, \neg\psi)(t_i)$$

Proof. It follows from axiom K_I (Section 4.2.4) and axioms II2 and II3.

B7) Int_I/Bel_S

Like with goals, an agent's individually motivated intentions must not contradict its socially motivated beliefs (see weak realism constraint I_I in Section 4.2.4).

[IB2] $\models Int_I(a_i, \phi)(t_i) \supset \neg Bel_S(a_i, \neg\phi)(t_i)$

B8) Bel_I/Int_S

Similarly, an agent's socially motivated intentions must not contradict its individually motivated beliefs.

[IB3] $\models Int_S(a_i, \phi)(t_i) \supset \neg Bel_I(a_i, \neg\phi)(t_i)$

The axioms that have been presented so far govern the agent's social mental shaping by imposing a degree of consistency on its mental state. When an agent is involved in a number of social actions and interactions, it is expected to make a range of decisions about which attitudes to adopt or to keep and which to abandon and/or modify. On the one hand, when confronted with mandatory socially motivated attitudes (Sections 4.3 and 5.2.2), the agent must adopt them and then decide which are the competing attitudes to be dropped and/or modified. On the other, when confronted with optional socially motivated attitudes, the agent is expected to select which of the competing attitudes is to be maintained. Accordingly, all the other conflicting attitudes will be dropped and/or modified. This decision can be based either on the agent's *preference* function or on a domain-dependent *influence* function (Cavedon and Sonenberg 1998). Thus, an agent may just locally decide to prefer one attitude among a set of competing ones. Or, alternatively, an attitude may be considered the most important and influential to an agent among a set of conflicting ones. This is an important issue that we leave for further investigation.

It now remains to discuss the primary forms of social cognitive relationships (see Definition 5.1 in Section 5.2.2; and Section 7), and to give some indications concerning the application of our model to structuring and analysing multi-agent systems (Section 8).

7. FORMS OF SOCIAL COGNITIVE RELATIONSHIPS

To illustrate the power and flexibility of the concept of social mental shaping, in this section a number of fundamental forms of social cognitive relationships will be introduced and examined by showing how they fit within the conceptual and formal framework. The chosen forms are fundamental in the sense that they represent the main ways in which social mental shaping may take place within social relationships between agents. Also, the chosen forms of social cognitive relationships underpin a wide variety of social phenomena, such as cooperation, negotiation and collaborative decision-making. Note that, for simplicity, the focus will be restricted to social relationships and the varying forms of social mental shaping that take place within them, even though such forms might well occur also outside of any pre-existing social relationship.

7.1. Imitation

Imitation is a form of social mental shaping grounded on desires. In this view, imitation takes place as long as agent a_i believes that another agent a_j has a desire towards ϕ, and this motivates a_i to adopt or to keep such a desire towards ϕ. Formally, we have:

$$\forall a_i, a_j, \forall t_i \, Imit(a_i, a_j, \phi)(t_i) \equiv \exists (r_i, r_j) \text{ s.t. } Infl(Des(a_i, \phi), rel(a_i, a_j, (r_i, r_j)))(t_i)$$

For example, the relationship between husband and wife may be such that they are mutually influenced to adopt each other's desires, perhaps in an effort to contribute to each other's happiness. In this view, they imitate each other. Imitation may occur mainly as a result of a variety of reasons related to love, friendship, mutual trust, or empathy. It refers to the spontaneous adoption of others' innermost psychological attributes and inclinations (Coleman 1990). More generally, imitation may take place either outside of any social relationship or within social relationships based on emotional (e.g. friendship) or normative exchange (e.g. authority relations) (Mitchell 1973). For example, should two agents be involved in a normative exchange, one of them, perhaps affected by the other's charm or social status, may spontaneously imitate and adopt some of the latter's desires.

7.2. Adoption

We call adoption that form of social mental shaping that refers to goals. Formally, we have:

$$\forall a_i, a_j, \forall t_i \, Adopt(a_i, a_j, \phi)(t_i) \equiv \exists (r_i, r_j) \text{ s.t. } \quad Infl(Goal(a_i, \phi), rel(a_i, a_j, (r_i, r_j)))(t_i).$$

For example, a salesman may adopt the goal of selling ten life insurance policies in a week because he believes that a rival salesman in the same company has the same goal. As a result of the competition between the two salesmen, therefore, the former adopts the latter's goal. From a more general perspective, adoption is a key form of social mental shaping that lies at the heart of a variety of social phenomena ranging from competitive to cooperative forms of social behaviour (Coleman 1990). On the one hand, like in the example above, an agent may be strategically influenced to adopt another's goal based on its own behavioural tactics suited to competitive environments (e.g. markets or negotiation processes) (Faratin et al. 1998). On the other, an agent's adoption of another's goal may be regarded as a pro-social behaviour leading to forms of joint cooperative activities (e.g. collaborative decision-making) (Panzarasa et al. 2001a). Furthermore, in social sciences, adoption has been conceptualised both in terms of spontaneous internalisation of others' goals and in terms of the effects that others' behaviours and/or expectations may have upon an agent's mental state (Baron et al. 1992). Firstly, an agent may spontaneously adopt another's goal as a result of variety of reasons related to emotional exchange (love, friendship, trust, pity, altruism and so forth). In such circumstances, adoption may be mainly regarded as a benevolent form of social mental shaping, whereby an agent adopts another's goal simply for the welfare of the latter. However, there are also circumstances in which goal adoption can be seen as a selfish form of social mental shaping that is instrumental to the achievement of individual intentions/goals. For example, should adoption be the inferential conclusion of an agent's practical reasoning regarding how to achieve some selfish ends, then it would represent a non-benevolent form of social mental shaping instrumental to the

agent's non-altruistic intentions (Panzarasa et al. 2001a). Secondly, goal adoption may occur as a result of an agent's request or simply as a result of an agent's intention to impact upon another's mental state.[15] In such cases, unlike spontaneous forms of goal adoption, an agent is induced to change its mental state by external pressures towards conformity to others' mental states (Baron et al. 1992). Most of the time, an agent's goal adoption induced by another's (explicit or implicit) request is instrumental to the achievement of (some of) the latter's intentions/goals.

7.3. Adhesion

Adhesion is that form of social mental shaping that is grounded on intentions: an agent may be motivated to hold an intention by the fact that it believes that another agent has that intention. Formally, we have:

$$\forall a_i, a_j, \forall t_i \; Adhes(a_i, a_j, \phi)(t_i) \equiv \exists (r_i, r_j) \text{ s.t. } Infl(Int(a_i, \phi), \; rel(a_i, a_j, (r_i, r_j)))(t_i)$$

For example, a secretary may adopt the intention that a report is completed by some deadline because he or she believes that his or her boss has this intention. In this view, the secretary adheres to the boss's intention. In this example, the secretary, by adhering to the boss's intention, will eventually behave in such a way that the intention will be fulfilled and the report completed by the deadline. In this case, the secretary's adhesion may be motivated by the intention that the boss's intentions are fulfilled. However, there may be also forms of adhesion that are simply induced by trust, respect, loyalty, etc., and therefore are unrelated to the intention that the other party achieves his or her own intentions. For example, the secretary may adopt his or her boss's intentions simply as a result of the former's confidence in the latter's expertise or, alternatively, the former's sensitivity about the latter's high social status. In these cases, the secretary may adopt the boss's intentions even though the former believes that the latter can fulfill his or her intentions in isolation, regardless of any assistance that others can provide. In turn, in these circumstances, the secretary's adhesion may be instrumental to the achievement of his or her selfish intentions/goals.

More generally, adhesion, like goal adoption, can be either benevolent or selfish (Coleman 1990). On the one hand, benevolent adhesion means that an agent adheres to another's intention because the former intends to contribute to the latter's welfare by fulfilling its intention. On the other, selfish adhesion means that an agent adheres to another's intention simply in order to have its own selfish intentions/goals fulfilled. Furthermore, like goal adoption, adhesion can be either spontaneous or motivated by others' requests (Baron et al. 1992; Milgram 1974). However, unlike goal adoption, adhesion can take place as a result not only of others' requests but also of the authority that an agent can exercise over another. In this case, when adhesion is intertwined with authority relations, an agent is obliged by another to internalise and adopt a relationship-based mandatory intention (Coleman 1990) (see Section 5.2.2). Typically,

[15]Note that an agent cannot be obliged by another to adopt a goal as a result of the authority that the latter can exercise over the former. The main reason why, unlike adhesion (see Section 7.3), goal adoption cannot be regarded as prescriptive (e.g. based on agents' authority/power) is that the only way to determine whether or not norms, prescriptions, rules and commands are being followed is to observe the agents' behaviour. Since in our formalisation goals are only loosely related to behaviour (see Section 4.2.3), and since agents' mental states are not observable entities, it would be impossible to determine whether or not an agent has adopted a goal (Clark 1998). As a result, a norm or an authority relation that relies on goal adoption would be mainly ineffective and fairly easy to escape. The same considerations also explain why only role-based intentions, and not goals, reflect the rules, norms, etc. that govern a multi-agent system (see Section 4.3.1).

adhesion to mandatory intentions is a normative form of social mental shaping that refers to the influence caused by the (explicit or implicit) rewards and punishments that an agent controls and to which another agent is subject (Baron et al. 1992).

7.4. Exploitation

In general, agents will neither store nor process information in costly ways when they can use their social environment as a convenient stand-in for the information-processing operations concerned. Particularly, within certain social relationships, it is often the case that an agent can easily lean on the information possessed by another agent to get the job done. In such cases, the agent extends its own mental state so as to include the other agent's beliefs. And such a mental extension is just based on the agent's belief that the other agent has a belief. Such a form of social mental shaping is here referred to as exploitation. Formally, we have:

$$\forall a_i, a_j, \forall t_i \, Expl(a_i, a_j, \phi)(t_i) \equiv \exists (r_i, r_j) \text{ s.t. } Infl(Bel(a_i, \phi), rel(a_i, a_j, (r_i, r_j)))(t_i).$$

Again, an example may help understand the meaning of the definition above. Fred, the muscle of a two person delivery team, may believe that there is a delivery schedule for the day because he believes that Ann, the van driver, has the schedule—and hence believes there to be a schedule. In fact, Fred may not have seen today's schedule (collecting and following the day's schedule is Ann's responsibility), but he comes to adopt this belief as a result of his relationship with Ann. Furthermore, and more interestingly, as Fred knows that Ann has the schedule, he also knows that, when necessary, he can rely on Ann to get the content of it and any further detail. In this view, Fred does not need to overload his mental state with the information owned by Ann. Simply, based on his relationship with her, Fred knows that, when necessary, he can exploit Ann's belief, and thus he can reduce the cognitive effort required to get his job done. In cognitive sciences, this form of social mental shaping is often described as reflecting a conception of the agent's mental state as escaping its "natural" confines and mingling with the physical and social environment (Clark 1998; Hutchins 1995). From a different perspective, in social and organisational sciences, exploitation is mainly regarded as related to such form of organisational memory often referred to as "transactive memory" (Carley 1990; Wegner 1987). This notion refers to the fact that agents, based on their interrelationships, develop a shared system for encoding, storing and retrieving information from different substantive domains. In this view, it is suggested that, drawing on their network defined in terms of "who knows who knows what," agents can exploit each other's beliefs and thus contribute to the system's performance by enhancing its efficiency and effectiveness (Wegner 1987).

8. SOCIAL MENTAL SHAPING IN CONTEXT

Having developed a formalisation of social mental shaping, it is now interesting to give some general indications as to how this formalisation can be used in multi-agent system research, both for theoretical and practical purposes. To this end, in this Section the model of social mental shaping will be applied to an extended example from a real-world domain, and it will be placed in the broader context of negotiation and practical reasoning in multi-agent systems (see Panzarasa et al. (2001b) for details).

The domain of interest here is the sale of a privately held company through the public offering of shares to relatively small investors and, possibly, through the involvement of an active investor. On the one hand, small investors are passive potential shareholders with virtually no monitoring capabilities. On the other, an active investor is an individual or group that is interested in investing in the company by purchasing a controlling interest and by bringing expertise or other resources to the firm. Finally, the selling agent (i.e., the representative of the company concerned) can be seen as motivated by the intention that shares in the company are sold, and through this, by the intention to release the maximum amount of revenue.

Building on Mello and Parsons (1998), four distinct strategies can be identified that may be used by the seller of shares in a privately held company. These are:

1. To search for an active investor, attempt to establish a price for the sale of a controlling block with that investor, and then offer the remainder of the shares to small investors ("Sequential 1").
2. To offer shares to small investors, but retain a controlling block to be subsequently offered to an active investor ("Sequential 2").
3. To negotiate with an active investor and simultaneously to offer shares to the small investors ("Parallel").
4. To sell all shares to small investors at a uniform price, without any attempt to involve an active investor ("Public Offering").

Each strategy reflects different patterns of social cognitive relationships between the agents involved. In order to examine these relationships, prototype agents representing the seller, the active investor and the small investors have been implemented using UM-PRS (Lee et al. 1994)—an implementation of the Procedural Reasoning System (Georgeff and Lansky 1987). Each selling strategy is specified as a "Knowledge Area" (KA), or predetermined plan. The details of these KAs are not relevant to this paper, but the strategy selected by the selling agent may have an impact on the mental shaping that can be exerted during the enactment of the strategy.

In the first three strategies, establishing a relationship with a potential active investor is essential. It is important for the selling agent to identify an active investor that will, in principle, be prepared to purchase a controlling share in the company. Should an acceptable price be agreed upon, the active investor will commit itself to pay that price for a controlling interest in the firm. Purchasing a controlling interest in the firm is a goal of the active investor's and is a candidate for being moved to intention status. The question remains: where does this goal come from? In its simplest form, the goal may be internally generated by the active investor, regardless of any relations with other agents. However, the model described in this paper provides for an alternative answer: the relationship established between the selling agent and the active investor can be regarded as a social cognitive relationship in which the active investor adopts the goal of the selling agent to have a controlling interest in the firm sold:

$$Adopt(active\text{-}investor, selling\text{-}agent, has(active\text{-}investor, controlling\text{-}interest))$$

The establishment of this relationship, and the related goal adoption by the active investor, will enable the selling agent to pursue one of the first three strategies. However, this is not the only social cognitive relationship that can be established between the selling agent and the active investor. Indeed, there are other forms of social cognitive relationships that are more closely related to the agreement as to the price at which shares are to be sold. In this respect, there are a number of variables that contribute to

an assessment of an acceptable price for shares in the firm. The "book-value" of a firm reflects the turnover and the assets of the company, and is information in the public domain. Furthermore, the active investor may bring further assets to the company in the form of expertise or other resources; this will tend to lower the price paid per share by the active investor, and is often referred to as the discount (Mello and Parsons 1998). Having a controlling interest in the company will also bring private benefits to the active investor; this will tend to raise the price paid per share, and is often referred to as the premium (Mello and Parsons 1998). It is in the interests of the active investor that the selling agent is aware of those assets that the active investor will bring to the company. Therefore, it is in the interests of the active investor to establish a relationship with the selling agent in which the active investor can influence the selling agent to believe that further assets will be brought to the firm:

$$Expl(\textit{selling-agent}, \textit{active-investor}, \textit{has-asset}\,(\textit{active-investor}, ?\textit{asset}))$$

Similarly, a social cognitive relationship may be established in which the active investor is influenced to exploit a belief of the selling agent's. Consider strategy "Sequential 2": the selling agent offers shares to the small investors, but retains a controlling share to be offered to the active investor. Even though, initially, the market value of the firm can only be estimated using the book-value, nonetheless after the sale of a minority stake to various small investors, the selling agent can gain a better estimate of the market value of shares in the company. Now, once the seller has gathered this new piece of information, it may be interested in making the active investor update its beliefs about the market value of shares. To this end, a social cognitive relationship can be established in which the active investor is influenced by the selling agent to exploit the latter's beliefs:

$$Expl(\textit{active-investor}, \textit{selling-agent}, \textit{market-value}(?\textit{value}))$$

These patterns of social cognitive relationships are now used to undertake a set of virtual experiments with simulated data. In particular, we will examine whether and to what extent the choice of a selling strategy can impact upon the negotiating agents' mental states and, ultimately, upon the generation of an agreement and the performance of a going public process. Since each selling strategy reflects different forms of social cognitive relationships, this study will enable us also to examine the impact that social cognitive relationships and social mental shaping can have upon the performance of joint behavioural processes. In this respect, an interesting interpretation and comparison of the selling strategies in terms of their underpinning social cognitive relationships can be as follows. With strategy "Sequential 1," the seller tries to impact upon the active investor's mental state by offering him or her a price with no additional information, and then approaches the small investors and influences them by offering a price and letting them know whether or not a controlling block has been sold. With strategy "Sequential 2," the seller impacts on the small investors' mental states by offering them a price with no additional information, and then impacts upon the active investor's mental state by offering a price and letting him or her know whether or not shares have been sold to the small investors. A variant of strategy "Sequential 2" can be introduced, namely "Strategy 2 + Influence", which differs from the former in that the seller tries to exert social influence more convincingly upon the active investor by providing him or her with more detailed information on which the price being offered is made contingent (e.g. the price already established with other investors). As a result of this, the active investor is expected to be more motivated to accept the offer received

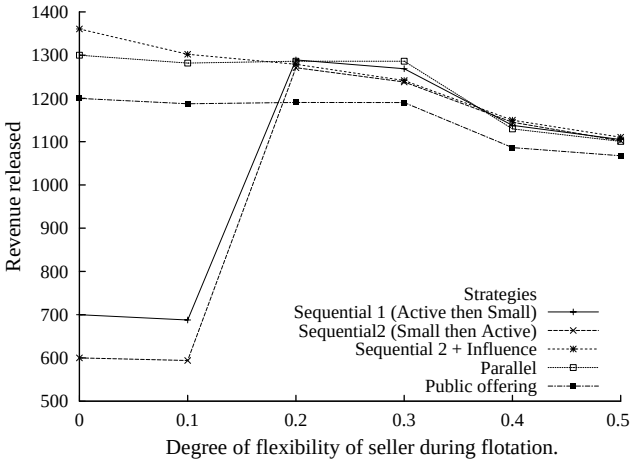

FIGURE 1. Revenue released from the sale of shares at different concession rates and using different selling strategies.

and act accordingly than would be the case if the seller had forwarded only a price without any further information that might justify it. With strategy "Parallel," the seller impacts simultaneously upon both the active investor's and the small investors' mental states, by letting them know about the parallel negotiation processes. Finally, with strategy "Public Offering," the seller approaches and influences the small investors only with a price and no further information.

Figure 1 illustrates the revenue released from the sale of the firm as the degree of flexibility of the selling agent changes when different strategies are used. The seller's flexibility is modelled as a concession rate reflecting the degree to which the seller will deviate from his or her ideal share price during negotiation. As is shown in Figure 1, strategies "Sequential 1" and "Sequential 2" perform poorly if the seller uses low concession rates during negotiation. An interpretation of these results is that, when establishing the share price with the active investor (in "Sequential 1") or the small investors (in "Sequential 2"), the seller gains information about the value of shares in the company. This information may bias the price that the seller may expect to receive during subsequent negotiation with the small investors (in "Sequential 1") or the active investor (in "Sequential 2"). This increased expectation of reward along with inflexibility may imply that the seller is unable to establish a share price with the small investors when using strategy "Sequential 1" and with the active investor when using strategy "Sequential 2." However, should strategy "Sequential 2 + Influence" be adopted, the seller may risk a less flexible strategy during negotiation to obtain a higher revenue. In fact, with this strategy, the seller influences the active investor by communicating an additional piece of information so as to justify the offer made. In turn, a greater amount of information makes social mental shaping more effective, and ultimately helps the seller to obtain higher revenue. Another result that is evident from Figure 1 is that, at medium and high concession rates, a public offering of all shares at a uniform price is associated with the worst performance. In such situations, a sequential sale is more effective as it allows the seller to obtain more information on which subsequent social mental shaping can be made contingent and thus more effectively exerted.

These preliminary experiments clearly illustrate how our model of social mental shaping can be used to investigate different social cognitive relationships between

agents, and how these relationships affect the behaviour of the agents involved. By viewing interactions in terms of influence processes, we abstract away from the specifics of any particular algorithm for cooperation, coordination or negotiation. Thus, our framework enables an application designer to concentrate on the essence and cognitive foundations of inter-agent social interactions, without worrying about how these relationships can actually be enacted. In this sense, the model is also a useful tool in the development and understanding of practical applications of multi-agent systems.

9. RELATED WORK

While BDI architectures conceptualise intentional behaviour, they say nothing about the social aspects of agents being situated in a multi-agent system (see Section 2). On the other hand, architectures for interacting agents have mainly focused on mechanisms and structures for managing the coordination process and have paid less attention to the agent's decision-making apparatus. There are, however, a number of architectures that attempt to incorporate both facets of agenthood into a single model (Cavedon et al. 1997; Lux and Steiner 1995). Such architectures are typically layered. They involve separate components for local decision-making and for social behaviour, combined within a control framework that seeks to strike a balance between the two. Unfortunately, in most cases, the link between the two main components is somewhat *ad hoc* in nature and has no clear underpinning conceptual model.

By focusing on the problem of sociality among autonomous agents, this work can be regarded as closely related to interaction-based architectures such as STEAM (Tambe 1997), GRATE* (Jennings 1995), and MECCA (Lux and Steiner 1995). However, our work does not deal with the incorporation of cooperative abilities into an agent framework through additional social constructs and operators (such as joint intentions and common goals). Rather, we work within a standard BDI architecture and provide a theoretical model that extends BDI logic so as to account for sociality among autonomous deliberative agents. In doing this, we broaden the set of an agent's mental attitudes so as to encompass not only internally motivated attitudes, but also socially motivated attitudes that the agent adopts as a result of its being situated within a social environment.

Our approach has some commonality with models based on social commitment and social attitudes (as discussed in Section 2). Castelfranchi (1995) states that both individual commitment and collective commitment are *internal* commitments of either an individual agent or a group of agents. Accordingly, he claims that we need a notion of social commitment that captures the dependence relationships between agents. Cavedon et al. (1997) introduce modalities corresponding to social attitudes. These concepts reflect relations between two agents (or teams), with respect to a proposition, and are modelled upon Castelfranchi's notion of generic commitments to performing actions of a given type. Thus, social attitudes turn out to be a useful means for incorporating commitment-based social interactions within a BDI framework. However, our notion of social mental shaping is more general, since it captures different forms of social interactions that do not necessarily reflect any commitment between agents (e.g. spontaneous help, and trust-based relations). In addition, our approach addresses the problem of how roles impact upon an agent's mental state, outside of any particular social relationship.

Our formal framework is informed by the work of Cavedon and Sonenberg (1998). They use roles as an abstraction that enables the agent designer to model the sphere of influence of one agent with respect to another. They associate goals with roles. This enables them to express generic social commitment: whenever an agent adopts a role,

this involves a commitment to taking on the goal attached to that role. So conceived of, roles provide a way to specify how the agent should balance competing obligations from different relationships, and overcome the tensions between personal preferences and social obligations. Our work extends Cavedon and Sonenberg's approach, in that it associates roles not only with goals but also with beliefs and intentions. In addition, we formalise the impact that an instantiation of a relation between roles has on an agent's mental state.

Our use of roles is similar to that of Werner (1989), in that he also associates roles with mental attitudes. Werner defines a role as a description of an abstract agent that is endowed with state information, permissions, responsibilities, and values. When an agent assumes a role, it internalises that role by adopting the mental attitudes that are attached to it. Likewise, our notion of role structure is similar to Werner's notion of social structure, in that both notions refer to a set of interrelated social roles. However, while Werner uses roles and social structure to develop a theory of social cooperation for multi-agent systems, we go further than he does and refer to these notions to cover a wider range of social inter-agent behaviour (e.g. social exchange, help, unilateral cooperation). Finally, our use of roles differs from that of Barbuceanu (1997), who associates roles with functions in an organisation, as well as with obligations towards others. Our definition of roles as a set of mental attitudes permits us to provide a unified framework in which, at a more general level, we can investigate the influence that the role structure has not only on the role-player's goals and intentions, but also on its beliefs.

10. CONCLUSIONS AND FUTURE WORK

The aim of this paper was to investigate the impact that sociality has on an agent's mental state, and to show how this impact can be formalised within a BDI framework. The term "social mental shaping" was introduced to refer to the phenomenon that the mere social nature of agents affects, and thereby alters, their mental states, thus motivating their behaviour. The process of social mental shaping may involve social roles, social relationships, or it may take place outside of any role or relationship. Its typical outcome is a modification of the social agent's mental state: the agent may adopt new socially motivated mental attitudes, or it may be provided with new reasons for keeping attitudes generated at an earlier stage.

Here, the focus has mainly been on that kind of social mental shaping that captures the impact that both roles and social relationships have on an agent's mental state. To address this, the structure of a multi-agent system was modelled in terms of roles and relations between roles. On the one hand, when an agent takes on a role, it may adopt the mental attitudes that are attached to that role. On the other, when an agent interacts with an acquaintance, it can be influenced by that agent to adopt or to keep a mental attitude. The novelty of our particular use of roles and social relationships rests on the fact that it allows us to capture important ingredients of sociality within a BDI logic. Interacting and deliberative agents can thus be modelled in terms of both their internally motivated mental attitudes, and the impact that other agents and the role structure of a multi-agent system have on their mental states. Indeed, roles and social relationships provide the agent with new socially motivated mental attitudes or new reasons for keeping individually motivated ones. In order for the social agent to behave in a coherent and rational manner, individually and socially motivated attitudes must be consistent with each other. In this respect, we introduced the main axioms

governing such consistency among attitudes. Finally, we concentrated on social cognitive relationships, and some of the main forms of these relationships have been explored.

Future work involves the investigation of the issues of *why* and *when* social mental shaping arises. Further attention also needs to be paid to how social mental shaping develops over time. An interesting scenario involves the case where social mental shaping is persistent for some time but is essentially dynamic. For example, an agent may join a multi-agent system, adopt some socially motivated mental attitudes, but later relinquish them by changing its role or dropping out of a social relationship.

Another important aspect of social mental shaping that deserves future investigation is the issue of *inconsistency*. In Section 6 we introduced a number of axioms underpinning consistency for attitudes. However, we need to study the conceptual mechanisms underlying an autonomous agent's behaviour when more than one role and/or social relationship are involved. As different roles and social relationships may have different impacts on the agent's mental state and some of them may well conflict (Cavedon and Sonenberg 1998), it is important to specify a local *preference function* in order for the agent to decide how to overcome this inconsistency. Further, different roles and social relationships may impose conflicting influences and responsibilities on the agent because adopting a socially motivated mental attitude precludes the agent from adopting a competing one. To model this, we need to investigate a domain-dependent *influence function* that asserts what is more influential to an agent in a given situation. For example, one relationship may be more influential as one of the involved agents has significant control or power over its acquaintance. Or, at a given time, one role may be more crucial to an agent than any other role it may also be playing. This also involves the investigation of decision theoretic techniques for imposing constraints on the agent's influence function; for example, one role may be considered more critical than another if occupying it brings about more utility to the role-player or the multi-agent system as a whole.

Moving on to the logic developed in this paper, there are a number of obvious areas for future work. We seek to provide a more detailed account of the relations between modalities expressing individually and socially motivated mental attitudes. Particularly, our focus will be on the cognitive modelling of roles, and on the complex process by which role-based attitudes are internalised by the role-player and tied to its internal attitudes in the course of the social mental shaping process. Further, in our model we have said nothing about socially motivated *joint* mental attitudes. For example, a group of agents may be influenced by another group or by an individual agent to adopt a joint attitude. Or, a group may adopt a joint attitude as a consequence of its members' adopting role-based attitudes. Future work needs therefore to highlight the key steps of the process in which individual agents' socially motivated attitudes are meshed together until a joint mental attitude ensues.

REFERENCES

BARBUCEANU, M. 1997. Coordinating agents by role-based social constraints and conversation plans. *In* Proceedings of the Fourteenth National Conference on Artificial Intelligence, pages 16–21.

BARNARD, C. I. 1938. The Functions of the Executive, Harvard University Press, Cambridge, MA.

BARON, R. S., N. L. KERR, and N. MILLER. 1992. Group Process, Group Decision, Group Action, Open University Press, Buckingham.

BELL, J. 1995. Changing attitudes. *In* M. Wooldridge and N. R. Jennings, editors, Intelligent Agents, Proceedings of the First International Workshop on Agent Theories, Architectures, and Languages, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, pages 40–55.

BELL, J., and Z. HUANG. 1997. Dynamic goal hierarchies. *In* L. Cavedon, A. Rao & W. Wobcke, editors, Intelligent Agent Systems: Theoretical and Practical Issues, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, pages 88–103.

BRATMAN, M. E. 1987. Intentions, Plans, and Practical Reasoning, Harvard University Press, Cambridge, MA.

BRATMAN, M. E., D. ISRAEL, and M. E. POLLACK. 1988. Plans and resource-bounded practical reasoning. Computational Intelligence, **4**(4):349–355.

CARLEY, K. M. 1990. Group stability: A socio-cognitive approach. *In* E. Lawler, B. Markovsky, C. Ridgeway and H. Walker, editors, Advances in Group Processes: Theory and Research, Vol. II, JAI Press, Greenwhich, CN, pages 1–44.

CARLEY, K. M., and A. NEWELL 1994. The nature of the social agent. Journal of Mathematical Sociology, **19**(4):221–262.

CASTELFRANCHI, C. 1995. Commitments: From individual intentions to groups and organisations. *In* V. Lesser, editor, Proceedings of the First International Conference on Multi-Agent Systems, AAAI Press and MIT Press, San Francisco, CA, pages 41–48.

CAVEDON, L., and L. SONENBERG. 1998. On social commitment, roles and preferred goals. *In* Proceedings of the Third International Conference on Multi-Agent Systems, pages 80–86.

CAVEDON, L., A. RAO, and G. TIDHAR. 1997. Social and individual commitment. *In* L. Cavedon, A. Rao & W. Wobcke, editors, Intelligent Agent Systems: Theoretical and Practical Issues, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, pages 152–163.

CHELLAS, B. 1980. Modal Logic: An Introduction, Cambridge University Press, Cambridge, MA.

CLARK, A. 1998. Being There. Putting Brain, Body, and World Together Again, The MIT Press, Cambridge, MA.

COLEMAN, J. S. 1990. Foundations of Social Theory, The Belknap Press of Harvard University Press, Cambridge, MA.

COHEN, P. R., and H. J. LEVESQUE. 1990. Intention is choice with commitment. Artificial Intelligence, **42**:213–261.

COHEN, P. R., and H. J. LEVESQUE. 1991. Teamwork. Nous, **25**(4):487–512.

DENNETT, D. C. 1987. The Intentional Stance, The MIT Press, Cambridge, MA.

DURFEE, E. H. 1999. Practically coordinating. AI Magazine, **20**(1):99–116.

ERMAN, L. D., F. HAYES-ROTH, V. R. LESSER, and R. REDDY. 1980. The HEARSAY-II speech understanding system: Integrating knowledge to resolve uncertainty. Computing Surveys, **12**(2):213–253.

FAGIN, R., J. Y. HALPERN, Y. MOSES, and M. Y. VARDI. 1995. Reasoning about Knowledge, The MIT Press, Cambridge, MA.

FARATIN, P., C. SIERRA, and N. R. JENNINGS. 1998. Negotiation decision functions for autonomous agents. International Journal of Robotics and Autonomous Systems, **24** (3–4):159–182.

FRANKLIN, S., and A. GRAESSER. 1997. Is it an agent, or just a program? A taxonomy for autonomous agents. *In* J. P. Mueller, M. J. Wooldridge, and N. R. Jennings, editors, Intelligent Agents III: Proceedings of the Third International Workshop on Agent Theories, Architectures and Languages, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, pages 21–35.

GEORGEFF, M. P., and A. L. LANSKY. 1987. Reactive reasoning and planning. In Proceedings of the Sixth National Conference on Artificial Intelligence, Seattle, WA, pages 677–681.

GRIFFITHS, N., and M. LUCK. 1999. Cooperative plan selection through trust, *In* F. J. Garijo and M. Boman, editors, Multi-Agent Systems Engineering—Proceedings of the Ninth European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, pages 162–174.

GROSZ, B., and S. KRAUS. 1996. Collaborative plans for complex group actions. Artificial Intelligence, **86**: 269–357.

HANDY, C. B. 1993. Understanding Organisations, Penguin Books, fourth edition, Harmondsworth.

HINTIKKA, J. 1962. Knowledge and Belief, Cornell University Press, Ithaca, N.Y.

HINTIKKA, J. 1972. Semantics for propositional attitudes. *In* L. Linsky, editor, Reference and Modality, Oxford University Press, Oxford.

HUCZYNSKI, A., and D. A. BUCHANAN. 1991. Organizational Behaviour. An Introductory Text, Prentice Hall, second edition, New York, New York.

HUTCHINS, E. 1995. Cognition in the Wild, The MIT Press, Cambridge, MA.

JENNINGS, N. R. 1995. Controlling cooperative problem solving in industrial multi-agent systems using joint intentions. Artificial Intelligence, **75**:195–240.

JENNINGS, N. R. 2000. On Agent-Based Software Engineering. Artificial Intelligence, **117**: 277–296.

KINNY, D., M. LJUNGBERG, A. RAO, E. SONENBERG, G. TIDHAR, and E. WERNER. 1994. Planned team activity. *In* C. Castelfranchi and E. Werner, editors, Artificial Social Systems, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, pages 227–256.

KRACKHARDT, D. 1987. Cognitive social structures. Social Networks, **9**:109–134.

KRAUS, S., K. SYCARA, and A. EVENCHIL. 1998. Reaching agreements through argumentation: A logical model and implementation. Artificial Intelligence, **104**:1–69.

LEE, J., M. J. HUBER, E. H. DURFEE, and P. G. KENNY. 1994. UM-PRS: An implementation of the procedural reasoning system for multirobot applications. *In* Proceedings of the Conference on Intelligent Robotics in Field, Factory, Service, and Space, Houston, TX, pages 842–849.

LESSER, V. R., and D. D. CORKILL. 1983. The distributed vehicle monitoring testbed: A tool for investigating distributed problem solving networks. AI Magazine, **4**(3):15–33.

LINDAHL, L. 1977. Position and Change: A Study in Law and Logic, D. Reidel Publishing Company, Dordrecht.

LUX, A. and D. D. STEINER. 1995. Understanding cooperation: An agent's perspective. *In* Proceedings of the First International Conference on Multi-Agent Systems, pages 261–268.

MARSH, S. 1994. Trust in distributed artificial intelligence. *In* C. Castelfranchi and G. Werner, editors, Artificial Social Societies, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, pages 94–112.

MELLO, A. S., and J. E. PARSONS. 1998. Going public and the ownership structure of the firm. Journal of Financial Economics, **49**:79–109.

MILGRAM, S. 1974. Obedience to Authority, Harper & Row, New York, N.Y.

MITCHELL, J. C. 1973. Networks, norms and institutions *In* Mitchell J. C., Social Network in Urban Situation, Manchester University Press, Manchester.

MUELLER, J. P. 1996. The Design of Intelligent Agents. A Layered Approach, Springer-Verlag, Berlin.

NORMAN, T. J., and C. A. REED. 2001. Delegation and responsibility. *In* Y. Lesperance and C. Castelfranchi, editors, Intelligent Agents VII, Springer-Verlag, Berlin.

PANZARASA, P., T. J. NORMAN, and N. R. JENNINGS. 1999. Modeling sociality in the BDI framework. *In* J. Liu and N. Zhong, editors, Intelligent Agent Technology: Systems, Methodologies, and Tools, Proceedings of the First Asian Pacific Conference on Intelligent Agent Technology, World Scientific Publishing, pages 202–206.

PANZARASA, P., N. R. JENNINGS, and T. J. NORMAN. 2001a. Formalising collaborative decision-making and practical reasoning in multi-agent systems. Journal of Logic and Computation, forthcoming.

PANZARASA, P., N. R. JENNINGS, and T. J. NORMAN. 2001b. Going public and the sale of shares with heterogeneous investors: Agent-based computational modelling and computer simulations. International Journal of Group Decision and Negotiation, **10**(5), 423–470.

PARSONS, S., C. SIERRA, and N. R. JENNINGS. 1998. Agents that reason and negotiate by arguing. Journal of Logic and Computation, **8**(3):261–292.

RAO, A. S., and M. P. GEORGEFF. 1991. Modeling agents within a BDI architecture. *In* R. Fikes and E. Sandewall, editors, Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning, Morgan Kaufmann Publishers, San Mateo, CA, pages 473–484.

RAO, A. S., M. P. GEORGEFF, and E. A. SONENBERG. 1992. Social plans: A preliminary report. *In* E. Werner and Y. Demazeau, editors, Decentralized AI-3, Elsevier, Amsterdam, pages 55–77.

SIERRA, C., N. R. JENNINGS, P. NORIEGA, and S. PARSONS. 1998. A framework for argumentation-based negotiation. *In* M. P. Singh, A. Rao and M. J. Wooldridge, editors, Intelligent Agents IV: Proceedings of the Fourth International Workshop on Agent Theories, Architectures and Languages, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, pages 177–192.

SERGOT, M. E. 1998. Normative positions. *In* P. McNamara and H. Prakken, editors, Norms, Logics and Information Systems, IOS Press.

SIMON, H. A. 1957. Models of Man: Social and Rational, John Wiley, New York, New York.

SINGH, M. P. 1995. Multiagent Systems: A Theoretical Framework for Intentions, Know-how, and Communications, Springer-Verlag, Berlin.

TAMBE, M. 1997. Towards flexible teamwork. Journal of Artificial Intelligence Research, **7**:83–24.

VON WRIGHT, G. H. 1980. Freedom and determination. North Holland Publishing Co., Amsterdam.

WALTON, D. N., and E. C. W. CRABBE. 1995. Commitment in Dialogue: Basic Concepts in Interpersonal Reasoning, State University of New York Press, New York, New York.

WEGNER, D. M. 1987. Transactive memory: A contemporary analysis of the group mind. *In* B. Mullen, G. R. Goethals, editors, Theories of Group Behavior, Springer-Verlag, New York, NY.

WERNER, E. 1989. Cooperating agents: A unified theory of communication and social structure. *In* L. Gasser and M. N. Huhns, editors, Distributed Artificial Intelligence, Pitman/Morgan Kaufmann, London, pages 3–36.

WOOLDRIDGE, M. 2000. Reasoning About Rational Agents, The MIT Press, Cambridge, MA.

WOOLDRIDGE, M., and N. R. JENNINGS. 1995. Intelligent agents: Theory and practice. The Knowledge Engineering Review, **10**(2):115–152.

WOOLDRIDGE M., and N. R. JENNINGS. 1999. Cooperative problem solving. Journal of Logic and Computation, **9**(4):563–592.

WOOLDRIDGE, M., N. R. JENNINGS, and D. KINNY. 2000. The Gaia methodology for agent-oriented analysis and design. Journal of Autonomous Agents and Multi-Agent Systems, **3**(3): 285–312.

APPENDIX: THE COMPLETE DEFINITION OF THE FORMAL FRAMEWORK

This Appendix introduces the quantified multi-modal logic that has been devised to formally express our model of social mental shaping. The logic is called *L*. *L* is a many-sorted, first-order, multi-modal language which both draws upon and extends the work described in Bell (1995), Bell and Huang (1997), Cohen and Levesque (1990), Panzarasa et al. (2001*a*), Rao and Georgeff (1991), and Wooldridge and Jennings (1999).[16] The main feature of *L* is that it can be used for reasoning about agents, roles, social relationships, and mental attitudes, with explicit reference to time points and intervals. *L* is first-order because if allows quantification over terms. Furthermore, it is multi-modal: it contains a set of modal operators, which allow to reason about and represent the beliefs, desires, goals and intentions of agents. Finally, *L* is based on a model of time that is linear.

The following sections are intended to detail the formal language *L*. Firstly, in Section A, a syntax will be developed, with an account of the symbols, terms, and well-formed formulae through which our model of social mental shaping has been expressed. Secondly, in Section B, the semantics of the language will be defined.

A1. SYNTAX

Definition 1. The language *L* contains the following symbols:

[16]Our use of time is consistent with Bell's (1995) and Bell and Huang's (1997) work. Our formalisation of individual mental attitudes draws upon Cohen and Levesque (1990) and Rao and Georgeff (1991). Like Wooldridge and Jennings' language (1999), our logic contains a modality that expresses mutual beliefs. However, our language extends the aforementioned formal frameworks in that it contains terms for reasoning about roles and relationships, and it explicitly addresses the formalisation of socially motivated attitudes. The most significant point of departure is that our logic is based on a linear view of time, whereas Wooldridge and Jennings use a branching temporal model.

1. a countable set *Const* of constant symbols, the union of the pairwise disjoint sets $Const_T$ (time point constants), $Const_I$ (interval constants), $Const_{Ag}$ (agent constants), $Const_{Roles}$ (role constants), $Const_{RelTypes}$ (relationship type constants), $Const_O$ (other constants);
2. a countable set *Var* of variable symbols, the union of the mutually disjoint sets Var_T, Var_I, Var_{Ag}, Var_{Roles}, $Var_{RelTypes}$, Var_O;
3. a countable set *Pred* of predicate symbols—each symbol $P \in Pred$ is associated with a natural number called its arity, given by $arity(P)$;
4. the operator symbols *Bel, Des, Goal, Int, In, RoleBel, RoleGoal, RoleInt*, $\in$ and $=$;
5. the punctuation symbols ")", "(", "[", "]", and comma ",".

Definition 2. A term is either a constant or a variable. The sort of a term is either *Ag*, *T*, *I*, *Roles, RelTypes*, or *O*. The terms of sort agent, time point, interval, role, relationship type and object (the sets $term_{Ag}$, $term_T$, $term_I$, $term_{Roles}$, $term_{RelTypes}$ and $term_O$, respectively) are defined as follows:

- $term_S$ is the minimal set s. t. $Const_S \cup Var_S \subseteq term_S$, where $S \in \{Ag, T, Roles, RelTypes, O\}$;
- $\{[u, u'] \mid u, u' \in term_T\} \cup Var_I \subseteq term_I$

We denote by Var the set of all variables, by *Const* the set of all constants, and by *Terms* the set of variables and constants. Note that we demand that a predicate P is applied to $arity(P)$ terms.

Definition 3. The syntax of well-formed formulae (*wffs*) of the language L is defined as follows:

- If $t, t' \in term_T$ then $(t < t') \in wffs$.
- If $u_1, \ldots, u_n \in Terms$, $P \in Pred$, and $i \in term_I$ then $P(u_1, \ldots, u_n)(i) \in wffs$.
- If $u, u' \in Terms$ and $i \in term_I$ then $(u = u')(i) \in wffs$.
- If $a \in term_{Ag}$, $r \in term_{Roles}$ and $i \in term_I$ then $In(a, r)(i) \in wffs$.
- If $a \in term_{Ag}$, $\phi \in wffs$, and $i \in term_I$ then $Bel(a, \phi)(i) \in wffs$.
- If $a \in term_{Ag}$, $\phi \in wffs$, and $i \in term_I$ then $Des(a, \phi)(i) \in wffs$.
- If $a \in term_{Ag}$, $\phi \in wffs$, and $i \in term_I$ then $Goal(a, \phi)(i) \in wffs$.
- If $a \in term_{Ag}$, $\phi \in wffs$, and $i \in term_I$ then $Int(a, \phi)(i) \in wffs$.
- If $r \in term_{Roles}$, $\phi \in wffs$, and $i \in term_I$ then $RoleBel(r, \phi)(i) \in wffs$.
- If $r \in term_{Roles}$, $\phi \in wffs$, and $i \in term_I$ then $RoleGoal(r, \phi)(i) \in wffs$.
- If $r \in term_{Roles}$, $\phi \in wffs$, and $i \in term_I$ then $RoleInt(r, \phi)(i) \in wffs$.
- If $\psi, \chi \in wffs$ then $\neg\psi \in wffs$ and $(\psi \vee \chi) \in wffs$.
- If $S \in \{Ag, T, I, Roles, RelTypes, O\}$, $x \in Var_S$, and $\phi \in wffs$ then $\exists x \phi \in wffs$.

Definition 4. Relations and functions on time points and intervals ($t, t' \in term_T$; $i, i' \in term_I$):

- $t = t' \equiv \neg(t < t') \wedge \neg(t' < t)$
- $t \leq t' \equiv (t < t') \vee (t = t')$
- $min([t, t']) \equiv min(t, t')$
- $max([t, t']) \equiv max(t, t')$
- $i < i' \equiv max(i) < max(i')$
- $i = i' \equiv max(i) = max(i')$
- $i \leq i' \equiv (i < i') \vee (i = i')$
- $i = i' \equiv (min(i) = min(i')) \wedge (max(i) = max(i'))$
- $i \subset i' \equiv (min(i) = min(i')) \wedge (max(i) < max(i'))$

- $i \subseteq i' \equiv (i \subset i') \vee (i = i')$
- $i + 1 \equiv [min(i) + 1, max(i) + 1]$ if $min(i) = max(i)$, $[min(i), max(i) + 1]$ otherwise

A2. SEMANTICS

The purpose of the semantics is to assign some formal meaning to the syntactic objects of the language. This section introduces the formal semantics of L. The semantics are presented in three main parts: the first defines a model structure for L, the second gives the semantic rules for L, and the third presents the notions of *validity* and *satisfiability* for formulae of L.

It is assumed that the actual world w_0 may be any of a set W of possible worlds. D_T is a set of time points. The worlds in W are thought of as possible worlds which share a common flow of time (D_T, r_{DT}), where $r_{DT} \subseteq D_T \times D_T$. D_I represents a set of intervals that are defined in terms of time points.

The world is populated by a non-empty set D_{Ag} of agents. Agents have beliefs, desires, goals and intentions. The beliefs of an agent are given by a *Belief-accessibility relation* B on W in the usual way. B maps agents, time and worlds to possible worlds frames. For world w, agent a and interval i, the conditions on $W_{(Bel,a,i,w)}$ and $R_{(Bel,a,i,w)}$ capture the idea that $W_{(Bel,a,i,w)}$ is the set of $R_{(Bel,\,a,\,i,\,w)}$-accessible worlds from w. Similarly, we assume that the desires, goals and intentions of agents are given by, respectively, Desire-, Goal-, and Intention-accessibility relations on W.

Finally, the world contains a set of objects, D_O, a set of roles D_{Roles} and a set of relationship types, $D_{RelTypes}$. Attached to roles are a number of mental attitudes. As happens with agents' beliefs, role-based beliefs are given by a *Role-Belief-accessibility relation* on W. R_B maps roles, time and worlds to possible worlds frames. For world w, role r and interval i, the conditions on $W_{(RoleBel,\,r,\,i,\,w)}$ and $R_{(RoleBel,\,r,\,i,\,w)}$ capture the idea that $W_{(RoleBel,\,r,\,i,\,w)}$ is the set of $R_{(RoleBel,\,r,\,i,\,w)}$-accessible worlds from w. Similarly, we assume that the goals and intentions attached to roles are given by, respectively, Role-Goal- and Role-Intention-accessibility relations on W.

Definition 5. The domain of quantification, D, is $D_{Ag} \cup D_T \cup D_I \cup D_{Roles} \cup D_{RelTypes} \cup D_O$.

The language thus allows quantification over agents, time points, intervals, roles, relationship types and objects. Note that D is fixed for all worlds.

Definition 6. An interpretation for constants, V, is a sort-preserving bijection V: $Const \rightarrow D$. A variable assignment, g, is a sort-preserving bijection g: $Var \rightarrow D$.

Definition 7. A model M is a structure:

$$\langle W, w_0, D_{Ag}, D_T, r_{DT}, D_I, D_{Roles}, D_{RelTypes}, D_O, In, B, D, G, I, R_B, R_G, R_I, v, \Phi \rangle$$

where:

- W is a non-empty set of possible worlds;
- w_0 is a distinguished member of W;
- D_{Ag} is a non-empty set of agents;
- D_T is a non-empty set of time points;
- $r_{DT} \subseteq D_T \times D_T$;
- $D_I = \{[t, t'] \mid t, t' \in D_T\}$ is a non-empty set of intervals;
- D_{Roles} is a non-empty set of roles;
- $D_{RelTypes} \subseteq D_{Roles} \times D_{Roles}$ is a non-empty set of relationship types: $\{(r, r') \mid r, r' \in D_{Roles}\}$;

- D_O is a non-empty set of objects;
- $In \subseteq D_{Ag} \times D_{Roles} \times D_I \times W$;
- B: $D_{Ag} \times D_I \times W \rightarrow \mathcal{P}W \times \mathcal{P}(W \times W)$ is such that:
  1. for $\alpha = (a, i, w)$, $B\alpha = (W_{(Bel,\alpha)}, R_{(Bel,\alpha)})$ is centred at w, that is $w \in W_{(Bel,\alpha)}$, and $(w, w') \in R_{(Bel,\alpha)}$ for any $w' \neq w$ in $W_{(Bel,\alpha)}$;
  2. $R_{(Bel,\alpha)}$ is serial, transitive, and euclidean;
- D: $D_{Ag} \times D_I \times W \rightarrow \mathcal{P}W \times \mathcal{P}(W \times W)$ is such that for $\alpha = (a, i, w)$, $D\alpha = (W_{(Des,\alpha)}, R_{(Des,\alpha)})$ is centred at w, that is $w \in W_{(Des,\alpha)}$, and $(w, w') \in R_{(Des,\alpha)}$ for any $w' \neq w$ in $W_{(Des,\alpha)}$;
- G: $D_{Ag} \times D_I \times W \rightarrow \mathcal{P}W \times \mathcal{P}(W \times W)$ is such that:
  1. for $\alpha = (a, i, w)$, $G\alpha = (W_{(Goal,\alpha)}, R_{(Goal,\alpha)})$ is centred at w, that is $w \in W_{(Goal,\alpha)}$, and $(w, w') \in R_{(Goal,\alpha)}$ for any $w' \neq w$ in $W_{(Goal,\alpha)}$;
  2. $R_{(Goal,\alpha)}$ is serial;
  3. $R_{(Goal,\alpha)} \cap R_{(Bel,\alpha)} \neq \varnothing$;
- I: $D_{Ag} \times D_I \times W \rightarrow \mathcal{P}W \times \mathcal{P}(W \times W)$ is such that:
  1. for $\alpha = (a, i, w)$, $I\alpha = (W_{(Int,\alpha)}, R_{(Int,\alpha)})$ is centred at w, that is $w \in W_{(Int,\alpha)}$, and $(w, w') \in R_{(Int,\alpha)}$ for any $w' \neq w$ in $W_{(Int,\alpha)}$;
  2. $R_{(Int,\alpha)}$ is serial;
  3. $R_{(Int,\alpha)} \cap R_{(Bel,\alpha)} \neq \varnothing$;
- R_B: $D_{Roles} \times D_I \times W \rightarrow \mathcal{P}W \times \mathcal{P}(W \times W)$ is such that for $\alpha = (r, i, w)$, $R_B\alpha = (W_{(RoleBel,r)}, R_{(RoleBel,r)})$ is centred at w, that is $w \in W_{(RoleBel,r)}$, and $(w, w') \in R_{(RoleBel,r)}$ for any $w' \neq w$ in $W_{(RoleBel,r)}$.
- R_G: $D_{Roles} \times D_I \times W \rightarrow \mathcal{P}W \times \mathcal{P}(W \times W)$ is such that for $\alpha = (r, i, w)$, $R_G\alpha = (W_{(RoleGoal,r)}, R_{(RoleGoal,r)})$ is centred at w, that is $w \in W_{(RoleGoal,r)}$, and $(w, w') \in R_{(RoleGoal,r)}$ for any $w' \neq w$ in $W_{(RoleGoal,r)}$.
- R_I: $D_{Roles} \times D_I \times W \rightarrow \mathcal{P}W \times \mathcal{P}(W \times W)$ is such that for $\alpha = (r, i, w)$, $R_I\alpha = (W_{(RoleInt,r)}, R_{(RoleInt,r)})$ is centred at w, that is $w \in W_{(RoleInt,r)}$, and $(w, w') \in R_{(RoleInt,r)}$ for any $w' \neq w$ in $W_{(RoleInt,r)}$.
- v: $Const \rightarrow D$ is an interpretation function for constants; and finally
- Φ: $Pred \times W \rightarrow \bigcup_{n \in \mathbb{N}} D^n$ is a function which gives the extension of each predicate symbol in each world, such that $\forall P \in Pred$, $\forall n \in \mathbb{N}$, $\forall w \in W$, if $arity(P) = n$, then $\Phi(P, w) \subseteq D^n$, i.e., Φ preserves arity.

Definition 8. Let g be a variable assignment, and let v be defined as above. Then the term valuation function V_g is defined as follows:

- $V_g(\tau) = v(\tau)$, for $\tau \in Const$;
- $V_g(\tau) = g(\tau)$, for $\tau \in Var$;
- $V_g(\tau) = [Vg(u), Vg(u')]$, for $\tau = [u, u'] \in termI$.

The semantics of the language are defined via the satisfaction relation, which holds between *interpretation structures* and formulae of the language. An interpretation structure is a triple $\langle M, w, g \rangle$, where M is a model, w is a world, and g is a variable assignment. The rules defining the satisfaction relation are given in Definition 10.

Definition 9. A formula ϕ is true at a world w in M (written $M, w \vDash \phi$) if ϕ is satisfied by all assignments g at w. If a formula ϕ is *valid* (satisfied by all interpretation structures), we write $\vDash \phi$, as usual.

Definition 10. A variable assignment g satisfies a formula ϕ at a world w in a model $M = \langle W, w_0, D_{Ag}, D_T, r_{DT}, D_I, D_{Roles}, D_{RelTypes}, D_O, In, B, D, G, I, R_B, R_G, R_I, v, \Phi\rangle$, written $M, w, g \models \phi$, as follows.

- $M, w, g \models \text{true}$
- $M, w, g \models t < t'$ iff $(V_g(t), V_g(t')) \in r_{DT}$
- $M, w, g \models P(u_1, \ldots, u_n)(i)$ iff $(V_g(u_1), \ldots, V_g(u_n)) \in (\Phi(P, w), V_g(i))$
- $M, w, g \models (u_1 = u_2)(i)$ iff $V_g(u_1) = V_g(u_2)$
- $M, w, g \models In(a, r)(i)$ iff $(V_g(a), V_g(r), V_g(i), w) \in In$
- $M, w, g \models Bel(a, \phi)(i)$ iff $\alpha = (V_g(a), V_g(i), w)$, $B\alpha = (W_{(Bel,a)}, R_{(Bel,a)})$ and $M, w', g \models \phi$ for all $(w, w') \in R_{(Bel,a)}$
- $M, w, g \models Des(a, \phi)(i)$ iff $\alpha = (V_g(a), V_g(i), w)$, $D\alpha = (W_{(Des,a)}, R_{(Des,a)})$ and $M, w', g \models \phi$ for all $(w, w') \in R_{(Des,a)}$
- $M, w, g \models Goal(a, \phi)(i)$ iff $\alpha = (V_g(a), V_g(i), w)$, $G\alpha = (W_{(Goal,a)}, R_{(Goal,a)})$ and $M, w', g \models \phi$ for all $(w, w') \in R_{(Goal,a)}$
- $M, w, g \models Int(a, \phi)(i)$ iff $\alpha = (V_g(a), V_g(i), w)$, $I\alpha = (W_{(Int,a)}, R_{(Int,a)})$ and $M, w', g \models \phi$ for all $(w, w') \in R_{(Int,a)}$
- $M, w, g \models RoleBel(r, \phi)(i)$ iff $\alpha = (V_g(r), V_g(i), w)$, $R_B\alpha = (W_{(RoleBel,r)}, R_{(RoleBel,r)})$ and $M, w', g \models \phi$ for all $(w, w') \in R_{(RoleBel,r)}$
- $M, w, g \models RoleGoal(r, \phi)(i)$ iff $\alpha = (V_g(r), V_g(i), w)$, $R_G\alpha = (W_{(RoleGoal,r)}, R_{(RoleGoal,r)})$ and $M, w', g \models \phi$ for all $(w, w') \in R_{(RoleGoal,r)}$
- $M, w, g \models RoleInt(r, \phi)(i)$ iff $\alpha = (V_g(r), V_g(i), w)$, $R_I\alpha = (W_{(RoleInt,r)}, R_{(RoleInt,r)})$ and $M, w', g \models \phi$ for all $(w, w') \in R_{(RoleInt,r)}$
- $M, w, g \models \neg\psi$ iff $M, w, g \not\models \psi$
- $M, w, g \models \psi \vee \chi$ iff $M, w, g \models \psi$ or $M, w, g \models \chi$
- $M, w, g \models \exists x\psi$ iff $M, w, g' \models \psi$ for some g' differing from g at most on x