Formalizing Collaborative Decision-making and Practical Reasoning in Multi-agent Systems

PIETRO PANZARASA, *Dept. of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK.*
Email: pp@ecs.soton.ac.uk

NICHOLAS R. JENNINGS, *Dept. of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK.*
Email: nrj@ecs.soton.ac.uk

TIMOTHY J. NORMAN, *Dept. of Computing Science, University of Aberdeen, Aberdeen AB24 3UE, UK.*
Email: tnorman@csd.abdn.ac.uk

Abstract

In this paper, we present an abstract formal model of decision-making in a social setting that covers all aspects of the process, from recognition of a potential for cooperation through to joint decision. In a multi-agent environment, where self-motivated autonomous agents try to pursue their own goals, a joint decision cannot be taken for granted. In order to decide effectively, agents need the ability to (a) represent and maintain a model of their own mental attitudes, (b) reason about other agents' mental attitudes, and (c) influence other agents' mental states. *Social mental shaping* is advocated as a general mechanism for attempting to have an impact on agents' mental states in order to increase their cooperativeness towards a joint decision. Our approach is to specify a novel, high-level architecture for collaborative decision-making in which the mentalistic notions of belief, desire, goal, intention, preference and commitment play a central role in guiding the individual agent's and the group's decision-making behaviour. We identify preconditions that must be fulfilled before collaborative decision-making can commence and prescribe how cooperating agents should behave, in terms of their own decision-making apparatus and their interactions with others, when the decision-making process is progressing satisfactorily. The model is formalized through a new, many-sorted, multi-modal logic.

Keywords: Multi-agent systems, BDI logic, joint mental attitudes, inter-agent social behaviour.

1 Introduction

Since Aristotle, it has been common to distinguish between two fundamental modes of reasoning: theoretical and practical. In its barest sense, theoretical reasoning is mainly used to describe a form of reasoning whose goal is knowledge and whose typical outcome is, at least, justified belief [45]. Practical reasoning, on the other hand, is usually intended to be directed towards conduct and, particularly, is expected to explain what it is for an agent to act for a reason [2]. However, both modes of reasoning can be described as an inferential process moving from a problem to an appropriate response. On the one hand, theoretical reasoning can be conceived as reasoning undertaken to answer a theoretical problem, i.e. a problem about what is the case. On the other, practical reasoning can be characterized as reasoning undertaken to find an answer to a practical problem, i.e. a problem about what is to be done.

In either case, the full statement of the problem, whether theoretical or practical, will involve giving all the relevant information and this provides the premisses from which a conclusion can be inferred that represents an answer to the problem. In this view, an agent's drawing that conclusion is an appropriate response to its asking a question, whether theoretical or practical.

In this paper, our focus will be on practical reasoning. Following the Aristotelian tradition, practical reasoning is here conceived of as a belief-based, intention-motivated, and action-guiding species of reasoning.[1] In essence, it coordinates intentions towards a state of affairs with beliefs about possible alternative courses of action that are means to achieve that state and with a practical judgement that recommends a prudent course of action, and it concludes in an intention to act according to the practical judgement. So conceived, practical reasoning is the vehicle of decision-making. It works out the requirements and steps of a decision by forming the sequences of possible paths of actions appropriate to a situation. Here, a decision is a composite concept specifying what practical judgement the agent has brought about through practical reasoning and how the agent is committed to acting in compliance with that judgement.

Most traditional work within the area of practical reasoning and decision-making has concentrated on solipsistic agents [25, 69]. However, with the advent of the Internet and other forms of network computing and applications that use cooperative agents working towards a common goal, multiple agents are increasingly becoming the norm [50]. In such cases, we need to re-consider and extend our notions of practical reasoning and decision-making so that they deal with the inherently social aspects of these classes of multi-agent systems. Given this, we are specifically concerned with issues that arise in the formal specification of practical reasoning and decision-making within a social setting, particularly within groups of agents that are primarily designed to exhibit specified intelligent behaviour as a collective. Collaborative decision-making (CDM) is perhaps the paradigm example of activity in multi-agent systems [11, 50]. It refers to a group of logically decentralized agents that cooperate to achieve objectives that are typically beyond the capabilities of any individual agent. In short, CDM has generally been viewed and modelled as a kind of distributed reasoning and search, whereby a collection of agents collaboratively go through the search space of a problem in order to find a solution [11, 33, 38].

Recently, a number of theoretical models have been proposed for investigating decision-making in a social setting [33, 48, 71, 82, 86]. However, none of these approaches cover the full breadth of social and cognitive activities that are typically involved in a CDM process. In some of these approaches, agents are seen as endowed with identifiable decision-making capabilities and are grouped together to form communities which cooperate to achieve both individual goals and the goals of the system as a whole [62, 86]. In these approaches, co-operative behaviour stems from predefined interactions between tightly coupled agents that cannot operate outside of the specific cooperation protocols specified in advance by the system designer. Therefore each agent has little knowledge of the system's overall objective or of general strategies for communication and coordination. In other approaches, the main focus is on coordination strategies between multiple decision-makers [15, 32, 34, 37, 78, 79]. Consequently, the mental apparatus and inferential mechanisms of agents are obscured under the mere assumption that agents have their own decision-making expertise which frequently has to be coordinated when the goals undertaken by individual agents are related.

[1] This definition encapsulates the notion of practical reasoning analysed in [3] and [24].

However, in the majority of multi-agent systems, community members are relatively autonomous in deciding what actions to perform *and* can reason about the process of coordination. Given this, the purpose of this paper is to put forward an alternative view of CDM. In this view, CDM is treated as a *multi-agent socio-cognitive process* in which a number of social and cognitive aspects are dynamically intertwined. That is, we propose a two-pronged approach to CDM where the mental modelling is intimately combined with an account of sociality and interaction capabilities. By *mental modelling* we mean an explicit account of the agents' mental attitudes, such as beliefs, desires, goals, intentions, and preferences, and the operations performed upon them. *Sociality*, on the other hand, refers to the agents' capabilities to interact with one another and to the mechanisms that they can use to coordinate their behaviour effectively within a common social environment.

It is our aim to provide an account of CDM that allows a step-by-step *dual-faceted* reconstruction of the entire process that leads intelligent autonomous agents to make a decision in a collaborative manner. To this end, we formalize a model of CDM that offers an insight into: (a) the agents' mental states and processes, and (b) a range of social behaviours that lead them to solicit, and take part in a decision-making process. Particularly, we will address the following questions:

(a) Why do agents engage in CDM in its most general form, and what are the reasoning mechanisms regulating it?
(b) Are the agents' mental states modified when they jointly reason about what to do, and if so, how does this happen?
(c) What is the relationship between the mental states of interacting agents?
(d) What are the coordination mechanisms that enable these relationships to be effected?

A key role in combining mental modelling with sociality will be played by the process of *social mental shaping* (Section 3.9) [66]. By this we mean the process by which the mere social nature of agents may impact upon their mental states and motivate their behaviour. This process can involve either roles (Sections 3.7 and 3.9.1) or social relationships (Sections 3.8 and 3.9.2) or simply other agents outside of any social relationship (Section 3.9.2). In our account, the social environment[2] in which an agent is located can be seen as a source of mental attitudes that the agent can adopt to complement or merely to change its individual mental state. To the extent that agents have social capabilities and can reason about other agents, their mental apparatus can be changed and/or complemented by internalizing *socially motivated* mental attitudes. Therefore, the complex interplay between the agent and its social environment turns out to be a process in which roles, agents and social relationships may complete and augment bare individual mental attitudes.

Another key characteristic of our approach is the distinction between CDM and social practical reasoning. On the one hand, CDM is meant to cover the whole process by which the agents recognize a potential for collaboration, organize themselves into a group, reason together about what to do and, finally, commit themselves to a given course of action. Meanwhile, social practical reasoning is clearly just one fundamental stage of CDM. It strictly refers to the mental apparatus and inferential processes through which agents reason about the appropriate means to achieve a given state of affairs.

[2]Here we use the term *social environment* to broadly refer to the pattern of social relations in which an agent is *actually and potentially* involved. Thus, it refers both to the set of social relationships (Section 3.9) that an agent has already established and to those that may be established as circumstances evolve.

Building upon [88] (see Section 9.3 for more details), our model of CDM consists of four stages:

(a) *the practical starting-point* — in which an agent is confronted with a problem concerning what is to be done and identifies a potential for cooperation over the resolution of that problem;
(b) *group generation* — in which a group is formed with decision-making purposes;
(c) *social practical reasoning* — in which each agent involved in the newly formed group attempts, via a set of social inferential processes, to find a solution to the problem set at the beginning;
(d) *negotiation* — in which the members of the group interact in an effort to make an agreement about a course of action to be performed by the group.

The purpose of this paper is, therefore, to provide a high-level agent architecture for CDM in which social practical reasoning plays a central role. In particular, our model identifies preconditions that must be fulfilled before CDM can commence and prescribes how cooperating agents should behave, in terms of their own decision-making apparatus and their interactions with others, when CDM is progressing satisfactorily. This work extends the state of the art in two ways. First, from a theoretical viewpoint, it identifies the fundamental structures and computational processes that underpin decision-making in a multi-agent system environment. Second, it can be viewed as an abstract specification for agent designers. That is, by interfacing mechanisms of individual and social behaviour, it represents a step towards the goal of helping practitioners to implement grounded real-world applications of models of decision-making for multi-agent systems. This link between the theoretical specification and its corresponding implementation is important if agents are to exhibit predictable and reliable behaviour both at the individual agent level and at the system level.

The remainder of this article is structured as follows. Section 2 outlines a number of desiderata that should be met by any principled account of CDM. Section 3 introduces a new quantified multi-modal logic that has been devised to formally express our model. This logic is used to reason about and represent the mental attitudes both of individual agents and of groups. Section 4 provides an overview of our four-stage model of CDM. Finally, Sections 5 through 8 analyse, and subsequently formalize, each stage. Section 9 situates our work within the related literature. Section 10 presents concluding remarks.

2 Desiderata for a theory of CDM

In a CDM process, decisions are the product of a variety of social actions and interactions and, as such, CDM is considerably different from the standard individual decision-making process. In this section, we will identify the desiderata that are required for an adequate account of decision-making in a social setting. These criteria will subsequently be used to evaluate our model (Section 10).

1. *Both individualistic and higher-order units of analysis and constructs are required*
   As social embeddedness and institutional theorists argue, individual agents both create and are products of social interaction [5, 16, 17, 55, 63]. That is, the individual agent is context-dependent and habits of mind and behaviour develop in a social environment. Individual mentalistic, motivational and behaviouristic concepts (i.e. individual actions, individual doxastic, motivational and deontic attitudes) are useful in defining some of the

processes involved in decision-making, goal setting, search for information, evaluation of information, search for alternatives, evaluation of alternatives, selection of the preferred alternative, and the implementation of it. However, individual concepts only define an individual's *local* decision-making behaviour and, as such, are an insufficient base on which to build a principled representation of CDM. The presence of other agents causes the dynamics of making the decision to differ from those of individual decisions. Decisions are made through processes of negotiation [49, 81], compromise, and bargaining and coalition formation [5]. Collaborative decisions are not necessarily those that any single member of the group would have made, or even the average. Therefore, to capture the dynamics of making a decision within a social setting we need representations that address higher-order units of analysis (i.e. groups, multi-agent actions, joint doxastic, motivational and deontic attitudes).

2. *Agents are autonomous, reactive and pro-active*
   First, agents will take part in CDM only if they choose to do so [36, 87]. That is, they are not required to be benevolent with an a priori disposition to cooperate whenever asked to. Second, agents should respond to any perceived change that occurs in their social environment and that affects their mental states and behaviour [18, 87]. Third, agents may take the initiative where appropriate and may exhibit opportunistic behaviour towards other agents sharing a social environment [87].
3. *Agents are self- and other-interested*
   First, individual agents seek maximal and efficient satisfaction of their own preferences [10, 58]. However, self-satisfaction is an insufficiently rich base on which to represent CDM. Agents are frequently torn between increasing self-satisfaction and their commitment to others [18]. Social commitment may be as valid a basis for CDM as self-interest. Agents are variably motivated by rationality, emotions, socialization, habit and social relationships, or some mixture of any of these. Commitment to others captures the three features that, according to Bratman [14], characterize truly shared cooperative activity. First, *mutual responsiveness*: each group member has a disposition to be responsive to the mental states and behaviour of the other agents involved in CDM. That is, local individual decision-making behaviour is influenced by expectations about the others, based on beliefs about the others' mental attitudes. Second, *commitment to the joint activity*: the agents participating in CDM are committed to acting together in order to make a joint decision. That is, each agent has an 'intention in favour of the joint activity', and will not drop such an intention at least until certain escape conditions become true [25, 27]. Third, *commitment to mutual support*: each agent is committed to supporting the others in their attempts to play their role in CDM. That is, although not benevolent (point 2), agents may decide to initiate social processes and, once committed to making a decision together, they will support one another during the performance of their joint activity. An adequate theory of CDM must describe all these features of the agents' commitments to each other, identify the key doxastic and motivational attitudes on which they are built, and tease out the fundamental forms of behaviour they bring about.
4. *Communication is essential*
   The need for communication pervades joint activity [63]; hence, CDM as well [88]. Joint intentions and goals cannot be adequately accomplished, and collective practical problems cannot be solved without adequate communication. The purposes of communication are to provide information and instructions, to influence and to integrate activities. Communication is interrelated with CDM effectiveness. Poor communication is frequently

cited as the basis of conflict within groups, and it is the cause of many other problems within social settings [63]. Good communication is also the basis for establishing effective relations and motivating agents to get involved in cooperative activity. Communication involves the use of networks, policies, and structures to facilitate coordination activities that help to reach a joint decision. An adequate theory of CDM should describe where and when communication is essential, although it should not prescribe the means through which it takes place.

5. *Cooperation can fail*
   Since agents are not assumed to be benevolent, there is no guarantee that cooperation is always successful [88]. Even when CDM is initially established, it can fail to progress satisfactorily for many different reasons. For example, some of the agents that are jointly committed to making a decision together may come later to adopt a goal or an intention that turns out to be incompatible with the ongoing cooperative activity. Alternatively, some unpredictable event may make the completion of CDM infeasible (e.g. the practical problem the group is jointly committed to solving may turn out to be unsolvable or beyond the control and capabilities of the agents involved). Therefore an adequate theory of CDM must not assume successful cooperation. Rather, failure must be accounted for. Furthermore, the key stages of the decisional process in which failure may occur must be identified, and the behaviour of the agents must be characterized in all those circumstances in which CDM cannot be successfully completed.
6. *Conflict is pervasive in CDM*
   Some degree of conflict can be expected in every CDM process [5, 63]. Although some organizational conflict may be functional (e.g. it may increase spontaneity in communication or enhance the stimulation of creativity and therefore productivity), it is likely that mismanaged conflict is dysfunctional to CDM. Among structural factors that can lead to conflict are interdependent tasks, differences in agents' goals, intentions and preferences, and differing expectations about others' mental attitudes and behaviour. The different types of conflict that may be involved in CDM include: inter-agent conflict (a friction between two or more agents) and role conflict (occurring when two or more roles have conflicting requirements). Such conflict among the agents involved in CDM decreases cooperation and may lead to actions that debar others from effectively performing their parts. An adequate theory of CDM must describe all those situations in which conflict may arise, and identify the appropriate resolution techniques that agents may adopt.
7. *CDM is a multi-stage process*
   Much work on group and organizational decision-making identifies the stages through which a collective is expected to progress in making a decision [61, 63]. Broadly speaking, these stages include: (a) task announcement and orientation: this stage involves recognizing a potential for social action and acquainting group members with the nature of the problem that requires a decision; (b) evaluation: this stage involves the search for alternative courses of action by group members and the judgement of these alternatives; (c) negotiation: group members typically possess somewhat different values and interests, and these differences must be resolved for the group to arrive at an agreement. An adequate theory of CDM must recognize that CDM involves sequential phases, and must identify the key behavioural processes that characterize these phases.

8. *CDM may involve the following of rules and routines that adapt to experience, rather than anticipatory choice*
   One of the oldest behavioural speculations concerning decision-making within organizations is that time and attention are scarce resources [60, 61]. Most work on organizational decision-making suggests that much organizational behaviour involves rule-following and history-dependent processes [60, 61, 63]. That is, organizations have standard operating procedures and relatively stable routines learned as appropriate in a particular situation. An adequate theory of CDM must account for decisions that are driven by rules reflecting history and encoding past experience. These behavioural rules and routines should define how to behave locally and towards other agents both when joint action is progressing as planned and when it runs into difficulty.

3 The formal framework

This section gives an overview of the formal framework in which our model of CDM will be expressed. A complete formal definition is given in the Appendix. The formalism used is a many-sorted first-order multi-modal language L which both draws upon and extends the work described in [8, 9, 25, 69, 88].[3] L is a many-sorted logic for reasoning about agents, actions, roles, social relationships, and mental attitudes, with explicit reference to time points and intervals.

Informally, the = operator is 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 existential quantifier are assumed to be introduced as abbreviations, in the obvious way. We also use the punctuation symbols ‘)’, ‘(’, ‘]’, ‘[’, and comma ‘,’.

3.1 Time

In L we have terms that denote *time points*, and we use $t_i, t_j, \ldots$ and so on as variables ranging over time points. Every occurrence of a formula φ is stamped with a time t_i, written $\varphi(t_i)$, 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 $i_i, i_j, \ldots$ 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). The usual connectives of linear temporal logic can be defined in the following way [8, 23]: $U(\varphi, \psi)(t_i)$ means ψ is satisfied until φ becomes satisfied; $\Diamond\varphi(t_i)$ means φ is eventually satisfied; $\Box\varphi(t_i)$ means φ is always satisfied.

[3] Our use of time and preferences is consistent with the work of Bell [8] and Bell and Huang [9]. Our formalization of individual mental attitudes draws upon [25] and [69]. Finally, our set-theoretic mechanism for relating agents and groups is similar to that of Wooldridge and Jennings [88]. However, our language extends the aforementioned formal frameworks in that it contains terms for reasoning about roles and relationships, and it explicitly addresses the formalization of joint mental attitudes (see Section 9.1 for a more detailed discussion of this matter).

$\forall t_i U(\varphi, \psi)(t_i) \equiv \exists t_j (t_i < t_j)$ s.t. $(\varphi(t_j) \wedge \forall t_k (t_i \leq t_k < t_j) \psi(t_k))$
$\forall t_i \Diamond \varphi(t_i) \equiv \exists t_j (t_i \leq t_j)$ s.t. $\varphi(t_j)$
$\forall t_i \Box \varphi(t_i) \equiv \forall t_j (t_i \leq t_j) \varphi(t_j)$[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, $(\varphi \wedge \psi)(i_i)$ abbreviates $(\varphi)(i_i) \wedge (\psi)(i_i)$.
- In complex sentences the same temporal terms are often repeated. In what follows we will adopt the convention that a missing temporal term is the same as the closest temporal term to its right. For example, *Goal* $(a_i,$ *Does* $(a_j, e_i))(t_i)$ states that at time t_i agent a_i has the goal that at time t_i agent a_j performs action e_i.

3.2 Agents and groups of agents

We have terms that denote *agents*, and we use $a_i, a_j, \ldots$ 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 characterize agents [36, 50]. 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 and take the initiative where appropriate. Fourth, agents have social ability: they can initiate social relationships with each other and will be mutually supportive during the execution of their joint actions.

In addition, we have terms that denote groups of agents, and we use $gr_i, gr_j, \ldots$ and so on as variables ranging over groups of agents. A group gr_i of agents is simply a non-empty subset of the set of agents. Agents and groups may easily be related to one another via simple set theory. With the $\in$ operator, we relate agents to groups of agents: $a_i \in gr_i$ means that the agent denoted by a_i is a member of the group denoted by gr_i. The operators $\subseteq$ and $\subset$ relate groups together, and have the obvious set-theoretic interpretation. We have:

$\forall gr_i, \forall gr_j, \forall i_i (gr_i \subseteq gr_j)(i_i) \equiv \forall a_i (a_i \in gr_i)(i_i) \supset (a_i \in gr_j)(i).$
$\forall gr_i, \forall gr_j, \forall i_i (gr_i \subset gr_j)(i_i) \equiv (gr_i \subseteq gr_j)(i_i) \wedge \neg (gr_i = gr_j)(i_i).$

Singleton $(gr_i, a_i)(i_i)$ means that, at i_i, gr_i is a singleton group with a_i as the only member:

$\forall gr_i, \forall a_i, \forall i_i$ *Singleton* $(gr_i, a_i)(i_i) \equiv \forall a_j (a_j \in gr_i)(i_i) \supset (a_j = a_i)(i_i).$

3.3 Actions and plans

In addition to terms denoting agents and groups, we have terms that denote *sequences of actions*, and we use $e_i, e_j, \ldots$ and so on as terms denoting sequences of actions. We distinguish between an action sequence type (an abstraction) and its occurrence in the world. The operator *Occurs* $(e_i)(i_i)$ states that action sequence e_i (type) happens at interval i_i. Complex

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

actions are defined in the usual way: Occurs $(e_i; e_j)(i_i)$ means that at i_i action sequence e_i is immediately followed by action sequence e_j; *Occurs* $(e_i|e_j)(i_i)$ means either e_i or e_j occurs at i_i; *Occurs* $(\varphi?)(i_i)$ is a test action which occurs if φ is true at i_i; *Occurs* $(e_i||e_j)(i_i)$ means that both action sequence e_i and action sequence e_j occur at i_i.

Actions may be distinguished depending on whether they can be performed by an individual agent (single-agent actions) or by a group of agents (multi-agent actions). To simplify the specification, we assume that an action sequence is either single-agent or multi-agent, but not both. A sentence of the form $Agts(gr_i, e_i)(i_i)$ states that at interval i_i the group denoted by gr_i are the agents required to perform the actions in the multi-agent action sequence denoted by e_i. $Agt(a_i, e_i)(i_i)$ means that a_i is the only agent of e_i at interval i_i. We have the following definition:

$$\forall a_i, \forall e_i, \forall i_i Agt(a_i, e_i)(i_i) \equiv \forall gr_i Agts(gr_i, e_i)(i_i) \supset \textit{Singleton}\ (gr_i, a_i)(i_i).$$

We formalize the performance of action sequences by agents and groups of agents by introducing the following operators:

$$\forall gr_i, \forall e_i, \forall i_i\ \textit{Do}\ (gr_i, e_i)(i_i) \equiv \textit{Occurs}\ (e_i)(i_i) \wedge Agts(gr_i, e_i)(i_i).$$
$$\forall a_i, \forall e_i, \forall i_i\ \textit{Does}\ (a_i, e_i)(i_i) \equiv \forall gr_i\ \textit{Do}\ (gr_i, e_i)(i_i) \supset \textit{Singleton}\ (gr_i, a_i)(i_i).$$

Along another dimension, actions may be further distinguished between those that are not state-directed and those that are [19, 29]. In the former case, actions are not motivated, monitored and guided by a mental representation of a state of the world (for example in some animals and in functional artefacts). In the latter case, actions require intelligent (cognitive) agents, that is, agents whose behaviour is regulated by a set of mental attitudes.

In what follows we are interested only in state-directed actions. These actions cannot be characterized simply as world-state transitions. Since they are regulated by mental attitudes, it is necessary to specify not only their *de facto* results, but also their *expected* and *intended results*. To this end, we introduce a derived operator *plan* that allows us to represent and reason about actions that individual agents or groups of agents perform in order to achieve a state of the world. Specifically, the operator $\textit{plan}(gr_i, e_i, \varphi(t_j))(t_i)$ expresses the fact that, at time t_i, action sequence e_i represents, for group gr_i, a plan to bring about state φ at time $t_j (t_i < t_j)$.[5] Formally, we have the following definition:

$$\forall gr_i, \forall e_i, \forall t_i, t_j (t_i < t_j)\ \textit{plan}(gr_i, e_i, \varphi(t_j))(t_i) \equiv$$
$$\exists t_h, t_k (t_i \leq t_h \leq t_k < t_j) \text{ s.t.}$$
$$Do(gr_i, e_i)(t_h, t_k) \wedge (Occurs(e_i)(t_h, t_k) \supset Occurs(\varphi?)(t_j))$$

Informally, we say that at time t_i action sequence e_i is a plan for group gr_i to achieve φ at $t_j (t_i < t_j)$ iff: (a) e_i will occur sometime before t_j; (b) gr_i will be the group required to perform e_i; and (c) if e_i occurs, then φ will be satisfied afterwards at t_j. As happens

[5] A more sophisticated definition of plans could have been adopted. For example, it might be useful to distinguish between the body and the preconditions of a plan. The body points to the method of carrying out the plan, whereas the preconditions refer to the circumstances under which the plan can be executed. Moreover, we could have made explicit representations of partial plans as well as hierarchical non-linear plans [42, 51]. These notions illustrate how a group of agents can decompose a higher-level state of the world into lower-level sub-states, which again can be decomposed into further lower-level states, until one finally reaches primitive plan types. However, for convenience, we will not attempt to represent such refinements in our model and we leave them to future work.

with the broader category of action sequences, we take state-directed actions to be either single-agent or multi-agent, but not both. A single-agent state-directed action is a single-agent action sequence e_i that at time t_i represents a plan for agent a_i to bring about state φ at time $t_j (t_i < t_j)$. Formally, we have the following definition:

$\forall a_i, \forall e_i, \forall t_i, t_j (t_i < t_j)$ ***plan***$(a_i, e_i, \varphi(t_j))(t_i) \equiv \forall gr_i$ ***plan***$(gr_i, e_i, \varphi(t_j))(t_i) \supset$
Singleton $(gr_i, a_i)(t_i)$.

The above definitions of single-agent and multi-agent state-directed actions capture the notion of actions that agents or groups eventually perform to satisfy certain states of the world. We also want to be able to describe the past execution of state-directed actions. To this end, we introduce the following operators:

$\forall gr_i, \forall e_i, \forall t_i \langle plan(gr_i, e_i, \varphi)\rangle(t_i) \equiv$
$\exists t_j, t_k (t_j \leq t_k < t_i)$ s.t.
Do$(gr_i, e_i)(t_j, t_k) \wedge$ (***Occurs***$(e_i)(t_j, t_k) \supset$ ***Occurs***$(\varphi?)(t_i))$

$\forall a_i, \forall e_i, \forall t_i,$
$\langle$***plan***$(a_i, e_i, \varphi)\rangle(t_i) \equiv \forall gr_i \langle$***plan***$(gr_i, e_i, \varphi)(t_i)\rangle \supset$ ***Singleton*** $(gr_i, a_i)(t_i)$.

Informally, we say that, at time t_i, the state of the world φ has been brought about as a consequence of the performance of action sequence e_i by group gr_i iff: (a) e_i occurred sometime in the past; (b) gr_i was the group required to perform e_i; and (c) φ was satisfied afterwards at t_i as a consequence of the occurrence of e_i.

3.4 Doxastic and motivational individual mental attitudes

Our analysis is based on a fairly standard BDI (belief, desire, intention) framework as found, for example, in [69] and [77]. Agents' mental states are here seen as sets of interrelated mental attitudes, among which there are doxastic attitudes (beliefs), motivational attitudes (desires, goals, intentions, preferences), and deontic attitudes (commitments). In this section, we will introduce individual doxastic and motivational attitudes and develop the technical apparatus for dealing with their semantics. Deontic attitudes will be dealt with in Section 3.6.

3.4.1 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 $Bel(a_i, \varphi)(t_i)$, 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 Hintikka's traditional possible-worlds semantics [46, 47]. 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.

We now express the semantic clause for Bel sentences. Formally, for model M, world w in M and variable assignment g, we have:

$M, w, g \models Bel(a_i, \varphi)(t_i)$ iff $M, w', g \models \varphi$ 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[6] [23]:

$\models Bel(a_i, \varphi)(t_i) \wedge Bel(a_i, (\varphi \supset \psi))(t_i) \supset Bel(a_i, \psi)(t_i)$
(closure under logical consequence: K_B)

$\models Bel(a_i, \varphi)(t_i) \supset \neg Bel(a_i, \neg\varphi)(t_i)$
(consistency axiom: D_B)

$\models Bel(a_i, \varphi)(t_i) \supset Bel(a_i, Bel(a_i, \varphi))(t_i)$
(introspection axiom: 4_B)

$\models \neg Bel(a_i, \varphi)(t_i) \supset Bel(a_i, \neg Bel(a_i, \varphi))(t_i)$
(negative introspection axiom: 5_B)

$\models \varphi(t_i) \supset Bel(a_i, \varphi)(t_i)$
(inference rule of necessitation)

The following conditions are imposed on the Belief-accessibility relation [25]:

Condition 1. Each $R_{(Bel,a_i,t_i,w)}$ is serial.
Condition 2. Each $R_{(Bel,a_i,t_i,w)}$ is transitive.
Condition 3. Each $R_{(Bel,a_i,t_i,w)}$ is euclidean.

A Belief-accessibility relation that satisfies conditions 1 to 3 validates the $D_B 4_B 5_B$ axioms. Furthermore, since axiom K_B is valid, it will be a theorem of any complete axiomatization of normal modal logic. Finally, the necessitation rule is a rule of inference in any axiomatization of normal modal logic [23].

3.4.2 Desires

An agent's desires are here conceived of as the set of states of the world that the agent wishes to bring about [8, 52]. To express an agent's desires, we introduce the modal operator $Des(a_i, \varphi)(t_i)$, 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, $Des(a_i, \varphi(t_j))(t_i)$ 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 set of an agent's 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 unrealizable desires; that is, desires that conflict with what it believes possible.

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

$$\models Des(a_i, \varphi)(t_i) \wedge Des(a_i, (\varphi \supset \psi))(t_i) \supset Des(a_i, \psi)(t_i) \qquad (K_D).$$

Furthermore, we have a necessitation property [23]:

$$\models \varphi(t_i) \supset \models Des(a_i, \varphi)(t_i).$$

[6] 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.

Again, axiom K_D and the necessitation rule are, respectively, a theorem and a rule of inference in any axiomatization of normal modal logic [23].

3.4.3 Goals

Goals can be defined as a set of consistent and realizable states of the world that an agent might be expected to bring about. They represent an agent's agenda that might motivate its current and future behaviour [9, 19]. Agents may choose their goals among their desires. However, as goals must be consistent and realizable whereas desires may be inconsistent and unrealizable (see Section 3.4.2), only the subsets of consistent and realizable desires can be moved up to goal-status, and also the selected subsets of consistent desires must be consistent with each other. Furthermore, an agent may have goals that are not desires; that is, there may well be states of the world that an agent does not wish to bring about, but that it *autonomously* chooses as potential candidates for motivating its behaviour. Typically, these are goals that are instrumental to the achievement of those goals that are also desires. Furthermore, goals may be adopted in response to changes in the physical and social environment. For example, an agent may be influenced to adopt a goal as a consequence of its taking on a role in a group, although it may not have a parallel desire.[7] To express an agent's goals, we introduce the modal operator $Goal(a_i, \varphi)(t_i)$, 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, $Goal(a_i, \varphi(t_j))(t_i)$ means that agent a_i has, at time t_i, the goal that φ holds at t_j, where $t_i \leq t_j$.

From this background, we can start formalizing the axiomatization for goals. Axioms K_G and D_G state that goals are, respectively, closed under implication and consistent:

$$\models Goal(a_i, \varphi)(t_i) \wedge Goal(a_i, (\varphi \supset \psi))(t_i) \supset Goal(a_i, \psi)(t_i) \qquad (K_G)$$
$$\models Goal(a_i, \varphi)(t_i) \supset \neg Goal(a_i, \neg\varphi)(t_i) \qquad (D_G)$$

We now introduce a *weak realism* constraint for goals [70]. Agents do not have goals towards propositions the negations of which are believed. That is, agents' goals do not contradict their beliefs.[8]

Formally, we have the following axiom:

$$\models Goal(a_i, \varphi)(t_i) \supset \neg Bel(a_i, \neg\varphi)(t_i) \qquad (G_1).$$

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

Condition 1. Each $R_{(Goal,a_i,t_i,w)}$ is serial.
Condition 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 $R_{(Bel,a_i,t_i,w)} \cap R_{(Goal,a_i,t_i,w)} \neq \emptyset$).

A Goal-accessibility relation that satisfies conditions 1 and 2 validates axioms D_G, and G_1. Again, axiom K_G is valid, and we have a necessitation property [23]:

$$\models \varphi(t_i) \supset \models Goal(a_i, \varphi)(t_i).$$

[7] This property contrasts with Kraus *et al.*'s framework [52], 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 [25] assume that all the agent's beliefs are also its goals. We do not have such a property. Indeed, our framework is more flexible, in that it allows an agent to believe φ, but not to adopt it as one of its goals at the same time.

[8] Note that a weak realism constraint is not the case with desires. An agent may have a desire to attain a state of the world that is believed to be always false.

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

3.4.4 Intentions

A fundamental characteristic of individual intentions is that they involve a special kind of 'self-commitment' to acting [13, 90]. 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. Fundamentally, we can distinguish between two different forms of intentions [8, 42], *Intention-to* and *Intention-that*, depending on whether the argument is respectively an action or a proposition.[9] That is, intentions can be subdivided into: (a) action-directed intentions (Intention-to), involving the performance of some action; and (b) state-directed intentions (Intention-that), involving the achievement of some state of affairs by performing some action. Like Grosz and Kraus [42], we believe that both types of intention commit an agent not to adopt conflicting intentions, and constrain replanning in case of failure [12]. And like Grosz and Kraus, we take Intention-that to be a fundamental means to correctly mesh collaborating agents' plans. However, in contrast to Grosz and Kraus, we believe that it is an Intention-that, rather than an Intention-to, that represents the basic intention operator that commits an agent to practical reasoning.[10] In our view, an Intention-that can be conceived of as the prior motivator of a line of practical reasoning. When an agent intends that a state of affairs holds, then it is committed to doing something to attain a world in which the intended state holds. Hence, an Intention-that commits an agent to find out, to decide, through practical reasoning, the appropriate means to attain a state of affairs [12].

Therefore, as in this paper we are primarily concerned with formalizing agents' practical reasoning processes, our focus will be exclusively on Intention-that. The modal operator $Int(a_i, \varphi)(t_i)$ is used to represent agent a_i's intention that proposition φ holds at time t_i. Like desires and goals, intentions can only be present-directed or future-directed, i.e. $Int(a_i, \varphi(t_j))(t_i)$ 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 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 achieve this subset at some point in the future. However, there may well be intentions that are not goals. Typically, these are intentions that the agent ought to adopt and does not autonomously choose as potential motivators of its behaviour. For example, an agent, in order to have some of its goals fulfilled, may be obliged to adopt an intention by another agent who has the authority to control the former's behaviour.

We can now formalize our axiomatization for intentions. Intentions are here taken to be closed under implication (K_I) and consistent (D_I):

$$\models Int(a_i, \varphi)(t_i) \wedge Int(a_i, (\varphi \supset \psi))(t_i) \supset Int(a_i, \psi)(ti) \quad (K_I)$$
$$\models Int(a_i, \varphi)(t_i) \supset \neg Int(a_i, \neg\varphi)(t_i) \quad (D_I)$$

As with goals, we introduce a weak realism constraint for intentions. Agents do not intend propositions the negations of which are believed. This ensures that agents' intentions do not

[9] Grosz and Kraus [42] 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 3.4.3).

[10] Grosz and Kraus [42] believe that a *potential* Intention-to stems from an agent's practical reasoning about how to perform some action to which it is already committed, whereas an Intention-to commits an agent to means-end reasoning and eventually to acting.

contradict their beliefs [52, 70]. Formally, we have the following axiom:

$$\models Int(a_i, \varphi)(t_i) \supset \neg Bel(a_i, \neg\varphi)(t_i) \quad (I_1)$$

Again, we have a necessitation property [23]:

$$\models \varphi(t_i) \supset \models Int(a_i, \varphi)(t_i).$$

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

Condition 1. Each $R_{(Int,a_i,t_i,w)}$ is serial.
Condition 2. $\forall w \exists w'$ s.t. $(w, w') \in R_{(Int,a_i,t_i,w)}$ iff $(w, w') \in R_{(Bel,a_i,t_i,w)}$ (or $R_{(Bel,a_i,t_i,w)} \cap R_{(Int,a_i,t_i,w)} \neq \emptyset$).

An Intention-accessibility relation that satisfies conditions 1 and 2 validates axioms K_I, D_I and I_1. Finally, the semantic clause for intentions is analogous to that of beliefs, desires, and goals.

3.4.5 Preferences

For our purposes, a key component of L is the modality *Pref* for expressing preferences between formulae. As we will see in Section 7, an agent's preference plays an active role in social practical reasoning, where an action is to be selected in order for a given intention to be fulfilled. The modal operator $Pref(a_i, \varphi, \psi)(i_i)$ means that agent a_i 'prefers' φ to ψ to hold for the interval i_i. The semantics of this modality are given in terms of a world preference as follows. According to von Wright's conjunction expansion principle [89], to say that an agent prefers φ to ψ is to say that the agent prefers the worlds in which φ holds and ψ does not hold to those in which φ does not hold and ψ holds. Formally, an agent prefers φ to ψ for interval i if it prefers $\varphi \wedge \neg\psi$-worlds to $\psi \wedge \neg\varphi$-worlds for i. However, this gives rise to the well-known paradoxes of 'conjunction and disjunction': if φ is preferred to ψ, then $\varphi \vee \xi$ is preferred to ψ, and φ is preferred to $\psi \vee \xi$. Since ξ can be whatever state of the world, either desirable or undesirable, these properties might sometimes describe unreasonable situations (see [9] for details). To avoid this, we need to impose some constraints on ξ so as to guarantee the *ceteris paribus* nature of preferences. The idea is that we need to be able to compare $\varphi \wedge \neg\psi$-worlds with $\psi \wedge \neg\varphi$-worlds that otherwise differ as little as possible from the actual world. To this end, we adopt a selection function cw that was originally introduced by Bell and Huang [9].

In line with Stalnaker-Lewis' treatments of conditionals [80], Bell and Huang [9] define a function $cw(w, [[\varphi]]_g^M)$ that gives the set of closest worlds to w in which φ is true, where $[[\varphi]]_g^M$ denotes the set of worlds in model M in which φ is satisfied by variable assignment g (i.e. $[[\varphi]]_g^M = \{w \in W : M, w, g \models \varphi\}$). We then define a function P_V that gives the value $p \in \mathbb{R}$ that agent a_i associates with a subset of states of the world, $w = \{w_0, w_1, \ldots, w_n\}$ in a given interval i_i. The agent's preferences over time are thus expressed in the following way: an agent a_i prefers φ to ψ at i_i, $Pref(a_i, \varphi, \psi)(i_i)$, iff at i_i the value a_i associates to the set of closest worlds to the world in which φ holds and ψ does not hold is greater than the value a_i associates to the set of closest worlds to the world in which ψ holds and φ does not hold. Formally, for model M, world w in M and variable assignment g, we have the following semantic condition for preferences:

$$M, w, g \models Pref(a_i, \varphi, \psi)(i_i) \text{ iff}$$
$$P_V(a_i, i_i, cw(w, [[\varphi \wedge, \neg\psi]]_g^M)) > P_V(a_i, i_i, cw(w, [[\psi \wedge \neg\varphi]]_g^M))$$

Given the semantics above, we have the following axioms [9]:

$\models Pref(a_i, \varphi, \psi)(i_i)$ iff $Pref(a_i, (\varphi \wedge \neg\psi), (\psi \wedge \neg\varphi))(i_i)$ (CE)
$\models Pref(a_i, \varphi, \psi)(i_i)$ iff $\neg Pref(a_i, \psi, \varphi)(i_i)$ (ASYM)
$\models Pref(a_i, \varphi, \varphi)(i_i)$ (IR)
$\models Pref(a_i, \varphi, \psi)(i_i) \wedge Pref(a_i, \psi, \chi)(i_i) \supset Pref(a_i, \varphi, \chi)(i_i)$ (TRANS)
$\models Pref(a_i, \varphi, \chi)(i_i) \wedge Pref(a_i, \psi, \chi)(i_i) \supset Pref(a_i, \varphi \vee \psi, \chi)(i_i)$ (DIS*)
$\models Pref(a_i, \varphi, \chi)(i_i) \wedge Pref(a_i, \varphi, \psi)(i_i) \supset Pref(a_i, \varphi, \chi \vee \psi)(i_i)$ (DIS**)

(CE) states the conjunction expansion principle. (ASYM), (IR), and (TRANS) establish, respectively, the asymmetry, irreflexivity and transitivity of preferences. Finally, (DIS*) and (DIS**) state disjunction principles for preferences.

3.5 Doxastic and motivational joint mental attitudes: mutual beliefs, joint desires, joint goals and joint intentions

As maintained in Section 2, an adequate model of CDM is required to provide a coherent set of data structures and conceptual mechanisms both at the individual and group level. This leads us to assume a *meso* perspective whereby both individual constructs and higher-level entities play a role in our attempt to formalize and reason about collaborative behaviour [19]. There are three conceptually distinct steps that, when taken together, constitute a meso approach to CDM: (a) identify the key concepts related to individual behaviour; (b) identify the key concepts related to collective behaviour, and (c) identify the relationship between individual and collective concepts. In Section 3.4 we have analysed the mental state of an individual agent in terms of its beliefs, desires, goals, intentions and preferences (point 1). In this section, we will introduce similar conceptual constructs to describe the doxastic and motivational mental attitudes of collective agents[11] (point 2). Particularly, in formalizing these higher-order attitudes, we will identify a principled representation of the key relationships between the individualistic level and the collective one (point 3).

For each group gr_i and formula φ, our language L includes the modal operators *M-BEL*$(gr_i, \varphi)(t_i)$, *J-DES*$(gr_i, \varphi)(t_i)$, *J-GOAL*$(gr_i, \varphi)(t_i)$ and *J-INT*$(gr_i, \varphi)(t_i)$. *M-BEL*$(gr_i, \varphi)(t_i)$ means that, at time t_i, the group gr_i has a mutual belief that proposition φ holds. A mutual belief is an infinite conjunction of an agent's belief about an agent's belief about an agent's belief and so forth, that φ holds [27]. A joint desire conveys the fact that two or more agents can be motivationally connected by the same state of the world that each of them wishes to bring about. A joint goal points to a state of the world that two or more agents consider both achievable and as a possible candidate for being moved up to joint intention-status. A joint intention conveys the idea that: (a) two or more agents are individually committed to achieving a particular state of the world; and (b) each of them intends the others to be individually committed to achieving that state.

More formally, the semantics for these operators can be defined as follows. First, we examine the semantics of every member of a group having a mental attitude towards a formula.

[11] In what follows, we will not formalize the notion of joint preferences. Indeed, there are two main reasons why in our model we do not need such a higher-order mentalistic concept. First, we take individual preferences to play a central role in forming practical judgements within social practical reasoning (Section 7.2). Second, a joint decision within a group stems from a process of negotiation among a number of agents trying to influence one another to perform some action. In our view, using joint preferences would have obscured the individual practical reasoning mechanisms that are involved in a multi-agent setting, and would have prevented us from formalizing the negotiation process through which an agreement is reached among a group of collaborating agents.

Following [71], we introduce the operator *E-BEL* $(gr_i, \varphi)(t_i)$, which means that, at time t_i, every member of the group gr_i believes that φ holds. We have the following definition:

$$\forall gr_i, \forall t_i \textit{E-BEL}\ (gr_i, \varphi)(t_i) \equiv \bigwedge_{\{a_i | a_i \in gr_i\}} Bel(a_i, \varphi)(t_i).$$

The semantics of *E-DES* (every agent of a group has a desire to attain a state of the world), *E-GOAL* (every agent of a group has a goal to attain a state of the world) and *E-INT* (every agent of a group has an intention to attain a state of the world) are defined analogously.

Now we say that in a group gr_i, at time t_i, it is mutually believed that φ, *M-BEL*(gr_i, φ) (t_i), iff at time t_i all members of gr_i believe that φ and all of them believe that all of them believe that φ and all of them believe that all of them believe that all them believe that φ and so on ad *infinitum* [88]. Let *E-BEL*$^0(gr_i, \varphi)(i_i)$ be an abbreviation for $\varphi(i_i)$, and let *E-BEL*$^{k+1}(gr_i, \varphi)(i_i)$ be an abbreviation for *E-BEL*$(gr_i,$ *E-BEL*$^k(gr_i, \varphi))(i_i)$. Then we have: *M-BEL*$(gr_i, \varphi)(i_i) \equiv$ *E-BEL*$^k(gr_i, \varphi)(i_i)$ for all $k \in \mathbb{N}$ (see [35] for details).

Against this background, we can now formalize joint desires, goals and intentions. In order to establish a joint desire/goal/intention towards a state of the world among the members of a group, a necessary condition is that all members of the group have the individual desire/goal/intention towards that state, and that it is a mutual belief in the group that all members have this desire/goal/intention. However, this condition is not sufficient to establish a joint desire/goal/intention. As an example, let us concentrate on the case of joint intentions. Imagine two agents who are individually committed (have the intention) to achieving the same state of the world. Although it might well be the case that there is a mutual belief among the two agents that both intend to achieve the same state, they might or might not intend that they share the same intention. Agents might simply find themselves holding the same intention. This suggests that what is needed to establish a joint intention towards a state is each agent's intention that the others have individual intentions towards that state. Additionally, there should be a mutual belief among the agents that this is so. The case of joint desires and joint goals is similar to that of joint intentions.

Let us summarize the above conditions for joint desires, goals, and intentions. A group has a joint desire towards φ iff: (a) each member has the desire towards φ; (b) it is mutually believed in the group that each member has a desire towards φ; (c) each member intends that the other members have a desire towards φ; and (d) it is a mutual belief in the group that (c). Joint goals and joint intentions are defined in the same way as joint desires. Formally, $\forall gr_i, \forall t_i$, we have the following definitions:

$$\begin{aligned}
\textit{J-DES}(gr_i, \varphi)(t_i) \equiv\ & \textit{E-DES}(gr_i, \varphi)(t_i) \wedge \\
& \textit{M-BEL}(gr_i, \textit{E-DES}(gr_i, \varphi))(t_i) \wedge \\
& \textit{E-INT}(gr_i, \textit{E-DES}(gr_i, \varphi))(t_i) \wedge \\
& \textit{M-BEL}(gr_i, \textit{E-INT}(gr_i, \textit{E-DES}(gr_i, \varphi)))(t_i)
\end{aligned}$$

$$\begin{aligned}
\textit{J-GOAL}(gr_i, \varphi)(t_i) \equiv\ & \textit{E-GOAL}(gr_i, \varphi)(t_i) \wedge \\
& \textit{M-BEL}(gr_i, \textit{E-GOAL}(gr_i, \varphi))(t_i) \wedge \\
& \textit{E-INT}(gr_i, \textit{E-GOAL}(gr_i, \varphi))(t_i) \wedge \\
& \textit{M-BEL}(gr_i, \textit{E-INT}(gr_i, \textit{E-GOAL}(gr_i, \varphi)))(t_i)
\end{aligned}$$

$$\begin{aligned}
\textit{J-INT}(gr_i, \varphi)(t_i) \equiv\ & \textit{E-INT}(gr_i, \varphi)(t_i) \wedge \\
& \textit{M-BEL}(gr_i, \textit{E-INT}(gr_i, \varphi))(t_i) \wedge \\
& \textit{E-INT}(gr_i, \textit{E-INT}(gr_i, \varphi))(t_i) \wedge \\
& \textit{M-BEL}(gr_i, \textit{E-INT}(gr_i, \textit{E-INT}(gr_i, \varphi)))(t_i)
\end{aligned}$$

Unlike some of the previous work in this area (e.g. [21]), we do not define joint mental attitudes as first-class entities that broadly imply, but do not clearly specify, any combination of the underpinning individual attitudes. Conversely, like other work in this area (e.g. [51]), our definition has the advantage of capturing the interplay between individual mental attitudes that is functionally relevant in generating a higher-order notion of joint mental attitude. Indeed, to model and formalize the joint mental attitudes involved in collaborative activity, it is necessary to model and formalize the individual agents' mental attitudes about each other's mental attitudes. In our characterization of the ontology of joint mental attitudes, we identify three main forms of individual agents' attitudes about others' attitudes. First, each agent's belief that each group member believes that each group member believes and so on that a certain mental attitude is held by each group member. Second, each agent's intention that each group member holds some mental attitude. Third, each agent's belief that each group member believes that each group member believes and so on that each group member intends that each group member holds some mental attitude.

In the last few years much work has been done both in Distributed Artificial Intelligence (DAI) and in the philosophy of mind towards a principled representation of the mentalistic apparatus involved in collaborative activity [14, 20, 42, 54, 84]. Although most of this work agrees in arguing that a formalization of cooperation is grounded upon the formalization of the mental states of the involved agents, there is however no consensus about the connection between individual and joint mental attitudes. Various constraints on, and relationships between, individual and joint mental attitudes have been proposed. For example, while Levesque *et al.* [54] require the agents of a given group to hold that group's goals and intentions, Kinny *et al.* [51] require a group's goals and intentions to be distributed among the constituent agents on the basis of their skills and capabilities. Likewise, Cavedon *et al.* [21] require a joint attitude of a team to entail a joint attitude of the same type in all its subteams.

Clearly, according to our definitions, a joint mental attitude of a group entails a mental attitude of the same type in all its constituent agents. We have, $\forall gr_i, \forall a_i \in gr_i$:

$$\models \textit{M-BEL}(gr_i, \varphi)(t_i) \supset Bel(a_i, \varphi)(t_i)$$
$$\models \textit{J-DES}(gr_i, \varphi)(t_i) \supset Des(a_i, \varphi)(t_i)$$
$$\models \textit{J-GOAL}(gr_i, \varphi)(t_i) \supset Goal(a_i, \varphi)(t_i)$$
$$\models \textit{J-INT}(gr_i, \varphi)(t_i) \supset Int(a_i, \varphi)(t_i).$$

The above theorems express a link between the mental state of an agent to the mental state of the group to which the agent belongs. These properties stem from our definition of joint attitudes, which ensures that mental attitudes of individual agents are propagated upwards to the group of which the agents are part. For example, according to our definition, a collective can be seen as having a joint intention towards φ in so far as all the agents that are part of the collective have the same intention towards φ. As a result, if a collective has an intention towards φ, then all the agents within that collective will have the same intention towards φ.

3.6 Social and joint commitments

Whereas an agent's individual intention towards a state of affairs entails the agent's commitment to acting towards the achievement of that state, a group's joint intention is not sufficient to ensure the group's commitment to performing a joint action. The reason is that a joint intention is not as strongly persistent as an individual intention. A joint intention does not bring about each group member's commitment to being part of the group and to acting in a

collaborative manner. Rather, it only ensures that each member is individually committed to acting, and intends all the others to be individually committed to acting. Hence, a joint intention can be dropped if one (or more) of the group members decides to leave the group for whatever reason. Thus, unless additional conditions are imposed on joint intentions in order to strengthen their persistence, there is no guarantee that a joint commitment to acting collaboratively will ensue. Such additional conditions should ensure at least the following three important facets concerning the relationships between the collaborating agents [18]. First, an agreement among the group members. Second, the right of the group to control each member's behaviour. Third, a deontic aspect concerning the obligation of each member towards the group.

In order to capture these three properties of joint persistent intentions (i.e. joint commitments), we shall now give a formalization of persistence (of a joint intention) in terms of a composite notion that builds upon the concept of social commitment. To this end, we take social commitment to be a primitive notion that expresses the relation between two groups of agents [18]. More precisely, $Comm(gr_i, gr_j, e_i)(t_i)$ is a three-argument relation, where gr_i is the committed group, gr_j is the other group to whom gr_i is committed, and e_i is the action sequence gr_i is committed to gr_j to doing. We define a social commitment between an individual agent and a group and a social commitment between two individual agents as special cases of social commitment between groups:

$$\begin{aligned}
\forall a_i, \forall gr_i, \forall e_i \forall t_i Comm(a_i, gr_i, e_i)(t_i) \quad &\equiv \quad \forall gr_j Comm(gr_j, gr_i, e_i)(t_i) \supset \\
&\quad Singleton(gr_j, a_i)(t_i). \\
\forall a_i, \forall a_j, \forall e_i, \forall t_i Comm(a_i, a_j, e_i)(t_i) \quad &\equiv \quad \forall gr_i, gr_j Comm(gr_i, gr_j, e_i)(t_i) \supset \\
&\quad (Singleton(gr_i, a_i)(t_i) \wedge \\
&\quad Singleton(gr_j, a_j)(t_i)).
\end{aligned}$$

Social commitments can also be expressed in terms of a formula φ that an agent is committed to a group to making true (commitments between two agents or two groups are defined in a similar way). In this case, we have:

$$\begin{aligned}
&\forall a_i, \forall gr_i, \forall t_i, t_j (t_i < t_j) Comm(a_i, gr_i, \varphi(t_j))(t_i) \equiv \\
&\exists e_i \text{ s.t. } [Comm(a_i, gr_i, e_i)(t_i) \wedge (plan(a_i, e_i, \varphi(t_j))(t_i) \vee \\
&\exists e_j \exists t_k (t_i < t_k < t_j) \text{ s.t.} \\
&(plan(a_i, e_i, plan(gr_i, e_j, \varphi(t_j))(t_k))(t_i) \vee \\
&plan(a_i, e_i, plan(\{gr_i, a_i\}, e_j, \varphi(t_j))(t_k))(t_i)))].
\end{aligned}$$

Informally, we say that, at time t_i, agent a_i is committed towards gr_i to making φ true at $t_j (t_i < t_j)$ iff there is some action e_i such that: (a) at t_i, a_i is committed to gr_i to performing e_i; and (b) either at t_i, e_i is a plan for a_i to achieve φ at t_j; or (c) at t_i, e_i is a plan for a_i to allow gr_i to achieve φ at t_j; or (d) at t_i, e_i is a plan for a_i to allow gr_i and a_i to achieve φ collaboratively at t_j.

We are now in a position to formalize the strongest motivational attitude to be considered in our paper: *joint commitment* [18, 31]. We say that a group of agents has a joint commitment to achieving a state of affairs as long as it has a persistent joint intention towards that state. As we shall see, a joint intention is persistent to the extent that it is strengthened by social-commitment relationships among the group members. That is, it is the cogent and normative nature of social commitments (that are in our notion of persistence) that makes a joint persistent intention guarantee the group's joint commitment to acting in a collaborative manner.

Since, according to our formal framework, intentions are primitive notions, and not all intentions are also goals, we cannot use Cohen and Levesque's approach [27] according to which joint commitments are a subset of joint goals (characterized by being persistent) and joint intentions are, in turn, a subset of joint commitments (characterized by the group acting in a particular joint mental state). Indeed, there might well be persistent joint intentions that are not persistent joint goals. Likewise, there might be joint intentions that cannot be characterized as persistent joint intentions, that is, intentions that are as stable as joint commitments. Furthermore, Cohen and Levesque [27] define joint commitments simply in terms of escape conditions without any account of the nature of the relationships between group members. Therefore, their notion fails to explain why a group of agents should be committed to acting in a collaborative way. What is needed to circumvent this shortcoming is a normative constraint on inter-agent behaviour that binds multiple agents into a unitary group where each member is committed to doing its own part towards the achievement of a given state.

Given this, our view is to follow Cohen and Levesque [27] in characterizing joint commitment in terms of persistence. However, unlike Cohen and Levesque, we conceptualize a joint commitment in terms of the persistence of a joint intention rather than of a joint goal. Most importantly, we model the persistence of a joint intention towards a state of affairs by adding the stronger condition that all group members must be individually committed to the group with respect to that state. Thus, our approach allows us: (a) to distinguish between joint goals and persistent joint intentions; (b) to distinguish between persistent and non-persistent joint intentions; and (c) to characterize the persistence of a joint intention in terms not only of some escape conditions but also of normative constraints among the collaborating agents.

DEFINITION 3.1
We say that at time t_i a group gr_i has a joint commitment to making φ true at $t_j (t_i < t_j)$ iff gr_i has a joint persistent intention towards $\varphi(t_j)$. In turn, a group gr_i may be said to have a joint persistent intention towards $\varphi(t_j)$ iff:

(a) in gr_i it is mutually believed that φ will be true at t_j;

(b) gr_i has the joint intention that φ will be true at t_j;

(c) each member $a_i \in gr_i$ is socially committed to gr_i to fulfilling the joint intention;

(d) in gr_i it is mutually believed that each member $a_i \in gr_i$ is socially committed to gr_i to fulfilling the joint intention; and

(e) it is true (and mutual belief in gr_i) that (b) will continue to hold until it is mutually believed in gr_i either that φ will not be true at t_j, or that at least one of the members has no longer the motivation to be part of the group and drops its commitment.

Note that (e) expresses the conditions under which the joint intention may be abandoned. As opposed to Cohen and Levesque [27], our escape conditions refer to the emergence of some new attitudes that are incompatible with the initial commitment. The motivation for this weaker escape condition comes, for example, from cases where a group member adheres to the group's joint intention, but subsequently it has to adopt another intention that is incompatible with the joint one. Condition (e), therefore, provides our model with a certain degree of flexibility as it accounts for all those cases in which the content of the escape conditions is not completely known when the agents endorse a joint commitment to achieving a state of affairs. In fact, in most cases, circumstances may change and it is not always possible to correctly predict the future and to specify in advance the content of the escape conditions under which a joint commitment may be dropped.

We can now give a formalization of the notion of joint commitment. $\forall gr_i, \forall t_i, t_j (t_i < t_j)$, we have:

$$\begin{array}{l}
\textit{J-COMM}(gr_i, \varphi(t_j))(t_i) \equiv \textit{M-BEL}(gr_i, \varphi(t_j))(t_i) \wedge \\
\textit{J-INT}(gr_i, \varphi(t_j))(t_i) \wedge \bigwedge_{\{a_i | a_i \in gr_i\}} [\textit{Comm}(a_i, gr_i, \varphi(t_j)) \wedge \\
\textit{M-BEL}(gr_i, \textit{Comm}(a_i, gr_i, \varphi(t_j)))](t_i) \wedge \gamma(t_i) \wedge \\
\textit{M-BEL}(gr_i, \gamma)(t_i),
\end{array}$$

where:

$$\begin{array}{ll}
\gamma \equiv & [\textit{J-INT}(gr_i, \varphi(t_j))(t_i, t_j) \vee \exists t_k (t_i < t_k \leq t_j) \text{ s.t.} \\
& ((\textit{M-BEL}(gr_i, \neg\varphi(t_j)) \vee \exists a_i \in gr_i \text{ s.t.} \\
& (\neg \textit{Comm}(a_i, gr_i, \varphi(t_j)) \wedge \\
& \textit{M-BEL}(gr_i, \neg \textit{Comm}(a_i, gr_i, \varphi(t_j))))(t_k) \wedge \\
& \forall t_h (t_i \leq t_h < t_k) \textit{J-INT}(gr_i, \varphi(t_j))(t_h))].
\end{array}$$

3.7 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. A role r_i 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* (for similar attempts to develop a cognitive modelling of roles, see [6, 22, 85]). Thus, attached to roles there are such mental attitudes as 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. Roles provide agents with much of the information and many of the goals and intentions that drive their behaviour. For example, assuming a role within an organizational unit may lead the agent to adopt the goal of being a contributor to the unit's success. The role can also provide the agent with some of the information it needs to execute its task within the unit. Likewise, taking on a role may influence the agent to commit itself to proceed forthwith to do what is required of it.

The cognitive characterization 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 intention of supervising the boss's correspondence. This refers to a job specification and the role-player may be obliged to adopt such an intention. However, there might well be the attached goal of being friendly with the other people in the business unit, and the secretary is just expected but not obliged to behave this way.

Note that our notion of mandatory role-based mental attitudes is consistent with the concept of *organizational commitment* [18]. When a member a_i is organizationally 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 organizational 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. In the light of the above conception, our characterization of roles in terms of mandatory mental attitudes

is therefore consistent with the prescriptive account of roles as sets of behavioural obligations based on the role-player's organizational commitment to the group.[12]

In what follows, we want to be able to express facts about the agents in roles.[13] In particular, in order to investigate the impact that a role in a multi-agent system can have on an agent's mental state, the logic proposed here is enriched by the operator $In(a_i, r_i)(t_i)$, which means 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.

3.8 Social relationships

In addition to roles, we have terms that denote *relationship types*, and we use (r_i, r_j), (r_j, r_k), ... and so on as variables ranging over relationship types. A relationship type represents a relationship abstraction between a pair of roles. For example, the roles 'boss' and 'subsidiary' can be linked by a particular type of social relationship that empowers one to dictate the work agenda of the other.

An instantiation of a relationship type gives rise to a *social relationship* between agents. For example, the 'boss-subsidiary' relationship type can give rise to a number of relationship instances — e.g. one involving agent a_i as the boss and agent a_j as subsidiary, another one involving the same agent a_i as the boss and agent a_k as subsidiary.

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:

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

3.9 Social mental shaping

Social mental shaping refers to the phenomenon that the mere social nature of agents affects their mental states, and thereby motivates their behaviour [66]. Its typical outcome is a modification of an agent's mental state, either via the adoption of a new *socially motivated* mental attitude or the modification of an individually motivated one. As outlined in [66], an agent's adoption of mental attitudes can be seen as partly driven by the social environment in which the agent is located. In particular, in [66] we claimed that roles and relationships can play an active causal role, and together they can govern an agent's behaviour in the same way that its individual mental state usually does. In this paper, we enrich and extend the formalism developed in [66] to a setting in which the process of social mental shaping may involve not

[12] Note that although our notion of roles is consistent with the account of roles as sets of behavioural obligations, 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 [57] and Sergot [74].

[13] Note that a *complete* cognitive characterization of roles requires the introduction of modal operators for formalizing the attached mental attitudes. These operators are expected to express the mental attitudes that an agent can internalize by adopting the roles to which they are attached. Like agents' mental attitudes (see Sections 3.4.1–3.4.5), the semantics of role-based mental attitudes should be expressed via accessibility relations between possible worlds. For simplicity, we will not explore this issue any further in this paper and leave it for future investigation.

only social roles and social relationships but also other agents outside of any relationship. In our view, an agent can be seen as a kind of *associative entity*, engaged in an iterated series of social actions and interactions aimed at *completing* its mental state. These social actions and interactions can be described as processes in which the social environment (i.e. roles and other agents within or outside social relationships) complements and augments an agent's bare individual mental attitudes.

In what follows we will introduce a modal operator $Infl$ to formalize the process of social mental shaping in its two basic forms: (a) social mental shaping based on social roles; and (b) social mental shaping occurring between agents. For simplicity, we write $Att(a_i, \varphi)(t_i)$ to indicate that agent a_i, at time t_i, has either a belief that φ holds, a desire towards φ, a goal towards φ, or an intention that φ holds.[14]

3.9.1 Roles and social mental shaping

We formalize the influence of a role on an agent's mental state by expressing the modal operator $Infl$ in terms of a_i, Att, and r_i, where a_i is an agent, r_i is a role and Att has the meaning outlined above. Formally, we have:

$$\forall a_i, \forall r_i, \forall t_i \; Infl(Att(a_i, \varphi), r_i)(t_i) \equiv In(a_i, r_i)(t_i) \wedge (In(a_i, r_i) \supset Att(a_i, \varphi))(t_i).$$

Informally, the meaning of $Infl(Att(a_i, \varphi), r_i)(t_i)$ 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) agent a_i occupies role r_i; and (b) agent a_i adopts or keeps the attitude $Att(a_i, \varphi)$ as a consequence of a_i's taking on role r_i. Note that our distinction between mandatory and optional role-based mental attitudes (see Section 3.7) is fundamental to the problem of preventing automatic attitude-adoption whenever an agent occupies a role in a multi-agent system. On the one hand, when there is a mandatory attitude attached to a role, the agent will *automatically* adopt such an attitude by occupying the role. In such a situation, the role-player is subjected to a social mental shaping process. On the other hand, when there are optional role-based attitudes, the role-player may decide *whether to adopt those attitudes or not*. In particular, there might be optional attitudes attached to roles that an agent decides not to adopt, for whatever reason.

3.9.2 Social mental shaping between agents

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:

(a) *Authority*. An agent may be influenced by another to adopt a mental attitude whenever the latter has the power to guide the behaviour of the former [7].

(b) *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 [53].

(c) *Trust*. An agent may be influenced by another to adopt a mental attitude merely on the strength of its confidence in the latter [41].

[14] Social mental shaping could be formalized also in terms of preferences. That is, an agent might be influenced either by roles or by social relationships to adopt preferences between formulae. As in what follows we are not concerned with such a form of social influence, our focus will be only on the socially driven adoption of beliefs, desires, goals and intentions.

(d) *Persuasion*. An agent may be influenced to adopt another agent's mental attitude via a process of bargaining, argumentation or negotiation [44, 52].

(e) *Threat*. An agent may be threatened to adopt a mental attitude on the basis of future negative interference or denied help [19, 52].

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 to to keep a mental attitude. However, not all modes of social influence can be suitably exercised in all circumstances. 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 (see our definition of social relationship in Section 3.8). For example, authority can be exercised even in the absence of any social relationship between agents (e.g. the authority exercised by the Prime Minister over the citizens). On the other hand, in most cases persuasion can be exercised when a social relationship already exists between the agents involved (e.g. friendship relationship). Furthermore, note that 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, each form of social mental shaping between agents — with and without social relationships — will be dealt with.

Social mental shaping outside social relationships. To formalize 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:

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

Informally, $Infl(Att(a_i, \varphi))(t_i)$ means that, at time t_i, agent a_i is socially influenced to hold mental attitude Att. This form of social mental shaping happens whenever an agent believes that another agent has a mental attitude, and for this reason it adopts or keeps that mental attitude. This covers several forms of social influencing, from imitation to spontaneous goal-adoption, from benevolent (not due) adhesion to emulation [19]. Note that in all these cases of social mental shaping the two agents *need not be in a social relationship*, as opposed to the form of social mental shaping detailed below.

Social relationships and social mental shaping. We now want to formalize how an agent's mental state can be influenced by its being within a social relationship with another agent. We have:

$$\forall a_i, a_j, \forall (r_i, r_j), \forall t_i Infl(Att(a_i, \varphi), rel(a_i, a_j, (r_i, r_j)))(t_i) \equiv rel(a_i, a_j, (r_i, r_j))(t_i) \wedge Att(a_j, \varphi)(t_i) \wedge Bel(a_i, Att(a_j, \varphi))(t_i) \wedge (Bel(a_i, Att(a_j, \varphi)) \supset Att(a_i, \varphi))(t_i).$$

Informally, if an agent a_i, which is in a social relationship of type (r_i, r_j) with another agent a_j, believes that a_j has a mental attitude Att, and for this reason it adopts or keeps Att, then we can say that a_i is influenced by its being situated within a social relationship with a_j. In general, this form of social mental shaping is based and depends on the agent's decision whether to adopt one of its acquaintance's mental attitudes or not. 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 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 attitudes*. For example, the boss is by right allowed to order other employees to perform particular activities. If a secretary decides to interact (i.e. establish a social relationship) with a boss, then it might well be the case that he or she ought to change his or her mental state so as to adopt some of the intentions imposed by the boss.

4 A formal model of CDM: an overview

In this section, we present an overview of our four-stage model of CDM, which we will then formalize by expressing it in the logic described in Section 3. Our model is based on Wooldridge and Jennings' formalization of the cooperative problem-solving process [88]. There are several similarities between our approach and Wooldridge and Jennings' work. First, we are inspired by implementation-based models for realizing cooperative systems, in that our aim is to identify the basic steps of the CDM process. Second, our approach is to characterize the mental attitudes of the agents involved in a CDM process. Third, our model aims at being comprehensive, in that it should cover the entire CDM process. However, in contrast to [88], our focus is primarily on the dual-faceted socio-cognitive processes that are involved in CDM. That is, on processes in which the mental modelling is strictly coupled with an account of sociality and the interaction capabilities of the cooperating agents (Section 1). We explicitly represent mechanisms for influencing other agents' mental states and behaviour in interactions between autonomous agents (Section 6). Furthermore, we provide the theoretical foundations of the inferential decision-making apparatus of the interacting agents, and we give a comprehensive account of practical reasoning processes within a social setting (Section 7). The four stages of our model are the following:

(a) *The practical starting-point*: As with the individual case, the CDM process begins when there is a state of the world that at least one agent intends to realize. That agent will then be confronted with a practical problem; that is, a problem about what to do. Insofar as the agent solves its problem in isolation, we will have a typical individual decision-making situation. CDM can occur whenever (a) there is a potential for cooperation over the resolution of a practical problem and (b) such a potential is relevant to the agent and can therefore be recognized (see Section 5).

(b) *Group generation*: During this stage, the agent that recognized a potential for cooperation at stage (1) will solicit assistance. If this stage is successful, then it will end up with a group characterized by a joint practical problem based on a joint commitment to achieving a state of the world (see Section 6).

(c) *Social practical reasoning*: During this stage, each member of the newly formed group will reason about what course of action the group should perform in order to fulfil its joint commitment and find an answer to the practical problem. If this stage is successful at least one agent will have a practical judgement about what action should be performed by the group, and will form the corresponding intention that the group performs that action (see Section 7).

(d) *Negotiation*: During this stage, each member will attempt to bring about an agreement within the group about the action to be performed. If this stage is successful, then an

action will be agreed upon and a joint intention along with a joint commitment of the whole group to acting accordingly will ensue (see Section 8).

Our model, although complete (in that it covers CDM from beginning to end), is aimed at providing a description of CDM in an *idealized* world. Among the assumptions that we have made and that will be detailed in what follows, there is one that is worth noting here as it deals with the overall structure of the model. Like in [88], we have assumed for simplicity that our model is strictly *sequential*; that is, each stage directly follows the one which precedes it without back-tracking. However, in reality a CDM process is inherently iterative, in that if one stage fails, the agents involved are allowed to return to previous stages.

Finally, our model does not deal with the agreed-upon action execution. The final stage, if successful, concludes with an agreement about an action to be performed by the group. This contrasts with other approaches that support the idea that the conclusion of either practical reasoning or decision-making should be an action (e.g. [1, 88]). However, an individual agent or a group of agents can make a decision to do something and at the same time be unexpectedly prevented from acting accordingly. Or, for whatever reason, they may decide to postpone the execution of the agreed-upon action. Or, after a decision has been made, they may even change their mind and decide not to perform the newly agreed action. Or they might fail to execute the action. In all these cases, we still have individual agents and/or groups that have made a decision, although not followed by action.

In our view, deciding simply means to give an answer to a practical problem. An answer to a practical problem involves (a) a practical judgement in favour of a specific action; and (b) a commitment to acting accordingly. Commitment, in turn, is simply a pledge or promise to perform a particular action, provided that circumstances do not change. However, in most realistic scenarios, agents are situated in time-varying environments: they may become aware of new information, and the external world may change. Therefore, circumstances may alter between the making of a commitment and the performance of the associated action. This suggests that decisions, built on commitments, might not be followed by action. Still, they provide an answer to a practical problem and in this sense can be viewed as the concluding element of CDM. Note that all this is consistent with the Aristotelian idea that successful practical reasoning concludes in action, just as theoretical reasoning concludes in justified belief. But what concludes in action need not have action as its conclusion.

5 The practical starting-point

The practical starting-point of CDM can be analysed into three components:

(a) The practical basis — that is, a characterization of *what an agent intends to achieve* (Section 5.1).

(b) The practical problem — that is, a problem about *what to do to fulfil an intention* in a given environment (Section 5.2).

(c) Recognition of the potential for cooperation — that is, the *identification by an agent of an opportunity to collaborate* with one or more agents over the resolution of a practical problem (Section 5.3).

5.1 The practical basis

The starting-point in CDM must involve some *state of the world* that at least one agent *intends* to achieve. An intention towards a state triggers the decision-making process as it provides the agent with a problem of what to do. That is, the agent is confronted with a situation in which it must decide what course of action is to be performed in order for it to attain a given state [12]. The agent will then consider to what extent the fulfilment of its intention is affected by the actions open to itself.

Decision-making can thus be regarded as directed at the realization of some state of affairs that an agent intends to bring about. For example, the agent might intend either to possess some object, or to attain some condition, or even to realize some situation not connected directly with the agent at all. The agent's intention will be said to constitute the *practical basis* of subsequent decision-making [39]. More precisely, an intention can be seen as the practical premiss of a piece of decision-making whenever it possesses *practical force* in raising a practical problem, that is a problem of what is to be done in order for the agent to achieve a given state of the world. At least one intention with practical force is required in any line of decision-making.

5.2 The practical problem

While the answer to theoretical problems is found in knowing something, in understanding, the answer to a practical problem is found in a decision concerning what to do in a given situation (Section 1). Once the agent has made a decision, the practical problem no longer presents itself as a problem. The decision may turn out to be either successful or unsuccessful. In the former case, the desired state of the world will be achieved and the agent's intention fulfilled. In the latter, a new practical problem will arise, and a new decision about what to do will have to be made.

A practical problem is related both to the particular agent and to the physical and social environment in which the agent is located [39]. First, it confronts a certain agent whose capacities, outlook, and achievements limit its actions. Second, it must be met by a decision and eventually by an action to be performed within a specified physical and social environment. The relation of practical problems to the agent's physical and social environment is an overriding factor in considering the principles and methods of decision-making. These principles and methods must enable the agent to resolve a practical problem under the limitations that its environment imposes.

5.3 Recognition of the potential for cooperation

A practical basis and a related practical problem are insufficient to give rise to a CDM process. On the one hand, the agent might be able and/or willing and/or socially influenced (e.g. obliged) to resolve its practical problem in isolation. On the other, although incapable and/or unwilling to resolve its problem in isolation, or even socially influenced to solve it in a collaborative manner, the agent might not be able to lean on other agents within the social environment in which it acts. What makes a CDM commence is: (a) the existence of a *potential* for cooperation over the resolution of a practical problem, and (b) the *recognition* of such a potential by the agent.

We explicitly separate out these two stages because a potential for cooperation is insuf-

ficient by itself to trigger subsequent phases of the CDM process. Even though an agent must be able to get other agents involved in a joint decision, nonetheless such a potential for cooperation must also be valuable, relevant to the agent, in order for it to be conveniently exploited. On the one hand, an agent might be aware of a potential for cooperation, but it might have no need to exploit it. On the other hand, an agent might need to cooperate with other agents but no cooperation might be attained.[15]

Potential for cooperation. There is a potential for cooperation insofar as (a) either an agent, say a_i, already is in a social relationship with at least another agent, say a_j, or a_i can bring about such a social relationship[16] with a_j; and (b) a_i has the ability to get a_j involved in some form of cooperation.[17] This ability points to the feasibility of a social mental shaping process through which a_i will endeavour to make a_j at least adhere to a_i's own intention to achieve a given state of affairs, say φ. As we shall discuss later (Section 6), a_j's adhesion to a_i's intention represents the first step towards the generation of a joint practical basis for subsequent CDM. In order for such a joint basis to be brought about, a social mental shaping process will be required through which a_j is influenced at least to adopt a_i's intention to achieve φ.

Against this background, a more precise definition of potential for cooperation can now be given in the following terms:

DEFINITION 5.1 (Potential for Cooperation)
With respect to agent a_i and a state of affairs φ that a_i intends to achieve, there is potential for cooperation iff:

(a) there is at least one agent a_j with which a_i already has a social relationship, or with which a_i believes that a relationship can be brought about; and

(b) a_i believes that it can exercise upon a_j a social mental shaping process through which a_j can be influenced to adhere to a_i's intention to achieve φ.

We now concentrate on the conditions under which a potential for cooperation can be recognized by an agent.

Recognition. A potential for cooperation must be valuable to an agent in order for it to be recognized and exploited. Essentially, recognition may take place whenever there is a social dependence relationship between two (or more) potentially cooperating agents. Social dependence, with respect to the achievement of a state φ, may occur for the following reasons:

---

[15] Here we are assuming that if cooperation can be established, the agent knows about such a potential. Thus we are ruling out all those circumstances in which an agent needs to cooperate with other agents, there is a potential for cooperation, but cooperation is not achieved due to the agent's being unaware of such a potential.

[16] In Section 3.9.2 we mentioned a number of modes of social influence between agents. An agent can obviously adopt one of these modes in order to establish a social relationship with another agent. The choice of the most effective mode will depend upon a number of factors, such as the physical and social environment, the agents' characteristics, etc. In what follows we will not discuss this issue any further, and leave it for future work.

[17] In our definition of potential for cooperation, we are referring to strictly cooperative activity [19]. Thus, we are not considering all those cases in which, for example, a number of agents are cooperating over the resolution of a practical problem, but only one of them — the central planner — is aware of such joint activity, while the others are unaware and/or even not interested in the results of the overall collaborative activity (orchestrated cooperation). For simplicity, we also rule out forms of emergent social activity. In particular, our model does not allow cooperation to be based on such weaker grounds as accidental or unaware cooperation, implicit communication, tacit agreement, non-mutual beliefs, etc. [19].

(a) An agent is *unable* to fulfil its intention to achieve φ in isolation and there are other agents that can assist it in the fulfilment of this intention. For example, an agent may have an intention to achieve a state that can be attained only through information accessible to another agent; without some form of social interaction with this other agent, the state cannot be achieved.

(b) An agent does not want (does not have the intention) to fulfil its intention to achieve φ in isolation and has the intention to fulfil it with other agents in a collaborative manner. Here the agent may well need no help as it may be able to achieve φ in isolation, but notwithstanding this ability it may well hold the intention to achieve φ collaboratively. There may be many reasons underpinning this intention towards cooperation. For example, the agent might believe that the fulfilment of the intention towards φ in isolation will eventually bring about a number of difficulties in the fulfilment of other intentions. Alternatively, it might believe that a collaboration with another agent will in some way turn out to be a better solution (e.g. it is supposed to bring about φ in a faster or more accurate way than acting in isolation). Or, the agent might just have the desire to act in a collaborative manner with another agent, although such a collaboration may not bring about a more effective fulfilment of the agent's intention.

(c) An agent is *influenced by the role(s)* it has adopted not to fulfil its intention towards φ in isolation, but to involve other agents in some form of collaboration aimed at achieving φ. Here, the agent may well be able to achieve φ in isolation, and it may also have no individually motivated intention to achieve φ collaboratively. However, it may be socially influenced by the adopted role(s) to establish some form of collaborative activity over the achievement of φ. For example, the head of a department might be expected to involve the head of another department in the resolution of a particular problem, where the role of the latter is to ensure that similar academic standards are applied throughout the university. In such a case, although the agent has the ability and is willing to act in isolation, the role it has taken on — i.e. head of department — induces it to seek cooperation.

(d) An agent is *influenced by other agents* (either within or outside social relationships) not to fulfil its intention towards φ in isolation, but to establish some form of collaboration aimed at achieving φ. Again, like in (c), here the agent may well be able and willing to achieve φ in isolation, but still be influenced by other agents to achieve φ collaboratively. For example, a secretary might be obliged by his or her boss to involve a colleague in the execution of a task. In this case, although the secretary may have the ability and be willing to execute the task in isolation, the relationship he or she has established with the boss induces him or her to seek cooperation.[18]

In the light of these observations, a more precise definition of recognition of a potential for cooperation can now be given in the following way:

DEFINITION 5.2 (Recognition of a potential for cooperation)
Given an agent a_i, and a state of affairs φ, whenever the conditions for a potential for cooperation with another agent, say a_j, are satisfied, a_i recognizes such a potential iff a_i depends on some form of collaborative activity with a_j for achieving φ.

[18]Note that points (c) and (d) are aimed at incorporating the prescriptive side of social dependence into a formal definition (see discussion below). Whenever an agent is required to establish cooperation, this requirement brings about social dependence between two or more agents. In what follows, we will model this normative aspect of social dependence by appealing to both social mental shaping based on roles (Section 3.9.1) and social mental shaping occurring between agents (Section 3.9.2).

In what follows, we will give a formalization of the above definitions. To this end, we need to enrich our formal framework by introducing two derived operators. On the one hand, the notion of potential for cooperation is couched in terms of the agent's ability to establish a social relationship and to exercise a social mental shaping process. On the other, the definition of recognition of a potential for cooperation is couched in terms of the notion of social dependence between agents, which in turn builds upon the concept of ability both for individual agents and groups of agents. Therefore, in order to formalize the above definitions, we need to introduce two derived modal operators that express the notions of *single-agent ability and multi-agent ability*.

There is currently no consensus in the literature about the most appropriate definition of ability (e.g. [76]). Rather than defining ability as a primitive modal operator, here we follow Wooldridge and Jennings [88] in adopting a definition that was originally proposed by Moore [64].

DEFINITION 5.3 (Single-agent ability)
Agent a_i can (has the ability to) achieve a state φ iff there is some action sequence e_i that is a plan for a_i either to achieve φ directly or to find out how to achieve φ.

Clearly, in our definition of single-agent ability, we allow for two different possibilities. First, a particular action sequence might be a plan for the agent to achieve state φ directly. Second, an action sequence might be a plan for the agent to get closer to φ.

In what follows, we now give a formal expression of the notion introduced above. We introduce the operator $Can(a_i, \varphi(t_j))(t_i)$ to express the fact that, at time t_i, agent a_i has the ability to achieve φ in isolation at time $t_j (t_i < t_j)$:

$$\begin{aligned}\forall a_i, \forall t_i, t_j (t_i < t_j) Can(a_i, \varphi(t_j))(t_i) \equiv \exists e_i \text{ s.t. } [&plan(a_i, e_i, \varphi(t_j))(t_i) \vee \\ &\exists e_j \exists t_k (t_i < t_k < t_j) \\ &\text{s.t. } plan(a_i, e_i, plan(a_i, e_j, \varphi(t_j))(t_k))(t_i)]\end{aligned}$$

Note that the notion of single-agent ability is not closed under conjunction; that is, $Can(a_i, \varphi(t_j))(t_i) \wedge Can(a_i, \psi(t_j))(t_i)$ need not imply $Can(a_i, \varphi(t_j) \wedge \psi(t_j))(t_i)$. For example, if agent a_i has the ability to buy pens for a dollar each, then if it has a dollar it can buy a black pen or a red one, but it has no ability to buy both [9].

We can now generalize the definition of single-agent ability to the multi-agent case:

DEFINITION 5.4 (Group ability)
A group gr_i can (has the ability to) achieve a state φ iff there is some action sequence e_i that is a plan for gr_i either to achieve φ directly or to find out how to achieve φ [88].

Once again, as happens in the single-agent case, we allow for two possibilities. First, there might be a particular action sequence that is a plan for the group to achieve state φ directly. Second, there might be some action sequence that is a plan for the group to get closer to state φ.

A formalization of the notion of group ability can be given in the following way:

$$\begin{aligned}\forall gr_i, \forall t_i, t_j (t_i < t_j) \textit{J-CAN}(gr_i, \varphi(t_j))(t_i) \equiv \exists e_i \text{ s.t. } [&plan(gr_i, e_i, \varphi(t_j))(t_i) \vee \\ &\exists e_j \exists t_k (t_i < t_k < t_j) \text{ s.t.} \\ &plan(gr_i, e_i, plan(gr_i, e_j, \varphi(t_j))(t_k))(t_i)].\end{aligned}$$

We can now turn to the concept of *social dependence* between agents [74], on which the notion of recognition of a potential for cooperation is based. Intuitively, if an agent intends

to achieve a given state of affairs, but it does not have the ability or does not intend to achieve it in isolation, or is socially influenced to achieve it collaboratively, then it may depend on some other agents with respect to the achievement of that state. More precisely, using the various definitions above, we can now state the conditions that characterize a social dependence relation between two potentially collaborating agents.

DEFINITION 5.5 (Social dependence)
Agent a_i depends on some form of collaborative activity with agent a_j, with respect to a given state φ, iff:

(a) a_i intends to bring about φ;

(b) a_i believes that a_i and a_j can jointly achieve φ; and

(c) either it is true (and believed by a_i) that a_i does not have the ability to achieve φ in isolation; or

(d) a_i does have the intention not to perform any of the action sequences that it believes are plans for a_i to bring about φ; or

(e) a_i is socially influenced by the role(s) it has adopted to collaborate with a_j in the achievement of φ; or

(f) a_i is socially influenced by other agents (either outside or within social relationships) to collaborate with a_j in the achievement of φ.

More formally, we have:

$$
\begin{array}{l}
\forall a_i, a_j, \forall t_i, t_j (t_i < t_j) DEP(a_i, a_j, \varphi(t_j))(t_i) \equiv \\
Int(a_i, \varphi(t_j))(t_i) \wedge Bel(a_i, \textit{J-CAN}(\{a_i, a_j\}, \varphi(t_j)))(t_i) \wedge \\
[(\neg Can(a_i, \varphi(t_j))(t_i) \wedge Bel(a_i, \neg Can(a_i, \varphi(t_j)))(t_i)) \vee \\
\forall e_i (Bel(a_i, plan(a_i, e_i, \varphi(t_j))) \supset Int(a_i, \neg \langle plan(a_i, e_i, \varphi) \rangle (t_j)))(t_i) \vee \\
\exists e_i \exists r_i \text{ s.t. } Infl(Int(a_i, \langle plan(\{a_i, a_j\}, e_i, \varphi) \rangle (t_j)), r_i)(t_i) \vee \\
\exists e_i \text{ s.t. } Infl(Int(a_i, \langle plan(\{a_i, a_j\}, e_i, \varphi) \rangle (t_j)))(t_i) \vee \\
\exists e_i \exists (r_i, r_j) \text{ s.t.} \\
Infl(Int(a_i, \langle plan(\{a_i, a_j\}, e_i, \varphi) \rangle (t_j)), rel(a_i, a_j, (r_i, r_j)))(t_i) \vee \\
\exists e_i \exists a_k (a_k \neq a_j) \exists (r_i, r_j) \text{ s.t.} \\
Infl(Int(a_i, \langle plan(\{a_i, a_j\}, e_i, \varphi) \rangle (t_j)), rel(a_i, a_k, (r_i, r_j)))(t_i)]
\end{array}
$$

In our definition of social dependence the modal operator $Infl$ is primarily aimed at capturing the *prescriptive sources of social dependence*. Such sources reflect the influence that regulations, norms and authority might have upon an agent's mental state and behaviour by making the agent socially dependent on others [19, 60, 61]. For example, an agent might be entirely capable and willing to fulfil its own intentions, but still be required to lean on another agent and establish some form of cooperation with it. The operator $Infl$ captures these aspects by appealing to the process of social mental shaping that is exercised either by roles or by other agents (within or outside social relationships). First, an agent may be required to depend on another agent and cooperate with it whenever a norm, or a regulation influences the former not to act in isolation. As was shown in Sections 3.7 and 3.9.1, we conceive a role as a set of attached mental attitudes that may have an impact upon the agent's mental state and behaviour through a process of social mental shaping. In this view, some of the norms that are reflected by the attitudes attached to roles may require the role-player to depend on

others and eventually establish cooperative activity with them.[19] Second, an agent may be induced to depend on others simply because it is subjected to the social influence exercised by another agent. This type of social influence may take place either outside or within social relationships (Section 3.9.2). In either case, an agent may be required by another agent to adopt the intention to achieve a state of the world in a collaborative manner, rather than in isolation. For example, should an agent be socially committed to another to achieving something, the latter may have the authority to change the former's mental state and induce it to seek cooperation. In this case, a social dependence relation is brought about by a social mental shaping process occurring between two agents.

We can now give a formal expression of the notion of potential for cooperation, from the perspective of an agent a_i, and with respect to another agent a_j and a state of affairs φ:

$$\begin{array}{l}\forall a_i, a_j, \forall t_i, t_j (t_i < t_j) PfC(a_i, a_j, \varphi(t_j))(t_i) \equiv \\ \exists (r_i, r_j) \exists t_k (t_i < t_k < t_j) \text{ s.t.} \\ [rel(a_i, a_j, (r_i, r_j))(t_i) \vee Bel(a_i, Can(a_i, rel(a_i, a_j, (r_i, r_j))(t_k)))(t_i)] \wedge \\ Bel(a_i, Can(a_i, Infl(Int(a_j, \varphi(t_j)), rel(a_i, a_j, (r_i, r_j)))(t_k)))(t_i).\end{array}$$

Informally, for an agent there is potential for cooperation with another over the attainment of a state of affairs whenever a social relationship between the two agents already exists or the former believes that a relationship can be brought about, and the former also believes that it can influence the latter to adopt the intention to achieve that state. Note that in our definition of potential for cooperation an agent needs to know *a priori* the identity of another agent with which it can cooperate to achieve a state of affairs. For simplicity, we expressed the argument of such notion in terms of two agents. However, we might well rewrite $PfC(a_i, a_j, \varphi(t_j))(t_i)$ as $PfC(a_i, \varphi(t_j))(t_i)$ by adding '$\exists a_j$' to the right-hand side of the identity above. This alternative way of defining a potential for cooperation leaves open to the agent the opportunity to find out the most appropriate acquaintance to cooperate with and thus allows for a searching process aimed at attracting help. Furthermore, we do not require that cooperation be established on an already existing social relationship between two agents. This does not preclude an agent from establishing some form of cooperation with another agent that is outside any of the former's existing social relationships.

Finally, we can give a formal expression of the notion of recognition of a potential for cooperation:

$$\begin{array}{l}\forall a_i, a_j, \forall t_i, t_j (t_i < t_j) RPfC(a_i, a_j, \varphi(t_j))(t_i) \equiv \\ PfC(a_i, a_j, \varphi(t_j))(t_i) \wedge DEP(a_i, a_j, \varphi(t_j))(t_i).\end{array}$$

Informally, one agent recognizes a potential to cooperate with another iff: (a) there is such a potential; and (b) the former is socially dependent on some joint activity with the latter.

6 Group generation

During this stage, the agent(s) that recognized the potential for cooperation will try to solicit assistance and involve other agents over the resolution of the practical problem. If this stage is successful, then it will end up with a group with decision-making purposes. However, merely identifying a potential for cooperation is insufficient to ensure that CDM will begin.

[19] For simplicity our focus is only on role-based prescriptions although a more comprehensive approach should also account for other types of prescriptions. For example, an agent may be required not to act in isolation by a norm that is not reflected in any of the roles it has taken on.

An agent might recognize its dependence on other agents over a particular state of affairs, and still decide not to solicit assistance from them. It could give up its own intention or revise it or even leave it temporarily to be fulfilled some time in the future. Therefore, we will have to make a number of assumptions about agents' behaviour. These assumptions are primarily concerned with the fundamental facets that characterize the process leading to the formation of a web of social and mental links binding together the members of a group with decision-making purposes. There are three main steps involved in this process:

(a) the generation of identical individual intentions to achieve a state of the world;
(b) the generation of a joint intention to achieve that state; and
(c) the generation of a joint commitment to acting in a collaborative manner towards the achievement of that state.

In what follows, each of these steps will be dealt with in turn.

6.1 Individual intention generation

To operate successfully towards the generation of a group with decision-making purposes, the agent who seeks assistance will have to influence[20] other agents to establish a new[21] group and act in a collaborative manner. We express this process of social influence in terms of the agent's attempt to exercise a social mental shaping process aimed at influencing other agents' mental states [66]. Here, social mental shaping is primarily intended to bring about a *joint practical basis*. This basis is meant to have practical force in raising a *joint practical problem*, that is, a problem of what is to be done by the group in order for it to achieve a state of the world (Section 5.1). Each group member's individual intention to achieve that state seems to be necessary for establishing a joint practical basis. An individual intention has practical force and involves some form of 'self-commitment' to acting towards its fulfilment. This suggests that the agent who recognized a potential for cooperation over the resolution of a problem of what to do in order to achieve some state, should attempt to seek assistance by inducing other agents to adopt *at least* its own *intention* to achieve that state. In our model, a group of agents having *identical individual intentions* is the first step towards the generation of a group with decision-making purposes. Accordingly, we can express the following assumption about agents' behaviour:

Assumption: (Individual intention generation) If agent a_i recognizes a potential for cooperation with agent a_j with respect to its intention of achieving φ, then (as long as it keeps recognizing such a potential) a_i will attempt to exercise a social mental shaping process aimed at making a_j hold:

(a) the intention to achieve φ; or, failing that, at least
(b) the belief that a_i has the intention to achieve φ.

We will now formalize the above assumption. For this purpose, we need to introduce a formal expression for the notion of an attempt by an agent to achieve a state of affairs by

[20] As mentioned in Section 3.9.2, the notion of influence can be regarded as a generalization of various forms of interaction such as those based on negotiation, persuasion, threat, authority, appeal to past rewards, appeal to prevailing practice, etc.

[21] We assume that the agent who seeks assistance will have to establish a *new* group instead of just adding new agents to a group that already exists or making an existing group adopt a new joint intention.

performing an action. To this end, we introduce the complex action *Attempt*$(a_i, e_i, \varphi, \psi)$ to express an attempt by a_i to achieve φ by performing e_i, at least achieving ψ (see also [26]):

$$Attempt(a_i, e_i, \varphi, \psi) \equiv ([Bel(a_i, \neg\varphi) \wedge Agt(a_i, e_i) \wedge Goal(a_i, Occurs(e_i; \varphi?)) \wedge Int(a_i, Occurs(e_i; \psi?))]?; e_i).$$

Our first main assumption about individual intention generation can now be stated in the following way:

$$\models \forall a_i, a_j, \forall t_i, t_j (t_i < t_j) RPfC(a_i, a_j, \varphi(t_j))(t_i) \supset$$
$$\exists t_k (t_i < t_k < t_j), \exists e_i, \exists (r_i, r_j) \text{ s.t. } (RPfC(a_i, a_j, \varphi(t_j)) \Leftrightarrow$$
$$Occurs[Attempt(a_i, e_i, Infl(Int(a_j, \varphi(t_j)), rel(a_i, a_j, (r_i, r_j))),$$
$$Infl(Bel(a_j, Int(a_i, \varphi(t_j))), rel(a_i, a_j, (r_i, r_j))))])(t_i, t_k).$$

Informally, if agent a_i recognizes a potential of cooperation with another agent a_j with respect to a state of affairs φ, then (as long as it keeps recognizing such a potential) it will attempt to impact upon a_j's mental state by making a_j adopt the intention to achieve φ, or, failing that, at least by making a_j believe that a_i has the intention to achieve φ.

Clearly, there might be groups in which a decision is made in a collective manner, although not every member is directly involved in the decision-making. For example, in many kinds of groups, some agents might agree to participate and give their contribution to a joint decision without sharing the same intentions of the group. They might not be interested in the final decision, but only in being part of the group. Or, they might be obliged to join the group and agree to be coordinated by other members, without even being aware that a decision is being made within the group. In all these cases, we have agents that are members of a group but do not share the group's ultimate intention. Most importantly, these agents are members of a group that makes a decision but are not decision-makers by themselves. Simply, they give their contribution for a decision to be made on their behalf by other members. Given this, in what follows we will not cover such situations of 'orchestrated' decision-making [19]. Our focus will be exclusively on decisions that are 'strictly' collaborative, that is, that are made by a number of agents who share the same practical problem (and intention) and are aware of being directly and individually involved in the decision-making process.

Finally, we recognize that in real-world situations, a process of social mental shaping aimed at generating individual identical intentions is not strictly necessary to account for CDM. In many kinds of groups, teams, and organizations, people participate in a CDM process without influencing one another to adopt a particular intention. Each agent might well have formed its own intention individually, outside of any social mental shaping process. It might simply happen that a collection of agents share the same practical problem and find it convenient to join their decision-making capabilities to solve that problem in a more effective manner. In these cases, CDM commences without the need to influence someone else's mental state towards the adoption of an intention. Note, however, that in this paper our aim is to give an account of the *entire* process of CDM, from the very beginning when an individual agent requests assistance by influencing others to adopt its own intention through to an agreement made within the group. Hence, our need to talk about social mental shaping and individual intention generation.

6.2 Joint intention generation

Agents sharing identical individual intentions are not necessarily motivated to act towards the fulfilment of their intentions in a collaborative manner. Consider the case of two scientists

holding the same intention (e.g. to find out the cause of a disease) [18]. Each scientist might have generated this intention in isolation, independently of the other, or he or she might have been influenced by the other to adopt the intention. However, in either case, they are not necessarily expected to collaborate in a coordinated way to fulfil their 'parallel' intention. Indeed, they might even be in competition with each other, and each of them might also intend (and have the goal/desire) that the other drops the intention. In this case, identical individual intentions entail no joint mentalistic notion to support any form of collective behaviour.

Thus something more seems to be necessary. Given two (or more) agents holding identical 'parallel' intentions, a necessary subsequent step towards CDM is that *each agent intends the other(s) to hold that intention*. That is, what is needed is a *joint intention* as a necessary condition for a group of agents to act together over the resolution of a practical problem. According to our definition of joint intentions (Section 3.5), if the group members are individually committed and intend to achieve a state φ, are aware of their being committed, intend that each member is equally committed, and finally are mutually aware of this, then the group can be said to jointly intend to achieve φ. By binding each agent's individual intention together in a common mental state, joint intentions provide a necessary cognitive step towards the generation of a group with decision-making purposes. Therefore, we can express our second assumption about agents' behaviour:

Assumption: (Joint intention generation). If agent a_i recognizes a potential for cooperation with agent a_j with respect to its intention to achieve φ, and is successful in exercising a social mental shaping process aimed at influencing a_j to adopt the same intention to achieve φ, then (as long as it keeps recognizing such a potential for cooperation) a_i will attempt to generate:

(a) a joint intention (held jointly by a_i and a_j) to achieve φ; or, failing that, at least
(b) a mutual belief (held jointly by a_i and a_j) that a_i intends that a_i and a_j jointly intend to achieve φ.

Expressing the above assumption formally, we have:

$$\begin{array}{l} \models \forall a_i, a_j, \forall t_i, t_j (t_i < t_j), \forall (r_i, r_j) RPfC(a_i, a_j, \varphi(t_j))(t_i) \wedge \\ Infl(Int(a_j, \varphi(t_j)), rel(a_i, a_j, (r_i, r_j)))(t_i) \supset \\ \exists t_k (t_i < t_k < t_j), \exists e_i \text{ s.t. } (RPfC(a_i, a_j, \varphi(t_j)) \Leftrightarrow \\ Occurs[Attempt(a_i, e_i, \textit{J-INT}(\{a_i, a_j\}, \varphi(t_j)), \\ \textit{M-BEL}(\{a_i, a_j\}, Int(a_i, \textit{J-INT}(\{a_i, a_j\}, \varphi(t_j)))))])(t_i, t_k). \end{array}$$

According to the constraints we have imposed on the logic of mental attitudes, joint intentions towards a state imply each group member's individual intention towards that state. However, they do not entail each member's desire and goal towards that state. We have a two-fold distinction:

(a) The group may have a joint intention to achieve a state and each member has an individual intention, desire and goal towards that state. This might occur, for instance, in groups based on mutual knowledge of mutual dependence. In these cases, all members might be interested (have a desire) in achieving an identical state of the world, and might also have the goal to achieve it.
(b) The group may have a joint intention to achieve a state and each member has an individual intention towards that state. However, some of the members may not hold the desire and/or the goal towards that state. For instance, in many kinds of groups, teams, and organizations, agents share identical intentions without sharing the same desires. Some

of the group members might not be interested in what they intend to bring about, but just in their personal benefits (rewards). Or, it might well be the case that some of the members are socially influenced (e.g. forced, obliged) to join the group and achieve something collaboratively without having the goal to do so. In these cases, the group's joint intention towards a state entails the members' individual intentions towards that state, but is not reinforced by the members' individual desires/goals towards that state. This can be accounted for within our formal framework, since our axiomatization of mental attitudes allows us to represent intentions that are not goals and/or desires at the same time (Section 3.4.4).

6.3 Joint commitment generation

Joint intentions are a necessary cognitive ingredient of CDM, since they provide a first (weak) foundation of collaborative activity within a group. However, although necessary, they are not still sufficient in order for a group with decision-making purposes to be established. This is due, more generally, to their inherent socio-cognitive weakness and, specifically, to the fact that they can be easily dropped by the agents (Section 3.7). In order for a joint intention to be fulfilled collaboratively, it must have a certain degree of stability, that is, it must not be dropped for whatever reason, at least until certain escape conditions become true [27]. However, the members of a group characterized simply by a joint intention have no cogent or normative constraints that can ensure such a stability by creating interpersonal obligations and/or obligations towards the group itself to act collaboratively. Consequently, each member could drop its own intention and exit the group without violating obligations nor frustrating expectations and rights. Thus, one more ingredient is necessary: the *joint commitment* of the group. Indeed, joint commitment determines the degree to which a group persists in holding a joint intention, and therefore controls the likelihood of the group's re-considering and dropping the intention.

Against this background, we can now express our third assumption about agents' behaviour:

Assumption: (Joint commitment generation). If agent a_i recognizes a potential for cooperation with agent a_j with respect to its own intention to achieve φ, and is successful in generating a joint intention (jointly held by a_i and a_j) to achieve φ, then (as long as it keeps recognizing such a potential for cooperation) a_i will attempt to generate:

(a) a joint commitment (jointly held by a_i and a_j) to achieving φ; and

(b) a mutual belief (jointly held by a_i and a_j) that a_i and a_j have the joint ability to bring about φ; or, failing that, at least

(c) a mutual belief (jointly held by a_i and a_j) that a_i believes that a_i and a_j have the joint ability to achieve φ; and

(d) a mutual belief (jointly held by a_i and a_j) that a_i has the intention that a_i and a_j endorse the joint commitment to achieving φ.

Formally, we have:

$$\models \forall a_i, a_j, \forall t_i, t_j (t_i < t_j), \forall (r_i, r_j) RPfC(a_i, a_j, \varphi(t_j))(t_i) \wedge$$
$$\textit{J-INT}(\{a_i, a_j\}, \varphi(t_j))(t_i) \supset$$
$$\exists t_k (t_i < t_k < t_j), \exists e_i \text{ s.t. } (RPfC(a_i, a_j, \varphi(t_j)) \Leftrightarrow$$
$$Occurs[Attempt(a_i, e_i, (\textit{J-COMM}(\{a_i, a_j\}, \varphi(t_j)) \wedge \textit{M-BEL}(\{a_i, a_j\},$$
$$\textit{J-CAN}(\{a_i, a_j\}, \varphi(t_j)))), \textit{M-BEL}(\{a_i, a_j\}, (Bel(a_i, \textit{J-CAN}(\{a_i, a_j\}, \varphi(t_j))) \wedge$$
$$Int(a_i, \textit{J-COMM}(\{a_i, a_j\}, \varphi(t_j))))))])(t_i, t_k).$$

Condition (a) assures the persistence of the joint intention of a_i and a_j towards φ. Condition (b) means that a_i and a_j are mutually aware of having the joint ability required to attain φ. We justify this condition in the context of our model, since we assumed that a_i seeks the assistance of a_j with respect to its intention to achieve a state φ, because a_i believes that a_i and a_j together have the joint ability to attain φ (Section 5.3). However, in a more general situation, condition (b) is not necessary for a joint commitment to be established. Indeed, a group might mutually believe that φ will be eventually true without the group's being able to achieve φ directly. It might well be mutually believed that there is another agent who is not a member of the group, or even another group that has the required ability and whose assistance can eventually be asked for.

If this stage is successful, then a group with decision-making purposes will have been generated whose members will be jointly committed to giving an answer to a joint practical problem. In our approach, CDM is triggered by individual attitudes, but once the group generation stage has been successfully executed, the process will be additionally guided by higher-order mental attitudes. At this stage, these higher-order mental attitudes are reflected in the group's joint commitment which, in turn, is grounded on the notions of joint intention and mutual belief (Section 3.6).

6.4 Motivation for group generation

In [18] it has been pointed out that, without the mutual belief in a mutual dependence and in the necessity to collaborate in order for a given state of the world to be achieved, joint commitment is unmotivated. In our view, mutual dependence is but one potential motivation for forming a group. Within our working example, the two scientists share the same intention to find out the cause of the disease. They mutually believe that they have that intention. Each of them might intend the other to be part of the group. They might be jointly committed to finding out the cause of the disease. However, they might well not be mutually dependent on each other. Indeed, each scientist may have his or her own motivation to agree to be part of the group, to intend the other to be part of the same group as well, and to be committed to the group. And such motivations need not be the same.

In our more general framework, the agent soliciting assistance is socially dependent on some joint activity with the others with respect to the achievement of a given state of affairs. However, the agents whose help is being sought might agree to form a group (and intend to be members of the group) for reasons that may be unconnected with the original request for assistance. Some of the motivations may be the following [52]:

(a) the agents might decide to be benevolent and give their help;

(b) the agents might be enticed with a promise of a future reward;

(c) the agents might be threatened;

(d) the agents might be convinced that taking part to the group will enable the achievement of their own goals.

Agents might also be obliged to cooperate over the resolution of a practical problem and thus to share a joint intention. This holds if and when either the agents whose assistance is requested are obliged to cooperate regardless of the specific agent seeking assistance, or the agent requesting assistance has the authority to involve other agents in some form of cooperation. These *prescriptive sources of cooperation* can be captured by our notions of social mental shaping based on roles and social mental shaping occurring between agents. First, roles might function in a 'normative' way as they might influence the role-players to adopt an intention either in given circumstances or whenever requested to do so by other agents. In this sense, roles may entail sets of behavioural *obligations* based on the mental attitudes attached to them, and these obligations may function as prescriptive sources of collaborative activity [66, 85]. Second, social mental shaping occurring between agents might function as a prescriptive source of cooperation as long as one of the two agents has the authority to force the other to join a group and achieve some state in a collaborative manner. For example, if a_i is socially committed to a_j with respect to the achievement of φ, a_i might be obliged by a_j to get involved in a collaborative activity aimed at achieving φ. In this case, a_i may be subjected to a social mental shaping process exercised by a_j and based on a_j's authority to make a_i accept the request for assistance, and adopt the intention to become a member of a group and collaborate.

7 Social practical reasoning

Once a group has been generated that is jointly committed to achieving a state of the world, a joint practical basis and a practical problem will ensue (see Sections 5.1 and 5.2 for the individual case). On the one hand, the group's joint commitment to achieving a state constitutes the joint practical basis as it possesses practical force in raising and supporting the remaining stages of CDM. On the other, the group's joint commitment towards a state triggers a joint practical problem as the group members are confronted with a problem of what is to be done by the group to achieve that state [39]. Giving an answer to this problem means to jointly decide what course of action the group should perform to fulfil its joint commitment. Such a decision can be seen as an agreement on what is to be done. However, agreement cannot be reached until the agents *individually* reason about what action the group should perform to fulfil its joint commitment. Such reasoning undertaken by individual agents being located within a social setting is what we call *social practical reasoning*. As we shall see, social practical reasoning points to the logical arguments and cognitive mechanisms by which a group's joint commitment is transformed into a group member's individual intention that the group performs a particular action to fulfil its joint commitment.

In evaluating our conception of practical reasoning, we need to distinguish between two different uses of 'practical reasoning' [2]. Like the broader term 'reasoning', 'practical reasoning' may designate either a process or the corresponding structure. Generally speaking, the *process of practical reasoning* is a process of passing from appropriate premisses to a practical conclusion that is aimed at action rather than truth (Section 1). An agent wants to know what to do in some situation, and if its practical reasoning is successful, it becomes committed to some policy or course of action. A decision about what to do is actually a commitment to performing a *reasoned* action. That is, a commitment to performing an action that is explained and justified by a line of practical reasoning.

On the other hand, the *structure of practical reasoning* is a practical argument as a structure of propositions. To avoid confusion, we will use 'practical inference" for the structure of practical reasoning. A practical inference has been often characterized as a structure of propositions whose main constituents are: (a) a motivational premiss; (b) a connecting premiss that relates the content of the motivational premiss to action; and (c) a conclusion favouring the action specified by the connecting premiss [2, 3, 12, 24, 39]. This type of inference is a *means-end* argument, in which the first premiss mentions an end and the second premiss some means to this end.[22] The 'practical' conclusion which results from the premisses would consist in favouring the use of the means to secure the end. Thus, the study of this kind of practical inference is relevant to the problems of explaining and understanding purposive behaviour and conduct - both of individual agents and of groups of agents.

Before embarking on a discussion of social practical inferences, in what follows we will briefly discuss and formalize the main forms of practical inferences for the individual agent [24]. This allows us to highlight the basic structure of individual practical reasoning, and therefore the main changes to which this structure is subjected when we have to represent practical reasoning undertaken in a social setting. In our account, the process of practical reasoning in a social setting must be *socially embedded* in order to do what is required of it when performed by an agent who is a member of a group. Therefore, the corresponding structure of such reasoning will be expected to be couched in terms of some higher-order representation of mentalistic notions that can ensure such social orientation.

7.1 Taxonomy of basic means-end practical inference types

In this section we will focus on different schemas of practical inferences for merely individual practical reasoning. The philosopher and logician Charles Sandres Peirce (1839–1914) contended that induction, deduction, and abduction are three distinct types of inference, although as his views developed, he occasionally introduced hybrid forms such as 'abductive induction' [68]. We will follow Peirce's view, with three important modifications. First, in our view, practical inferences can be distinguished into deductive and non-deductive inferences. Second, we distinguish non-deductive practical inferences into abductions and inductions. Third, we treat practical inductions as special cases of practical abductions.

7.1.1 Deductive practical inferences

Deductive practical inferences can be regarded as instantiations of the following general pattern:

Major premiss: At time t_i, agent a_i intends to attain φ.

Minor premiss: At time t_i, agent a_i believes that the only way to attain φ is to perform action sequence e_i.

Conclusion: Therefore, at time t_i, agent a_i intends to perform e_i.

[22] In the present paper we shall not be dealing with the so-called 'rule-case syllogisms' [3, 24], that is, with those practical inferences that subsume an individual action under a general rule of action by the intermediary of a particular fact-stating premiss (e.g. (a) In all cases, if circumstances C hold, then do action e_i; (b) Circumstances C hold in this case; (c) Therefore, do action e_i in this case).

The first and second premisses of this form of practical inference are, respectively, an intention to achieve some state of the world and a practical judgement about the appropriate means to achieve that state. In turn, the practical judgement contains a *practical necessity*, namely a necessary relation between the end mentioned in the first premiss and the means to secure that end. Note that in this form of practical inference the conclusion is a *deductive consequence logically inferred* from the first two premisses. Deductive practical inferences support their conclusions in such a way that the conclusions must be true, given true premisses; they convey conclusive evidence.

7.1.2 Non-deductive practical inferences

In formulating the basic schema for non-deductive means-end practical inferences, we need to draw a distinction between a *decision* and a *choice*. Every choosing is a deciding, but not every deciding is a choosing. Generally speaking, to decide to do something is to make up one's mind to do it, where making up one's mind involves giving at least minimal consideration to the question of what you are going to do. As we conceive of it, explicitly choosing to do action sequence e_i is explicitly deciding to do e_i in preference to some alternative action sequence.[23]

Schematically, the line of practical inference leading to a decision that is a choice ideally includes steps like these:

1. At time t_i, agent a_i intends to attain φ.
2. At time t_i, agent a_i believes that performing action sequence e_i is a way to attain φ.
3. At time t_i, agent a_i believes that performing action sequence e_j is another way to attain φ.
4. At time t_i, agent a_i believes that it cannot perform both e_i and e_j simultaneously.
5. Shall agent a_i perform e_i or e_j?

   Conclusion: At time t_i, agent a_i intends to perform e_i.

In this form of practical inference, although the conclusion is clearly motivated by a set of premisses, it seems clear that it is not a deductive consequence logically inferred from any of the premisses the agent actually considered. The agent's conclusion seems, in fact, to be a logically 'free' step: it accords with its preferences, it is based on them, but *it is not deductively inferred* from anything at all. The conclusion transcends the information of its premisses and generates a commitment that is not based upon any practical necessity encoded there at all. This can be contrasted with practical deductive inferences, which can be thought of as extracting, explicitly in their conclusions, practical necessities that were already contained in the premisses. In short, while the conclusion of a deductive practical inference is an intention reflecting a decision that is not a choice, non-deductive practical inferences reflect a deciding that is also a preference-based choosing.

7.1.3 Forms of non-deductive practical inferences

Non-deductive means-end practical inferences can be distinguished into:

[23] Note that doing and not doing a given action can be considered alternative means only insofar as both bring about the same results. In particular, given the agent's intention to achieve φ and an action e_i that, if performed, can attain φ, if not doing e_i prevents the agent from attaining φ, then 'doing e_i' and 'not doing e_i' are not alternative means for achieving φ.

(a) abduction-based practical inferences; and

(b) induction-based practical inferences.

This taxonomy reflects the type of non-deductive inferential process the individual agent uses in selecting an action.

Practical abduction. Peirce maintained that there occurs in science as well as in everyday life a distinctive pattern of theoretical reasoning wherein explanatory hypotheses are formed and accepted [68]. He called this kind of theoretical reasoning 'abduction'. Along this view, we call practical abduction, or *practical inference to satisfactory action*, the structure corresponding to the reasoning process of passing from an intention to achieve a state of affairs to an intention to perform an action that can *satisfactorily* fulfil the prior intention.

This reasoning process includes the following steps:

(a) an initial process of coming up with a set of alternative actions that are plans for the agent to attain a state of the world;

(b) a process of critical evaluation wherein a practical judgement is made as to which action should be performed; and

(c) the agent's intention to perform that action.

More formally, the structure of this process is reflected by a practical abduction, which we take to be a distinctive kind of practical inference that adheres to the following pattern:

1. At time t_i, agent a_i intends to attain φ.
2. At time t_i, agent a_i believes that $ACT_{a_i} = \{e_i, e_j \ldots\}$ is a set of alternative action sequences that a_i can perform to attain φ.
3. At time t_i, agent a_i believes that e_i is a satisfactory means to achieve φ (i.e. a_i prefers to perform action sequence e_i rather than the other action sequences included in ACT_{a_i}).

   Conclusion: Therefore, at time t_i, agent a_i intends to perform e_i.

The core idea is that a prior intention to achieve a state of the world is transformed into a conclusive intention to perform an action that satisfactorily can achieve that state of the world. Therefore, an abductive practical inference aims at a *satisfactory* decision about what is to be done, one that can be confidently accepted.

Practical induction. Inductive practical inference can be treated as an instance of abductive practical inference. Specifically, an inductive practical inference is the structure corresponding to the reasoning process of passing from an intention to achieve a state of affairs to an intention to perform an action that, on the basis of the *observed performance*, can satisfactorily fulfil the prior intention. An inductive practical inference adheres to this general pattern:

1. At time t_i, agent a_i intends to achieve φ.
2. At time t_i, agent a_i believes that, at time $t_j < t_i$, action sequence e_i turned out to be a satisfactory means to attain φ.
3. At time t_i, agent a_i believes that action sequence e_i is still a plan to attain φ.
4. At time t_i, agent a_i believes that action sequence e_i is a satisfactory plan to attain φ.

   Conclusion: Therefore, at time t_i, agent a_i intends to perform e_i.

In contrast to the way that other types of abductive practical inferences look forward in terms of expectations about the future, inductive practical inferences are history-dependent inferences that look backward to experience. Therefore the latter can be called *backward-directed* inferences (as opposed to all the other abductive inferences that are *forward-directed* inferences). As we shall see (Section 7.2.2.2), backward-directed inferences play a central role within many forms of collectives (e.g., organizations), in which rule- and procedure-oriented inferences (i.e. history-dependent inferences) characterize most decisional processes and choice behaviour.

To say that inductive practical inferences can be analysed as special cases of abductive practical inferences has the following meaning. First, the second step of an inductive practical inference, saying that action sequence e_i is believed to have been a satisfactory means to achieve φ, can be thought of as a conjunction of three beliefs: (a) the belief that e_i was preferred to other alternative actions; (b) the belief that e_i, once performed, brought about φ; and (c) the belief that the past performance of e_i was not biased, that is, it was a representative outcome. Second, the fourth step of an inductive practical inference, saying that agent a_i believes that e_i is currently a satisfactory plan to attain φ, reflects the belief that past performance is an appropriate informative basis on which a practical judgement can be expressed. In this view, a practical induction is just a practical abduction in which a history-dependent preference function has been specified and used to form a conclusive intention. However, should action sequence e_i, once performed, fail to attain its target, or should agent a_i come to believe that past performance is no longer an accurate basis on which to select actions, then a_i may not pass from the past performance of e_i to the intention to perform e_i. Thus, as happens with practical abduction, practical induction is based upon a comparison between alternative means to attain a state. And this comparison is hidden in a history-dependent preference function according to which what turned out to be satisfactory in the past is also satisfactory when compared to current available alternatives.

7.2 Taxonomy of basic social means-end practical inference types

In the light of the above conception of individual practical inferences, this section concentrates on the main forms of means-end social practical inferences. Broadly speaking, a social practical inference may be defined as the structure of propositions that correspond to the reasoning process that a member of a jointly committed group undertakes in order to give an answer to a joint practical problem. A social practical inference has, minimally, three sorts of constituents: (a) a motivational premiss based on the group's joint commitment to achieving a state of affairs; (b) a connecting premiss that relates the content of the motivational premiss to a multi-agent action; and (c) a conclusion favouring the multi-agent action specified by the connecting premiss. As in the individual case, this type of inference can be seen as a *means-end* argument, since an end and some means to this end are mentioned in its premisses, whereas its conclusion would consist in favouring the use of the means to secure the end. More specifically, the conclusion of a social practical inference will support the performance of some multi-agent action that is a plan for the group to fulfil its joint commitment [83].

In our account, jointly reasoning about what is to be done inherently reflects the group members' taking an intentional stance towards one another and the group as a whole. That is, the process of a social practical reasoning emerges from each group member's reasoning about and representing the other members and the group itself in intentional terms, namely

as cognitive agents endowed with mental attitudes. Correspondingly, the *structure* of social practical reasoning — i.e. the social practical inference — will reflect both the group's and the individual agent's mental state, and therefore will be formalized in terms of both higher-order and individual mental attitudes.

In what follows, in developing our conception of social practical inference, we will distinguish, as with the individual case, between two main types of inferences: deductive and non-deductive inferences.

7.2.1 Social deductive practical inferences

For a given state φ, and an agent a_i that is a member of a group gr_i, the basic schema for *social deductive practical inference* is the following:

1. At time t_i, group gr_i is jointly committed to achieving φ; hence, a_i intends to achieve φ.
2. At time t_i, group member a_i believes that the only way for gr_i to achieve φ is to perform action sequence e_i.

   Conclusion: Therefore, at time t_i, agent a_i intends that gr_i performs e_i.

Formally, we have:

$$\models \forall gr_i, \forall a_i \in gr_i, \forall t_i, t_j (t_i < t_j), \forall e_i [\textit{J-COMM}(gr_i, \varphi(t_j))(t_i) \wedge$$
$$Bel(a_i, (\varphi(t_j) \Leftrightarrow \langle plan(gr_i, e_i, \varphi)\rangle(t_j)))(t_i)] \supset Int(a_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j))(t_i).$$

Clearly, here the intention that gr_i performs action sequence e_i is a *deductive consequence* logically inferred from the two premisses. That is, to the extent that agent a_i believes that there is only one action sequence e_i that the group can perform in order to achieve φ, the conclusion of a social practical inference can be seen as logically deduced from the premisses. In such a case, the premisses give evidential support to an intention based on the practical necessity that, unless e_i is performed by group gr_i, gr_i cannot achieve φ.

7.2.2 Social non-deductive practical inferences

Whenever a member of a jointly committed group believes that there is more than one action that the group can perform to fulfil its joint commitment, the conclusion of the social practical inference cannot be seen as logically deduced from the premisses. As happens with the individual case, whenever a choice is involved there is no question of the practical inference being formally valid. Preference-based social practical inferences are inferences that aim at satisfactory decisions about what is to be done. As in the individual case, the conclusions of social non-deductive inferences might well be warranted, justified, or rendered acceptable, although not deduced, by a set of premisses [24, 39, 45].

7.2.2.1 *Abduction-based social practical inferences*

For a given state φ, and an agent a_i that is a member of a group gr_i, the basic schema for abductive social practical inference is the following:

1. At time t_i, group gr_i is jointly committed to achieving φ; hence, a_i intends to achieve φ.
2. At time t_i, agent a_i believes that $ACT_{gr_i} = \{e_i, e_j, \ldots\}$ is a set of possible alternative action sequences that are plans for gr_i to attain φ.

3. At time t_i, agent a_i believes that e_i is a satisfactory plan for gr_i to achieve φ (i.e. a_i prefers that gr_i performs action sequence e_i rather than the other action sequences included in ACT_{gr_i}).

   Conclusion: Therefore, at time t_i, agent a_i intends that gr_i performs e_i.

Formally, we have:

$$\begin{array}{l}\models \forall gr_i, \forall a_i \in gr_i, \forall t_i, t_j (t_i < t_j), \forall e_i \\ [\textit{J-COMM}(gr_i, \varphi(t_j))(t_i) \wedge Bel(a_i, plan(gr_i, e_i, \varphi(t_j)))(t_i) \wedge \\ \bigwedge_{\{e_j | e_j \neq e_i\}} (Bel(a_i, plan(gr_i, e_j, \varphi(t_j))) \supset \\ Pref(a_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j), \langle plan(gr_i, e_j, \varphi)\rangle(t_j)))(t_i)] \supset \\ Int(a_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j))(t_i).\end{array}$$

A number of alternative criteria for selecting the satisfactory action have been developed (e.g., maximin criteria; regret criterion; expected monetary value; maximum expected utility) [58, 72]. However, in our schema, no such criterion is mentioned. This is because what is satisfactory is often determined by each agent's mental attitudes, and the circumstances in which the choice is being made. When, at step 3, we say that 'at time t_i, agent a_i believes that e_i is a satisfactory plan for gr_i to achieve φ', we are implicitly providing a straightforward way to generalize over specific choice criteria. Therefore, our schema reflects into a general preference function the rationale underlying a number of more specific choice criteria.

Finally, our conception of social abductive practical inferences captures one of the oldest behavioural speculations about decision-making in organizations: the idea that time and attention are scarce resources [60, 61]. Neither all alternative actions that can be performed to achieve a state of the world nor all the consequences of any of them can be known by the decision-maker [61]. Since only a few alternatives can be considered simultaneously, actions are selected less by choices among alternatives than by decisions with respect to search of new alternatives. However, there may be situations in which careful and intelligent social practical reasoning requires the generation of as many appropriate alternatives as possible. This might especially be called for when the cost of making a mistake is high, or when there is plenty of time to consider the alternatives.

7.2.2.2 *Induction-based social practical inferences*

Induction-based social practical inferences are special cases of abduction-based social practical inferences. For a given state φ, and an agent a_i that is a member of a group gr_i, the basic schema for *inductive social practical inferences* is the following:

1. At time t_i, group gr_i is jointly committed to achieving φ; hence, a_i intends to achieve φ.
2. At time t_i, agent a_i believes that, at time $t_j < t_i$, action sequence e_i turned out to be a satisfactory plan for gr_i to attain φ.
3. At time t_i, agent a_i believes that e_i is still a plan for gr_i to attain φ.
4. At time t_i, agent a_i believes that e_i is a satisfactory plan for gr_i to attain φ.

   Conclusion: Therefore, at time t_i, agent a_i intends that gr_i performs e_i.

Social inductive practical inferences are built on two core ideas of organizational decision-making. The first is that organizations tend to devote more attention to plans that fail to meet targets than they do to plans that meet targets [60, 61]. When a plan is successful, the search for new ones is reduced. When a plan fails, on the other hand, a search is undertaken for another one. This search for new alternatives continues until a satisfactory plan is discovered.

Thus, as is reflected in our conception of inductive practical inferences, alternative plans are not compared with each other so much as they are reviewed sequentially and accepted or rejected on the basis of their *observed past performance*.

The second core idea of organizational decision-making concerns the role of adaptive rules (Section 2). Much organizational choice behaviour involves rule-following more than calculation of consequences [28, 60, 61, 65]. Organizations have standard decisional procedures, some formally specified and some less formal. Procedures are followed because they have been learned as appropriate in a particular situation or as part of a particular role, rather than because they reflect a deeper comparison between several alternatives as a basis for a subsequent decision [60]. Awareness of the role that rule-following plays in organizational decision-making has directed attention to the processes by which rules and relatively stable organizational routines are created and changed [55]. The idea is that rules and routines encode experiential wisdom and reflect the lessons of history, in that they are the outcome of trial an error learning and the selection and retention of prior behaviour [28, 60].

To see organizations as prone to devote attention primarily to plans that fail to meet their targets, and as driven by rules, routines and procedures reflecting history and past experience, is to argue that an adequate account of CDM must characterize and describe the key social reasoning processes as inherently inductive. Indeed, we modelled social inductive inferences as reflecting history-dependent preferences based on past experience. That is, the agent's practical judgement about the satisfactory action (step 4) is based upon the evidence concerning the past execution of that action. Given this, to say that a social practical induction corresponds to a process of rule-following means that: (a) rules encode history-dependent preference functions; (b) there is a fit between the situation in which a rule is applied and the situation in which it has developed; and (c) rules are insensitive to changes in the mental state of rule-followers [60, 61].

So far, we have considered social deductive and non-deductive practical inferences that are related to the reasoning processes undertaken by the members of a jointly committed group. The conclusion of these processes is an intention favouring the performance of an action by the group. However, in order for the group to actually perform an action, the agents' practical reasoning processes must be 'socially connected'. This involves a *coordination problem* that is captured and formalized in the following fourth stage of our model.

8 Negotiation

Having displayed the basic types of social practical inferences, we need now to illustrate the process through which the members of a jointly committed group reach an *agreement* about what course of action the group should perform to fulfil its joint commitment. As was shown in Section 7, each agent will conduct a social practical reasoning process, either in a deductive or non-deductive form. A key role in this process is played by the agent's practical judgement, which is a belief concerning what action ought to be performed by the group in order for a given state of the world to be achieved. As a result, each agent will end up with an intention that the group performs the action specified by the practical judgement.

However, it might well be the case that each agent has a different view about what course of action the group should perform. In this case, each of the group members will have inconsistent intentions that the group performs differing actions. It is therefore necessary for the agents to come to some form of agreement about exactly which action the group will perform. There could be different forms of disagreement between the agents. On the one hand, the agents might disagree about the practical necessity upon which a deductive practi-

cal inference is based. On the other, the agents might disagree about the number and identity of the alternative actions that have to be considered before a choice can be made as well as about the choice criterion that is to be used to select the satisfactory alternative. In either case, however, an agreement about an action is called for if an answer to the joint practical problem is to be given.

A common assumption in the multi-agent system research community, and particularly in the multi-agent planning literature, is that a group of agents is endowed with an entire, pre-computed joint plan, which the group members will carry on executing until certain conditions arise [42, 71]. This is the idea of a joint plan as specifying a complete sequence of actions which need only be successfully performed to achieve some state. Once we look closely at the real-world behaviours of planning agents, however, it becomes clear that there is a rather complex interplay between the plan and the supporting social environment [52, 82]. This interplay goes well beyond the obvious fact that specific actions, once performed, may not have the desired effect and may thus require some re-thinking about how to achieve some specific states. In such cases, the original joint plan is still a complete, though fallible, specification of a route to success. In many cases, however, the joint plan turns out to be something more flexible, and much more dependent on a number of social actions and interactions between the agents. Individual agents deploy general strategies that incorporate social interactions with their acquaintances as an intrinsic part of a joint-plan generation process. At the outset, such a process can clearly involve explicitly formulated potential joint plans that are represented in the agents' mental states. But even in these cases, the plans function more like one ingredient of the whole generation process than complete recipes for success.

In the light of these observations, it seems that what is needed for describing an agreement generation about an action is a dual-faceted process that allows the agreed-upon action to result from both the individual agents' decision-making apparatus and a number of social actions and interactions between the agents. We believe that *negotiation* can capture these underpinning foundations of the process of jointly deciding what course of action a group should perform. Negotiation is the process by which an agreement is made by two or more parties [49]. Agents usually make proposals and counter-proposals; they might suggest modifications and receive requests of amendments of their own proposals; they might have objections to one or more of the alternative actions. Negotiation can range over a number of quantitative and qualitative aspects of actions. Each successful negotiation is therefore expected to resolve a number of different issues to the satisfaction of each agent. For example, a trade-off between contrasting issues might be required in order for the agents to come to an agreement [49].

In what follows, we will not develop a formalized analysis of negotiation. Following [88], we will simply focus on the key underpinning properties, structures, and processes that appear to be common to most forms of negotiation.

First, the minimum self-evident pre-condition required in order for negotiation to take place is that at least one agent has successfully conducted a social practical reasoning process. That is, negotiation cannot commence unless at least one agent:

(a) has come to form a practical judgement, that is, a belief that some action is either satisfactory or the only available action that can be performed by the group to achieve a given state of the world; and

(b) has come to the corresponding practical conclusion favouring the practical judgement; that is, the intention that the group performs the action suggested in the practical judgement.

Thus, against this background, a first assumption can be formulated that reflects the above pre-condition:

(Assumption: Minimum pre-condition of negotiation). Given a group of agents jointly committed to achieving a state of the world, there will eventually follow a state in which the agents will hold their commitments iff at least one of them maintains the intention that the group performs some action in order to fulfil its joint commitment.

Formally, we have:

$$\begin{array}{l}\models \forall gr_i, \forall t_i, t_j (t_i < t_j) \textit{J-COMM}(gr_i, \varphi(t_j))(t_i) \supset \\ \exists t_k (t_i < t_k < t_j) \text{ s.t. } [\textit{J-COMM}(gr_i, \varphi(t_j))(t_i, t_k) \Leftrightarrow \\ \exists a_i \in gr_i, \exists e_i \text{ s.t. } Int(a_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j))(t_k)].\end{array}$$

Informally, this assumption means that the agents will not keep their joint commitment for ever. They will eventually drop their commitment unless one of them comes up with a possible way of fulfilling it.

Once one of the agents has generated an intention that the group performs some action, negotiation commences when that agent attempts to generate a mutual belief within the group about the content of its own intention. That is, negotiation will be triggered by an agent's attempt to bring about a state where it is mutually believed by the group that at least one action is a candidate for being moved up to the agreed-upon action status. We can therefore write our second assumption in the following way:

(Assumption: Making intentions known). If a member a_i of a group jointly committed to attaining a state of the world holds the intention that the group performs action sequence e_i, then (as long as the group keeps its joint commitment and a_i its intention) there will follow a state in which a_i will attempt:

(a) to bring about a state where it is mutually believed in the group that a_i intends that the group performs e_i; or, failing this, at least

(b) to bring about a state where it is mutually believed in the group that a_i believes that e_i is either satisfactory or the only available action sequence that the group can perform to fulfil its joint commitment.

Formally, we have:

$$\begin{array}{l}\models \forall gr_i, \forall a_i \in gr_i, \forall t_i, t_j (t_i < t_j), \forall e_i \\ [\textit{J-COMM}(gr_i, \varphi(t_j)) \wedge Int(a_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j))](t_i) \supset \\ \exists t_k (t_i < t_k < t_j), \exists e_k \text{ s.t. } ([\textit{J-COMM}(gr_i, \varphi(t_j)) \wedge \\ Int(a_i \langle plan(tr_i, e_i, \varphi)\rangle(t_j))] \Leftrightarrow Occurs[Attempt(a_i, e_k, \\ \textit{M-BEL}(gr_i, Int(a_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j))), \textit{M-BEL}(gr_i, Bel(a_i, (\varphi(t_j) \Leftrightarrow \\ \langle plan(gr_i, e_i, \varphi)\rangle(t_j))) \vee (Bel(a_i, plan(gr_i, e_i, \varphi(t_j))) \wedge \\ \bigwedge_{\{e_j | e_j \neq e_i\}} (Bel(a_i, plan(gr_i, e_j, \varphi(t_j))) \supset Pref(a_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j), \\ \langle plan(gr_i, e_j, \varphi)\rangle(t_j)))))])(t_i, t_k).\end{array}$$

Informally, the above assumption means that the agents who have successfully conducted a social practical reasoning process will attempt to generate a mutual belief in the group about their practical conclusion, i.e. the intention that the group performs a particular action. Failing this, they will at least generate a mutual belief within the group about their practical judgements. These may be expressed in the form either of a practical necessity (in the case of

a deductive reasoning) or of a preference over alternative actions (in the case of non-deductive reasoning).

Finally, in order for negotiation to come to a conclusion, a social mental shaping process must be exercised aimed at generating an agreement about what course of action should be performed. On the one hand, if an agent intends that the group performs a particular action, then it will try to take the whole group closer to an agreement about performing that action. On the other, if an agent has an objection to some action, it will try to prevent an agreement about that action being performed by the group. Against this background, we can express our third assumption:

(Assumption: Attempt to influence). If a member a_i of a group jointly committed to bringing about a state of the world intends that the group performs action sequence e_i, and believes that another member a_j intends that the group acts differently, e.g. that it performs action sequence e_j, then (as long as the group keeps its joint commitment and a_i its intention and belief) there will follow a state in which a_i will attempt to exercise a social mental shaping process upon a_j aimed at:

(a) making a_j intend that the group performs e_i; or, failing that, at least

(b) making a_j intend that the group does not perform e_j.

Formally, we have:

$$\begin{array}{l}
\models \forall gr_i, \forall a_i, a_j \in gr_i, \forall e_i, e_j, \forall t_i, t_j (t_i < t_j)[\textit{J-COMM}(gr_i, \varphi(t_j)) \wedge \\
Int(a_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j)) \wedge Bel(a_i, Int(a_j, \langle plan(gr_i, e_j, \varphi)\rangle(t_j)))](t_i) \supset \\
\exists e_k, \exists (r_i, r_j), \exists t_k (t_i < t_k < t_j) \text{ s.t.} \\
([\textit{J-COMM}(gr_i, \varphi(t_j)) \wedge Int(a_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j)) \wedge \\
Bel(a_i, Int(a_j, \langle plan(gr_i, e_j, \varphi)\rangle(t_j)))] \Leftrightarrow \\
Occurs[Attempt(a_i, e_k, Infl(Int(a_j, \langle plan(gr_i, e_i, \varphi)\rangle(t_j)), rel(a_i, a_j, (r_i, r_j))), \\
Infl(Int(a_j, \neg\langle plan(gr_i, e_j, \varphi)\rangle(t_j)), rel(a_i, a_j, (r_i, r_j))))])(t_i, t_k).
\end{array}$$

Informally, this assumption means that not only will agents attempt to make their intentions known within the group. They will also try to exert a social influence process upon their acquaintances, aimed at changing their beliefs about how the group should act, and ultimately impacting upon their intentions.

If negotiation is successful, then the whole process of CDM will end up with an agreement about an action to be performed by the group. Such an agreement implies that the agents share the same intention that the group performs a given action. However, as we maintained in Section 3.6, sharing identical individual intentions entails no joint mentalistic notion supporting any form of collective behaviour. What is needed is the group's joint commitment to acting in a specified manner. Therefore, in compliance with our definition of joint commitments (Section 3.6), we say that an *agreement* reached by a group gr_i at time t_i about an action sequence e_i represents the outcome of a CDM process — that is, a joint decision — iff, at time t_i, gr_i has a joint persistent intention (i.e. a joint commitment) that action sequence e_i will eventually be performed. More formally, $\forall gr_i, \forall e_i, \forall t_i, t_j (t_i < t_j)$, we say that at time t_i group gr_i has made a joint decision to perform action sequence e_i that is a plan for gr_i to

achieve φ at t_j iff:

$$\begin{array}{l}
\textit{M-BEL}(gr_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j))(t_i) \wedge \\
\textit{J-INT}(gr_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j))(t_i) \wedge \\
\bigwedge_{\{a_i | a_i \in gr_i\}} [Comm(a_i, gr_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j)) \wedge \\
\textit{M-BEL}(gr_i, Comm(a_i, gr_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j)))](t_i) \wedge \\
\delta(t_i) \wedge \textit{M-BEL}(gr_i, \delta)(t_i),
\end{array}$$

where

$$\begin{array}{l}
\delta \equiv [\textit{J-INT}(gr_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j))(t_i, t_j) \vee \\
\exists t_k (t_i < t_k \leq t_j) \text{ s.t.} \\
((\textit{M-BEL}(gr_i, \neg\langle plan(gr_i, e_i, \varphi)\rangle(t_j)) \vee \exists a_i \in gr_i \text{ s.t.} \\
(\neg Comm(a_i, gr_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j)) \wedge \\
\textit{M-BEL}(gr_i, \neg Comm(a_i, gr_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j))))(t_k) \wedge \\
\forall t_h (t_i \leq t_h < t_k) \textit{J-INT}(gr_i, \langle plan(gr_i, e_i, \varphi)\rangle(t_j))(t_h))].
\end{array}$$

Our definition of agreement-based joint decision reflects the following properties (Section 3.6):

(a) in gr_i it is mutually believed that e_i will be performed by gr_i in order to bring about φ;

(b) gr_i has the joint intention to perform e_i in order to bring about φ;

(c) each member $a_i \in gr_i$ is socially committed to gr_i to fulfilling the joint intention;

(d) in gr_i it is mutually believed that each member $a_i \in gr_i$ is socially committed to gr_i to fulfilling the joint intention;

(e) it is true (and mutual belief in gr_i) that (b) will continue to hold until it is mutually believed in gr_i either that φ will not be brought about as a consequence of gr_i's performing e_i, or that at least one of the members no longer has the motivation to be part of the group and drops its commitment.

Should the group be successful in making a joint decision, the whole CDM process would conclude with a *transformation of commitments and intentions*. That is, a transformation of a joint commitment, and the corresponding joint intention, to achieving a state of affairs into a joint commitment, and the corresponding joint intention, to performing an action that is a plan for the group in order to bring about that state. As with the individual case, such a conclusive joint commitment/intention might or might not be followed by action. What constitutes the essence of CDM is this process of transformation of joint commitments/intentions. In our account, the outcome of CDM is simply a decision, that is, an answer to a joint practical problem. And a decision is a composite concept that inherently reflects an intention-based joint commitment generated through social practical reasoning processes.

9 Related work

The focus of our work was on the formalization of the decision-making process performed by a number of cognitive agents jointly committed to acting together. Therefore, the work shares common research issues with three main Artificial Intelligence (AI) areas: formal languages; formal models of mental attitudes; and models of decision-making within multi-agent systems. In the following subsections, we present related work in these areas and situate our research in the relevant literature.

9.1 The formal language

In this paper we have developed a many-sorted multi-modal first-order language that draws upon and extends the previous work of Bell and Huang [8, 9], Wooldridge and Jennings [88], and Cohen and Levesque [25]. We will now consider the relationship with this work.

Our logic most closely resembles that of Bell and Huang [8, 9]. Like their CA language, it is a many-sorted logic with explicit reference to time points and intervals. Time is taken to be composed of points and is assumed to be discrete and linear. Like Bell and Huang, we expressed the semantics of the preference operator through closest-worlds functions that avoid the counter-intuitive properties of simpler possible-worlds semantics for preferences. However, our logic extends their language in several respects. First, it contains terms denoting groups of agents, and provides a set-theoretic mechanism for relating agents and groups. Second, it contains terms denoting roles and social relationship types. Third, it contains modal operators denoting joint mental attitudes, and formalizes a set of operations to be performed upon them.

Like Cohen and Levesque's logic [25], our language has modalities for representing mental attitudes, and a mechanism for describing the structure and occurrence of complex actions. As with Bell and Huang, our logic extends Cohen and Levesque's language in that it gives a formal account of sociality, in terms both of groups of agents and of joint mental attitudes. Semantically, the logic is very similar to that of Cohen and Levesque in that both share a linear view of time. However, in their logic, time does not explicitly appear in propositions, whereas time is explicit in our logic. This expressibility of our system enables us to characterize different types of mental attitudes. For example, we can express goals towards propositions that will be true at some specific time in the future or intentions towards propositions with different time points.

Like Wooldridge and Jennings' language [88], our logic contains terms that express groups of agents and modalities that express joint mental attitudes. However, our logic is different from their language. 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.

9.2 The mental model

Various works in AI research support the idea that agents can be modelled in terms of their mental states [87]. One particularly common approach is to model agents as BDI systems [69]. The first difference to highlight when comparing different approaches is the varying usage of mental attitudes. Cohen and Levesque [25] and Wooldridge and Jennings [88] refer only to two primitive mental attitudes — i.e. beliefs and goals — and define all other attitudes in terms of these two. Rao and Georgeff [69] use three primitive attitudes: beliefs, goals, and intentions. In all these cases, the definitions are not sufficient for a suitably articulated description of the mental states and behaviour of the agents involved in a CDM process. We used a wider definition of attitudes; in particular, beliefs, desires, goals, intentions, and preferences. Our mental model strictly resembles that of Kraus *et al.* [52]: like their formalism, our logic has modalities for representing all those mental attitudes. Like in [52], desires may be inconsistent; goals and intentions are consistent, closed under consequences, and do not contradict beliefs. However, we have taken beliefs to be closed under consequences, whereas in [52] agents are not assumed to be omniscient. Moreover, in [52] every goal is also a desire: we do not have such an axiom, as in our framework an agent can adopt a goal that may well

not represent a desired state of affairs. Finally, like in [52], in our model there may be intentions that are not goals; however, in contrast to [52], we did not assume that an agent adopts all its goals as intentions. Other researchers [69] who did not model an agent's mental state within a social setting assumed that every intention is also a goal.

As described in Section 3.4.4, we distinguished between two types of intention, i.e. Intention-to and Intention-that. Many other formalizations, such as that of Cohen and Levesque [25], also make this distinction. For example, in [25] there are intentions with, respectively, an action expression and a proposition as their arguments. However, in [25] intending to bring about a state of the world means being committed to doing some sequence of actions after which that state holds. In contrast, in our model, the agent, once adopting an Intention-that, knows whether it is capable of fulfilling it in isolation or not. Should it be unable to act on its own, the agent will have to look for assistance from other agents. What we required is only that Intentions-that do not contradict beliefs, which means that the agent does not have intentions towards propositions the negations of which are believed. Our distinction between intentions is also consistent with that of Grosz and Kraus [42]. In [42], there are four types of intentions, Intentions-to and -that, and potential Intentions-to and -that. In their framework, potential intentions are used to represent an agent's mental state when it is considering whether to adopt an intention or not. In our framework, those mental attitudes that are potential candidates for being moved up to intention-status are formalized through the notion of goals. In contrast to [42], in our model an Intention-to does not commit an agent to practical reasoning. We have taken Intentions-that to play such a role, in that they induce the agent to look for the appropriate way to achieve the intended state of the world. Like in [42], our notion of Intention-that plays a key role in coordination problems. However, in [42] an Intention-that forms 'the basis for meshing sub-plans, helping one's collaborator, and coordinating status updates' (p. 282); in our approach, it forms the basis for socially connecting a number of agents who are jointly performing social practical reasoning processes. Indeed, we have taken Intentions-that to be the major premisses of both deductive and non-deductive social practical inferences.

In Section 3.5 we have formalized doxastic and motivational joint mental attitudes. Our account of mutual beliefs is similar to that of many other systems, such as that of Cohen and Levesque [27]. Like in [27], mutual beliefs are an infinite conjunction of beliefs about others' beliefs about others' beliefs (and so on to any depth) about some proposition. In contrast to [21], we did not model joint goals and joint intentions as first-class entities. Rather, we followed systems such as [27, 51, 71], in which joint mental attitudes clearly build upon the underpinning individual mental attitudes of the agents involved. However, in contrast to [27, 51, 71], we did not model joint goals/intentions as shared individual goals/intentions plus mutual beliefs. In Section 3.5 we explained why such a characterization is too weak to account for truly joint mental attitudes and we formalized them in terms of our proposed additional requirements.

Our notion of social commitment builds upon Castelfranchi's work [18]. We used this notion to formalize the higher-order concept of joint commitment. Like in [18], joint commitments reflect a web of social commitment relations between the agents and the group. However, in contrast to [18], we did not model joint commitments in terms of mutual dependence between the agents. Rather, we allowed for a variety of motivations (such as a disposition to help, authority, etc.) that may move agents into endorsing a social commitment towards others and hence generating a joint commitment to acting together. Furthermore, our notion of joint commitments is different from that of Cohen and Levesque [25, 27]. In [27],

joint commitments are particular forms of joint goals, and joint intentions are formalized as joint commitments of a group of agents sharing a particular mental state. In our model, we have joint intentions that may not be joint goals, and joint intentions that may not entail joint commitments. We formalized joint commitments as the strongest motivational attitudes of our framework. Finally, in our model the escape condition on which joint commitments are built is weaker and more flexible than in [25, 27]. According to [27], a group of agents will drop a joint commitment if they come to mutually believe that a given precondition is not true. Such a precondition, in their account, must be known by the agents from the beginning when they endorse the joint commitment. In contrast, in our formalization of joint commitments, we allowed for a weaker notion of escape condition, whereby the reasons for abandoning the joint commitment include, for example, the emergence of some new attitudes that are incompatible with the initial commitment.

9.3 Models of decision-making within multi-agent systems

Many aspects of CDM have been studied by researchers from a variety of disciplines, such as DAI [11, 27, 38, 40, 59, 82, 86, 87, 88], economics [10, 58], organizational behaviour theory [28, 55, 60, 61, 63], philosophy [14, 83, 84], and sociology [5, 17, 43, 81]. We can distinguish between two main categories of models of CDM: (a) high-level formal architectures for decision-making within a social setting; and (b) implementation architectures aimed at helping practitioners to realise software systems for managing coordination among a number of agents in real-world domains. The model developed in this paper lies in the former category, in that it aims at developing the theoretical foundations of CDM by using a formal language. However, there are a number of similarities also with the latter category. Indeed, like most implementation-oriented models, in our approach CDM is conceived of as constituted of a number of stages. In Section 2 we have identified three main phases that an adequate theory of CDM should account for: task announcement and orientation; evaluation; and negotiation [63].

Our characterization of CDM as a four-stage process strictly resembles that of Wooldridge and Jennings [88]. Like in [88], we have four stages that cover the whole process of CDM; we formalized CDM in an idealized world; we assumed that the four stages are not iterative. Furthermore, like in [88], we have a (first) stage concerning the recognition of a potential for cooperation and one (fourth) describing the negotiation process between agents. However, in contrast to [88], our model is more comprehensive, in that it captures the underpinning motivations and social processes of each stage. For example, our first stage — 'the practical starting-point' — starts from the very beginning of CDM, that is, the formulation of a practical problem that motivates at least one agent to seek assistance from others. Furthermore, we distinguished between a potential for cooperation which can be detected by one (or more) agent and the recognition of such a potential, which occurs whenever a social dependence relation exists between the agents involved. The most significant point of difference is that we formalized the social practical reasoning processes that agents undertake when they have to make a joint decision. Such processes are obscured in [88], and no account is given as to how a transformation of joint intentions/commitments occurs within a social setting that gives rise to the final joint decision. Moreover, we identified the key joint mental and motivational attitudes that guide a group generation, and we characterized the relationships between them. In [88] these attitudes are synthetically identified through the notion of Pre-Team that captures the agents' commitment to collective action if group generation is successful. Particularly,

in [88] Pre-Team expresses a mutual belief within a group that: (a) the group has the ability to achieve a state of affairs, and (b) each member has a commitment to achieving that state. In the second stage of our model, we formalized a step-by-step process of endorsement of a collective mental state, where identical individual intentions are strengthened by a joint intention, and this in turn is strengthened by a joint commitment. Finally, in [88] the last stage of CDM describes the joint performance of the agreed-upon action. In Section 4, we have explained why our model does not deal with action execution.

A number of approaches to cooperative activity have been developed which can be revisited so that we can evaluate how our model stands against them. For example, Bratman [14] outlines three main features of shared cooperative activity: mutual responsiveness; commitment to joint activity; and commitment to mutual support. Our model is consistent with this trio of aspects. First, our modelling of agents as reactive cognitive entities that act on the basis of their mental representations of other agents' mental attitudes accounts for some degree of responsiveness to the changes that occur within the social environment. Second, our characterization of groups in terms of higher-order doxastic and motivational attitudes allowed us to account for commitment both to joint activity and to mutual support.

One interesting area of investigation in DAI to which we can compare our model is multi-agent planning, particularly that work that has concentrated on multiple agents' mental attitudes for coordinating their activities [20, 42, 71]. For example, Grosz and Kraus [42] develop a formal model of collaborative plans and specify the mental states of the participants in a collaborative activity that handles complex actions. Like their approach, we provided the minimal mental state requirements that a group of agents must meet in order to continue to successfully perform collaborative activity. In [42] the focus is on collaborative plan definition, and details are given as to how collaborative activity rests eventually on the actions of the individual agents involved. Our focus was on the reasoning processes that individual agents undertake when jointly committed to acting in a collaborative manner. In [42] the formulation of full and partial SharedPlans is aimed at providing a definition of collaborative plans in which knowledge about how to act, ability to act and commitment to joint activity are distributed among group members. In contrast, we did not concentrate on details about the articulation of a collaborative plan for group action. Rather, the major goal of our work was to provide a clear conceptual framework in which a particular form of collaborative activity — i.e. CDM — can be evaluated in terms of its motivation, dynamics, mental state requirements, and social interaction and reasoning processes undertaken by the participants.

The idea that the agents involved in a CDM process perform a social practical reasoning process leads to Tuomela's definition of social practical inference [83]. Like in [83], our schema of social practical inference has as its first premiss an intention to achieve a state of the world. Moreover, like in [83], our schema contains a practical judgement expressing a belief about an appropriate means to achieve the intended state. However, there are a number of significant points of difference between our schema and Tuomela's. First, in [83] the agent that performs a social practical inference already knows the identity of an action that the group has decided to perform. Indeed, each agent believes that all the others will do their parts of that action, and it is mutually believed within the group that each agent holds such a belief [83, p. 217]. Furthermore, in [83] the agent concludes its practical inference by performing its part of the agreed-upon action. In contrast, in our model each agent reasons about the appropriate action and ends up with a practical judgement which does not reflect any form of agreement within the group. The conclusion of our social inference is the agent's intention that the group performs an action (that, in turn, is based on the agent's practical judgement).

In our model, it is this intention that leads the agent to interact with its acquaintances in order to reach an agreement about the action to be performed by the group. Finally, in [83] only deductive inferences are considered. We also accounted for non-deductive inferences where agents' preferences play a key role within the selection of a satisfactory action.

10 Concluding remarks

In this paper, we have presented an abstract formal model of decision-making in a social setting, and we have described all aspects of the process, from recognition of a potential for cooperation through to joint decision. In a multi-agent environment, where self-motivated autonomous agents try to pursue their own goals, a joint decision cannot be taken for granted. In order to decide effectively, agents need the ability to (a) represent and maintain a model of their own mental attitudes, (b) reason about other agents' mental attitudes, and (c) influence other agents' mental attitudes. Social mental shaping has been advocated as a general mechanism for attempting to have an impact on agents' mental states in order to increase their cooperativeness towards a joint decision. Our aim was to consider a number of issues that have hitherto been neglected in social sciences as well as in AI. For example, we have defined the conditions under which there is potential for initiating social interaction and such a potential can be recognized by agents. We presented a logical model of joint attitude generation, whereby identical individual attitudes have been grouped together to form joint attitudes, through to the strongest motivational attitude we have considered in this paper, i.e. joint commitment. In addition, we have formalized different types of social reasoning processes, both deductive and non-deductive. Particularly, we provided a clear conceptual model that allowed us to represent the structure of abductive and inductive reasoning processes undertaken for practical purposes within a distributed social environment.

In Section 2 we set out a number of properties that an adequate theory of CDM should exhibit. We will now briefly revisit those desiderata and evaluate whether our model is consistent with them.

1. Both individualistic and higher-order units of analysis and constructs are required

Our model builds upon both individualistic and social constructs. Our formal language contains terms denoting individual agents and groups of agents. The agent's local decision-making apparatus is extended with higher-order doxastic, motivational, and deontic attitudes — mutual beliefs, joint desires, joint goals, joint intentions, joint commitments — that capture the dynamics of inter-agent processes within a common social setting.

2. Agents are autonomous, reactive, and pro-active

First, autonomy is captured by our characterization of agents as entities that decide whether to interact or not. When requested to interact, they might either accept or reject such a request of assistance on the basis of their own motivations. In our model, autonomy is mainly captured by the notion of 'attempt'. Since agents are not required to cooperate, the agent that recognizes a potential for cooperation is expected to attempt to get other agents involved in a cooperative activity by influencing their mental states (stage 2). Moreover, since agents are not required to accept other agents' proposals, attempts are also advocated during negotiation when an agent has to interact with others in order to make its preferences known within the group. Second, reactivity is captured by our characterization of agents as cognitive entities, that is, entities endowed with a set of mental attitudes that represent the physical and social environment in which agents are located. These attitudes reflect any perceived change oc-

curring within the environment and will in turn regulate the agents' behaviour in a reactive manner. Third, proactiveness is captured by the agents' ability to initiate social processes whenever they deem it appropriate. Since agents take an intentional stance [30] towards the others, they will try to perform a variety of social actions and interactions on the basis of their beliefs about their acquaintances' mental attitudes and behaviour.

3. Agents are self- and other-interested

In our model agents are self-interested in that they attempt to satisfy their own needs and preferences. For example, when an agent is not able to fulfil an intention in isolation, it will seek assistance from other agents. Moreover, during negotiation, agents will try to make their preferences known within the group and to influence the group to perform their preferred plan. However, agents are also other-interested: they may decide to be benevolent or they may be moved into action by deontic attitudes such as social commitments. For example, when the group is jointly committed to making a decision, each member has a social commitment towards the group and its behaviour will be accordingly regulated by social normative constraints. Once jointly committed to making a decision, agents will also be mutually responsive (i.e. their behaviour is regulated by expectations about their acquaintances), committed to acting in a collaborative manner (i.e. they have a persistent joint intention to perform a joint activity that is necessary to reach a joint decision), and committed to mutual support (i.e. they have a joint persistent intention that the others do their part in getting the whole group close to a joint decision).

4. Communication is essential

Although we have not explicitly considered communicative actions, our notion of social mental shaping inherently builds upon the assumption of a number of agents communicating within a common social setting. Our model describes communicative actions in terms of their effect rather than of the means through which they take place. Indeed, an agent's mental state may be influenced by a number of interaction strategies (persuasion, threat, negotiation, etc.) each based upon different patterns of communication. Finally, our model is consistent with most current theories of communications [4, 26, 77]. For example, Cohen and Levesque [26] gave an account of illocutionary acts as attempts to bring about some mental attitudes in a conversation participant. This is consistent with our formalization, since in our model we used the notion of attempt at a number of points to characterize agents' actions aimed at influencing their acquaintances' mental states.

5. Cooperation can fail

In our model cooperation is not taken for granted. There are a variety of stages at which the CDM process may falter. Indeed, the overall formal structure of our model — mainly based on a number of assumptions — reflects the need to ensure that the whole process comes to a conclusion, as there is no *a priori* guarantee that each stage is always successful. For example, the agent that recognizes a potential for cooperation may be unable to exercise a social mental shaping process upon other agents. Or, some of the agents that are jointly committed to making a joint decision may come later to adopt an intention that turns out to be incompatible with the ongoing cooperative activity. In this situation the joint commitment to acting in a collaborative manner will be dropped by the group and no joint decision will be made by that group. Moreover, negotiation may fail to get the group closer to an agreement. Agents may fail to influence one another; in this case, no answer to the joint practical problem will be found.

6. Conflict is pervasive in CDM

Our model reflects some degree of conflict in a number of stages. Conflict may arise between two potentially cooperating agents when one of the two attempts to make the other adopt an intention that is incompatible with the latter's own intentions. Conflict permeates the negotiation process whereby each agent has its own preference and beliefs as to the action to be performed by the group and consequently attempts to make the group perform that action. In our account, the stability of a joint commitment rests on the group members' being committed to one another and to the group as a whole. Whenever any form of conflict arises between any pair of the involved agents so that at least one of the two drops its commitment to the common endeavour, then no joint commitment will exist any more within the group and CDM will consequently falter. Conflict is strictly related to the agents being autonomous and (partly) self-interested. Agents are not required to interact; they decide to do so on the basis of their differing preferences and needs. Moreover, once a cooperative activity has been established, agents' behaviour is (partly) regulated by their disposition to satisfy their differing preferences and needs. Hence, in order for cooperation to take place successfully, some degree of conflict among differing preferences and needs (goals, intentions, etc.) must be overcome.

7. CDM is a multi-stage process

Our model explicitly identifies four stages in which CDM can be decomposed. The first two stages — 'the practical starting-point' and 'group generation' — are mainly related to task announcement and orientation. They involve the recognition of a potential for cooperation and a number of social actions and interactions aimed at generating a group with decision-making purposes. The third stage — 'social practical reasoning' — can be described as an evaluation phase. Each group member will search for alternative courses of action that, if performed, can give an answer to the practical problem. These alternative actions are then evaluated until a practical judgement is formed, and an intention is generated that favours that practical judgement. The fourth stage deals with negotiation. Each group member will attempt to get the group closer to an agreement about the action to be performed. If this phase is successful, then a joint decision will be made and the group will be jointly committed to performing the agreed-upon action.

8. CDM may involve the following of rules and routines that adapt to experience, rather than anticipatory choice

The third stage of our model formalizes different structures of social practical reasoning. Among these, induction-based practical inferences play a key role. As maintained in Section 7.2.2.2, these inferences may be seen as reflecting rule-following processes and history-dependent preferences. Therefore, in our account, CDM is modelled as inherently encoding not only anticipatory calculations of consequences, but also standard procedures, relatively stable routines, conventions, and rules encoding experiential wisdom.

In summary, the main contributions of this paper are:

(a) We described and formalized the properties of decision-making within a social setting and linked these properties to the mental attitudes of the agents involved.

(b) We specified the minimal mental state requirements needed for the agents to continue to perform a CDM process.

(c) We combined the representation of a mental model with an account of sociality and interaction capabilities.

(d) Unlike previous work, we explicitly represented and reasoned about mechanisms for influencing other agents' mental attitudes and behaviour in interactions between autonomous agents.

(e) The reasoning about influencing was integrated in a multi-agent setting where it was used to reconcile conflicting mental attitudes and guide the agreement generation process.

There are a number of issues that we intend to address in future work. A refinement of our model is obviously needed, including a more detailed treatment of the process of recognizing a potential for cooperation, the process of group generation, and the process of negotiation. For example, we assumed that the agent that recognizes a potential for cooperation and attempts to generate a group with decision-making purposes is also the one that subsequently attempts to generate a joint intention and a joint commitment within the group. Perhaps a more realistic characterization of this stage should account for the possibility also of other agents' attempts to generate such joint mental attitudes. Furthermore, our model of CDM deals with a number of cognitive agents that are aware of being involved in a decision-making process. However, CDM may also be an emergent form of cooperation and, as such, it may involve unaware agents. A more comprehensive theory of CDM must therefore account also for such emergent social phenomena that do not necessarily entail the full representation in the minds of the participants of what is jointly being performed. We also envisage providing a more detailed account of the relations between different modalities. Particularly, our focus will be on the relations between individual and joint mental attitudes, and on the complex process by which the latter is formed over time in the course of the CDM process. In our model we have said nothing about the process through which a joint mental attitude is formed. Future work needs therefore to highlight the key steps of the process through which individual mental states mesh together until a joint mental attitude ensues. Most importantly, perhaps, future research will comparatively evaluate various persuasive arguments for dynamically influencing agents' mental attitudes (along the lines of [52, 67]). Verifying the effectiveness of different arguments under different conditions may lead to a set of criteria for selecting the most appropriate argument at any stage of CDM. This will also help practitioners to develop cooperative systems, by indicating the argument types that can be used for increasing the willingness of agents to cooperate and for getting the most out of the negotiation process.

Acknowledgements

The authors would like to thank the anonymous referees for their valuable and detailed comments on earlier versions of this paper.

The authors are also deeply thankful to Mrs Jane Spurr for her kind assistance in the editing process.

References

[1] E. M. Anscombe. *Intention*. Blackwell, Oxford, 1957.

[2] R. Audi. A theory of practical reasoning. *American Philosophical Quarterly*, **19**, 25–39, 1982.

[3] R. Audi. *Practical Reasoning*. Routledge, London and New York, 1989.

[4] K. Bach and R. M. Harnish. *Linguistic Communication and Speech Acts*. MIT Press, Cambridge, MA, 1979.

[5] S. B. Bacharach and E. J. Lawler. *Power and Politics in Organizations: The Social Psychology of Conflict, Coalitions, and Bargaining*. Jossey-Bass, San Francisco, CA, 1980.

[6] M. Barbuceanu. Coordinating agents by role-based social constraints and conversation plans. In *Proceedings of the 14th National Conference on Artificial Intelligence*, pp. 16–21, 1997.

[7] C. I. Barnard. *The Functions of the Executive*. Harvard University Press, Cambridge, MA, 1938.

[8] J. Bell. Changing attitudes. In *Intelligent Agents, Post- Proceedings of the ECAI-94 Workshop on Agent Theories, Architectures, and Languages*, M. Wooldridge and N. R. Jennings, eds. pp. 40–55. Springer, Berlin, 1995.

[9] J. Bell and Z. Huang. Dynamic goal hierarchies. In *Intelligent Agent Systems: Theoretical and Practical Issues*, L. Cavedon, A. Rao & W. Wobcke, eds. pp. 88–103. Springer-Verlag, Berlin, 1997.

[10] K. Binmore. Modeling rational players I & II. *Economics and Philosophy*, **3**, 9–55; **4**, 179–214, 1987.

[11] A. H. Bond and L. Gasser, eds. *Readings in Distributed Artificial Intelligence*. Kaufmann, San Mateo, CA, 1988.

[12] M. E. Bratman. *Intentions, Plans, and Practical Reasoning*. Harvard University Press, Cambridge, MA, 1987.

[13] M. E. Bratman. What is intention? In *Intentions in Communication*, P. R. Cohen, J. Morgan and M. E. Pollack, eds. pp. 15–31. MIT Press, Cambridge, MA, 1990.

[14] M. E. Bratman. Shared cooperative activity. *Philosophical Review*, **101**, 327–41, 1992.

[15] S. Cammarata, D. McArthur and R. Steeb. Strategies of cooperation in distributed problem solving. In *Proceedings of the International Joint Conference On Artificial Intelligence*, pp. 767–770, Karlsruhe, 1983.

[16] K.M. Carley. Knowledge acquisition as a social phenomenon. *Instructional Science*, **14**, 381–438, 1986.

[17] K.M. Carley. Group stability: A socio-cognitive approach. *Advances in Group Processes*, **7**, 1–44, 1990.

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

[19] C. Castelfranchi. Modelling social action for AI agents. *Artificial Intelligence*, **103**, 157–182, 1998.

[20] C. Castelfranchi and R. Falcone. Basic mental attitudes of a collaborating agent: Cognitive primitives for MAS. In *MAAMAW-99*, pp. 188–209. Springer-Verlag, Berlin, 1999.

[21] L. Cavedon, A. Rao and G. Tidhar. Social and individual commitment. In *Intelligent Agent Systems: Theoretical and Practical Issues*, L. Cavedon, A. Rao & W. Wobcke, eds. pp. 152–163. Springer-Verlag, Berlin, 1997.

[22] L. Cavedon and L. Sonenberg. On social commitment, roles and preferred goals. In *Proceedings of the Third International Conference on Multi-Agent Systems*, pp. 80–86, 1998.

[23] B. Chellas. *Modal Logic: An Introduction*. Cambridge University Press, Cambridge, MA, 1980.

[24] D. S. Clarke. *Practical Inferences*. Routledge and Kegan Paul, London, 1985.

[25] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, **42**, 213–61, 1990.

[26] P. R. Cohen and H. J. Levesque. Rational interaction as the basis for communication. In *Intentions in Communication*, P. R. Cohen, J. Morgan and M. E. Pollack, eds. pp. 221–256. MIT Press, Cambridge, MA, 1990.

[27] P. R. Cohen and H. J. Levesque. Teamwork. *Noûs*, **25**, 487–512, 1991.

[28] R. M. Cyert and J. G. March. *A Behavioral Theory of the Firm*. Prentice-Hall, Englewood Cliffs, NJ, 1963.

[29] D. Davidson. *Essays on Actions and Events*. Clarendon Press, Oxford, 1980.

[30] D. C. Dennet. *The Intentional Stance*. MIT Press, Cambridge, MA, 1987.

[31] B. Dunin-Keplicz, R. Verbrugge. Collective commitments. In *Proceedings of the Second International Conference on Multi-Agent Systems*, M. Tokoro, ed. pp. 56–63. Menlo Park, CA. 1996.

[32] E. H. Durfee. *Coordination of Distributed Problem Solvers*. Kluwer Academic Publishers, Boston, MA, 1988.

[33] E. H. Durfee, V. R. Lesser and D. D. Corkill. Trends in cooperative distributed problem solving. *IEEE Trans. On knowledge and Data Engineering*, **1**, 63–83, 1989.

[34] E. H. Durfee and T. A. Montgomery. Coordination as distributed search in a hierarchical behavior space. *IEEE Trans. On Systems Man and Cybernetics*, **21**, 1363–1378, 1991.

[35] R. Fagin, J. Y. Halpern, Y Moses and M. Y. Vardi. *Reasoning about Knowledge.* MIT Press, Cambridge, MA, 1995.

[36] S. Franklin and A. Graesser. Is it an agent, or just a program? A taxonomy for autonomous agents. In *Intelligent Agents III*, J. P. Mueller, M. J. Wooldridge, and N. R. Jennings, eds. pp. 21–35. Springer Verlag, Berlin, 1997.

[37] J. R. Galliers. A strategic framework for multi-agent co-operative dialogue. In *Proceedings of the 8th European Conference on Artificial Intelligence*, pp. 415–20. Pitman, London, 1988.

[38] L. Gasser. An overview of DAI. In *Distributed Artificial Intelligence: Theory and Praxis*, N. M. Avouris and L. Gasser, eds. pp. 9–30. Kluwer Academic Publishers, Boston, MA, 1992.

[39] D. P. Gauthier. *Practical Reasoning*. Oxford University Press, Oxford, 1963.

[40] M. P. Georgeff and F. F. Ingrand. Research on Procedural Reasoning Systems. Technical Report, AI Center, SRI International, Menlo Park, CA, Final Report, phase 1, 1988.

[41] N. Griffiths and M. Luck. Cooperative plan selection through trust, In *Multi-Agent Systems Engineering - Proceedings of the Ninth European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, F. J. Garijo and M. Boman, eds. pp. 162–174, **1647** of *Lecture Notes in Artificial Intelligence*, Springer-Verlag, Berlin, 1999.

[42] B. Grosz and S. Kraus. Collaborative plans for complex group actions. *Artificial Intelligence*, **86**, 269–357, 1996.

[43] R. Grunder. On the actions of social groups. *Inquiry*, **19**, 443–454, 1976.

[44] C. B. Handy. *Understanding Organisations*, 4th edn. Penguin Books, Harmondsworth, 1993.

[45] G. Harman. Practical reasoning. *Review of Metaphysics*, **29**, 431–463, 1976.

[46] J. Hintikka. *Knowledge and Belief*. Cornell University Press, Ithaca, NY, 1962.

[47] J. Hintikka. Semantics for propositional attitudes. In *Reference and Modality*, L. Linsky, ed. Oxford University Press, Oxford, 1972.

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

[49] N. R. Jennings, P. Faratin, M. J. Johnson, T. J. Norman, P. O'Brien and M. E. Wiegand. Agent-based business process management. *International Journal of Cooperative Information Systems*, **5**, 105–130, 1996.

[50] N. R. Jennings, K. Sycara and M. Wooldridge. A roadmap of agent research and development. *Autonomous Agents and Multi-Agent Systems*, **1**, 275–306, 1998.

[51] D. Kinny, M. Ljungberg, A. S. Rao, E. Sonenberg, G. Tidhar and E. Werner. Planned team activity. In *Artificial Social Systems - Selected Papers from the Fourth European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW- 92*, C. Castelfranchi and E. Werner, eds. pp. 226–256. Volume **830** of Lecture Notes in Artificial Intelligence, Springer-Verlag, Heidelberg, 1992.

[52] S. Kraus, K, Sycara and A. Evenchil. Reaching agreements through argumentation: A logical model and implementation. *Artificial Intelligence*, **104**, 1–69, 1998.

[53] V. R. Lesser and D. D. Corkill. The distributed vehicle monitoring testbed: A tool for investigating distributed problem solving networks. In *AI Magazine*, **4**(3), 15–33, 1983.

[54] H. J. Levesque, P. R. Cohen, and J. H. T. Nunes. On acting together. In *Proceedings of the Eighth National Conference on Artificial Intelligence, AAAI-90*, pp. 94–99, 1990.

[55] B. Levitt and J. G. March. Organisational learning. *Annual Review of Sociology*, **14**, 319–340, Palo alto, CA, Annual Reviews, 1988.

[56] D. K. Lewis. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA, 1969.

[57] L. Lindahl. *Position and Change: A Study in Law and Logic*. D. Reidel Publishing Company, Dordrecht, 1977.

[58] R. D. Luce and H. Raiffa. *Games and Decisions*. John Wiley, New York, NY, 1957.

[59] A. Lux and D. D. Steiner. Understanding cooperation: An agent's perspective. In *Proceedings of the First International Conference on Multi-Agent Systems*, pp. 261–268, 1995.

[60] J. G. March. Decisions in organizations and theories of choice. In *Perspectives on Organization Design and Behavior*, A. Van de Ven and W. Joyce, eds. pp. 205–244. Wiley Interscience, New York, NY, 1981.

[61] J. G. March and H. A. Simon. *Organisations*. Wiley, New York, NY, 1958.

[62] J. L. McClelland and D. E. Rumelhart. *Parallel Distributed Processing*. MIT Press, Cambridge, MA, 1986.

[63] R. D. Middlemist and M. A. Hitt. *Organisational Behaviour, Managerial Strategies for Performance*. West Publishing Company, 1988.

[64] R. C. Moore. A formal theory of knowledge and action. In *Readings in Planning*, J. F. Allen, J. Hendler and A. Tate, eds. pp. 480–519. Morgan Kaufmann Publishers, San Mateo, CA, 1990.

[65] R. Nelson and S. Winter. *An Evolutionary Theory of the Firm*. Harvard University Press, Cambridge, MA, 1982.

[66] P. Panzarasa, T. Norman and N. R. Jennings. Modeling sociality in the BDI framework. In *Intelligent Agent Technology: Systems, Methodologies, and Tools. Proceedings of the First Asian Pacific Conference on Intelligent Agent Technology (IAT-99)*, J. Liu and N. Zhong, eds. pp. 202–206. World Scientific Publishing, 1999.

[67] S. Parsons, C. Sierra and N. R. Jennings. Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, **8**, 261–292, 1998.

[68] C. S. Peirce. *Writings of Charles S. Peirce: A Chronological Edition*. **5**, 1884–1886, 1993.

[69] A. S. Rao and M. P. Georgeff. Modeling agents within a BDI architecture. In *Proc. of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR '91)*, R. Fikes and E. Sandewall, eds. pp. 473–484. Morgan Kaufmann, Cambridge, MA, 1991.

[70] A. S. Rao and M. P. Georgeff. Decision procedures for BDI logics. *Journal of Logic and Computation*, **8**, 293–342, 1998.

[71] A. S. Rao, M. P. Georgeff and E. A. Sonenberg. Social plans: A preliminary report. In *Decentralized AI - 3*, E. Werner and Y. Demazeau, eds. 55–77, Elsevier, Amsterdam, 1992.

[72] E. Rasmusen. *Games and Information: An Introduction to Game Theory*. Basil Blackwell, Oxford, 1990.

[73] J. R. Searle. *Intentionality: An Essay in Philosophy of Mind*. Cambridge University Press, Cambridge, MA, 1983.

[74] M. E. Sergot. Normative positions, In *Norms, Logics and Information Systems*, P. McNamara and H. Prakken, eds. IOS Press, 1998.

[75] J. Sichman. *Du Raisonnment Social Chez les Agents*. PhD thesis, Polytechnique - LAFORIA, Grenoble, 1995.

[76] M. P. Singh. Group ability and structure. In *Decentralized AI-2, Proceedings of the Second European Workshop on Modelling Autonomous Agents in a Multi- Agent World (MAAMAW-91)*, Y. Demazeau and J. P. Mueller, eds. pp. 57–76. Elsevier Science Publishers, Amsterdam, 1991.

[77] M. P. Singh. *Multiagent Systems: A Theoretical Framework for Intentions, Know-how, and Communications*. Springer-Verlag, Berlin, 1995.

[78] R. G. Smith. *A Framework for Distributed Problem Solving*. UMI Research Press, 1980.

[79] R. G. Smith and R. Davis. Frameworks for cooperation in distributed problem solving. In *IEEE Trans. On Systems Man and Cybernetics*, **11**, 61–70, 1981.

[80] R. Stalnaker. A theory of conditionals. *Studies in Logical Theory, American Philosophical Quarterly*, **2**, 98–122, 1968.

[81] A. Strauss. *Negotiations: Varieties, Contexts, Processes, and Social Order*. Jossey-Bass, San Francisco, CA, 1978.

[82] M. Tambe. Towards flexible teamwork. *Journal of Artificial Intelligence Research*, **7**, 83–24, 1997.

[83] R. Tuomela. *A Theory of Social Action*. Reidel, Boston, MA, 1984.

[84] R. Tuomela and K. Miller. We-intentions. *Philosophical Studies*, **53**, 115–137, 1988.

[85] E. Werner. Co-operating Agents: A unified theory of communication and social structure. In *Distributed Artificial Intelligence*, **II**, M. Huhns and L. Gasser, eds. pp. 3–36. Kaufman and Pitman, London, 1989.

[86] M. Wooldridge. A Knowledge-theoretic approach to distributed problem solving. In *ECAI 98. 13th European Conference on Artificial Intelligence*, John Wiley & Sons, Ltd., New York, NY, 1998.

[87] M. Wooldridge and N. R. Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, **10**, 115–152, 1995.

[88] M. Wooldridge and N. R. Jennings. Cooperative problem solving. *Journal of Logic and Computation*, **9**, 563–592, 1999.

[89] G. H. von Wright. *The Logic of Preference*. Edinburgh University Press, Edinburgh, 1963.

[90] G. H. von Wright. *Freedom and Determination*. North Holland Publishing Co., Amsterdam, 1980.

Appendix

A The formal framework: a complete definition

Here we detail the formal language used throughout the paper. First, a syntax will be developed, with an account of the symbols, terms, and well-formed formulae through which our model has been expressed. Second, we will present the semantics of the language, and the satisfaction rules of the formulae of the language will be defined.

Syntax

DEFINITION A.1

The language L contains the following symbols:

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), $Const_{Gr}$ (group constants), $Const_{Roles}$ (role constants), $Const_{RelTypes}$ (relationship type constants), $Const_E$ (action sequence 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_{Gr}, Var_{Roles}, $Var_{RelTypes}$, Var_E, 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 $Bel, Des, Goal, Int, Comm, Pref, In, Agts, \in, and =$;
5. the punctuation symbols ")", "(", "[", "]", and comma ",".

DEFINITION A.2
A term is either a constant or a variable. The sort of a term is either $Ag, Gr, T, I, Roles, RelTypes, E$ or O. The terms of sort agent, group, time point, interval, role, relationship type, action sequence and object (the sets $term_{Ag}, term_{Gr}, term_T, term_I, term_{Roles}, term_{RelTypes}, term_E$, 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 $S \in \{Ag, T, Gr, Roles, RelTypes, E, O\}$;
- $\{[u, u'] | u, u' \cup term_T\} \cup Var_I \subseteq term_I$

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 A.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 $e \in term_E$ and $i \in term_I$ then $Occurs(e)(i) \in wffs$.
- If $gr \in term_{Gr}, e \in term_E$ and $i \in term_I$ then $Agts(gr, e)(i) \in wffs$.
- If $a \in term_{Ag}, r \in term_{Roles}$ and $i \in term_I$ then $In(a, r)(i) \in wffs$.
- If $a \in term_{Ag}, gr \in term_{Gr}$ and $i \in term_I$ then $a \in gr(i) \in wffs$.
- If $a \in term_{Ag}, \varphi \in wffs$, and $i \in term_I$ then $Bel(a, \varphi)(i) \in wffs$.
- If $a \in term_{Ag}, \varphi \in wffs$, and $i \in term_I$ then $Des(a, \varphi)(i) \in wffs$.
- If $a \in term_{Ag}, \varphi \in wffs$, and $i \in term_I$ then $Goal(a, \varphi)(i) \in wffs$.
- If $a \in term_{Ag}, \varphi \in wffs$, and $i \in term_I$ then $Int(a, \varphi)(i) \in wffs$.
- If $gr, gr' \in term_{Gr}, e \in term_E$, and $i \in term_I$ then $Comm(gr, gr', e)(i) \in wffs$.
- If $a \in term_{Ag}, \varphi, \psi \in wffs$, and $i \in term_I$ then $Pref(a, \varphi, \psi)(i) \in wffs$.
- If $\psi, \chi \in wffs$ then $\neg\psi \in wffs$ and $(\psi \vee \chi) \in wffs$.
- If $S \in \{O, Ag, Gr, T, I, Roles, RelTypes, E\}, x \in Var_S$, and $\varphi \in wffs$ then $\exists x\varphi \in wffs$.

DEFINITION A.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.

DEFINITION A.5
We define the operators of Dynamic Logic in the following way ($i, i' \in term_I$):
- $Starts(i, i') \equiv min(i) = min(i) \wedge max(i) \leq max(i')$;
- $Ends(i, i') \equiv min(i) \geq min(i') \wedge max(i) = max(i')$;
- $Meets(i, i') \equiv max(i) + 1 = min(i')$;

- $Contains(i, i') \equiv min(i') \geq min(i) \wedge max(i') \leq max(i)$.

DEFINITION A.6
Complex actions are defined in the following way ($e, e' \in term_E$):
- $Occurs(e; e')(i) \equiv \exists i', i''(Starts(i', i) \wedge Meets(i', i'') \wedge Ends(i'', i) \wedge Occurs(e)(i') \wedge Occurs(e')(i''))$ (sequential action);
- $Occurs(e|e')(i) \equiv Occurs(e)(i) \vee Occurs(e')(i)$ (nondeterministic choice action);
- $Occurs(e||e')(i) \equiv Occurs(e)(i) \wedge Occurs(e')(i)$ (parallel action);
- $Occurs(\varphi)(i) \equiv \varphi(i)$ (test action).

Semantics

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. A group over D_{Ag} is a non-empty subset of D_{Ag}. The set of all such groups is D_{Gr}. Agents and groups can be related to one another via simple set theory. Agents have beliefs, desires, goals, and intentions. The beliefs of an agent are given by a *belief-accessibility relation* on W in the usual way. B maps agents, time and worlds to possible-worlds frames. For world w, agent a and time point t, the conditions on $W_{(Bel,a,t,w)}$ and $R_{(Bel,a,t,w)}$ capture the idea that $W_{(Bel,a,t,w)}$ is the set of $R_{(Bel,a,t,w)}$-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.

Agents have local preferences. An agent a prefers φ over ψ at time t in world w, if the value $p \in \mathbb{R}$ that a associates to the set of closest worlds to w in which φ is true and ψ is false, $cw(w, \varphi\neg\psi)$, is greater than the value $p' \in \mathbb{R}$ that a associates to the set of closest worlds to w in which ψ is true and φ is false, $cw(w, \psi \wedge \neg\varphi)$.

The set of all primitive action types is D_E. Every primitive action e is associated with an agent, given by $Agt(e)$. Finally, the world contains a set of objects, D_O, a set of roles D_{Roles}, and a set of relationship types, $D_{RelTypes}$.

DEFINITION A.7
The domain of quantification, D, is $D_{Ag} \cup D_T \cup D_I \cup D_{Gr} \cup D_{Roles} \cup D_{RelTypes} \cup (D*_E) \cup D_O$, where $D*_E$ denotes the set of non-empty sequences over E. If $n \in \mathbb{N}$, then the set of n-tuples over D is D^n.

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

DEFINITION A.8
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 A.9
A model M is a structure:
$\langle W, w_0, D_{Ag}, D_T, r_{DT}, D_I, D_{Gr}, D_{Roles}, D_{RelTypes}, D_E, D_O, Occurs, Agt, In, B, D, G, I, Comm, P_V, v, \Phi\rangle$,
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;
- $r_{DT} \subseteq D_T \times D_T$;
- $D_I = \{[t, t'] | t, t' \in D_T\}$ is a non-empty set of intervals;
- D_{Gr} is a non-empty set of groups;
- 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: $\{(r, r') | r, r' \in D_{Roles}\}$;
- D_E is a non-empty set of primitive action types;
- D_O is a non-empty set of objects;
- $Occurs \subseteq D*_E \times D_I \times W$;

- $Agt : D_E \rightarrow D_{Ag}$ gives the agent of each primitive action;
- $In \subseteq D_{Ag} \times D_{Roles} \times D_I \times W$;
- $B : D_{Ag} \times D_I \times W \rightarrow \mathcal{P}W \times \mathcal{P}(W \times W)$ 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 $W_{(Des,\alpha)}$;
- $G : D_{Ag} \times D_I \times W \rightarrow \mathcal{P}W \times (W \times W)$ 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. $R_{(Goal,\alpha)} \cap R_{(Bel,\alpha)} \neq \emptyset$;
- $I : D_{Ag} \times D_I \times W \rightarrow \mathcal{P}W \times \mathcal{P}(W \times W)$ 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. $R_{(Int,\alpha)} \cap R_{(Bel,\alpha)} \neq \emptyset$;
- $Comm \subseteq D_{Gr} \times D_{Gr} \times D *_E \times D_I \times W$;
- $P_V : D_{Ag} \times D_I \times \mathcal{P}W \rightarrow \mathbb{R}$ is such that for $w = \{w_0, w_1, \ldots, w_n\} \in W$ and $\alpha = (a, i, w), P_V\alpha = p \in \mathbb{R}$ is the value agent a associates with the set of worlds w in a given interval i;
- $v : Const \rightarrow D$ is an interpretation function for constants; and finally
- $\Phi : Pred \times W \rightarrow \cup_{n \in \mathbb{N}} D^n$ 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 A.10
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 $\tau \in Var$;
- $V_g(\tau) = [V_g(u), V_g(u')]$, for $\tau = [u, u'] \in term_I$.

The semantics of the language are defined via the satisfaction relation, '$\models$', 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 A.12.

DEFINITION A.11
A formula φ is true at a world w in M (written $M, w \models \varphi$) if φ is satisfied by all assignments g at w. If a formula φ is *valid* (satisfied by all interpretation structures), we write $\models \varphi$, as usual.

The rules defining the satisfaction relation make use of three additional functions. First, we denote by $[[\varphi]]^M_g$ the set of worlds in model M in which φ is satisfied by variable assignment g; i.e. $[[\varphi]]^M_g = \{w \in W | M, w, g \models \varphi\}$. We now define a function cw, of type $W \times \mathcal{P}(W) \rightarrow \mathcal{P}(W)$, where $cw(w, [[\varphi]]^M_g)$ is the set of closest worlds to w in which φ is true.

Second, a function is defined that returns all the primitive actions referred to in an action sequence:

$$\text{actions}\,((e_1, \ldots, e_n)) \equiv \{e_1, \ldots, e_n\}$$

Finally, a function is defined that returns the agents required for an action term:

$$agents(e) \equiv \{a | \exists e \in actions(V_g(e)) \text{ s.t. } Agt(e) = a\}, \text{ where } e \in term_E.$$

DEFINITION A.12
A variable assignment g satisfies a formula φ at a world w in a model $M = \langle W, w_0, D_{Ag}, D_T, r_{DT}, D_I, D_{Gr}, D_{Roles}, D_{RelTypes}, D_E, D_O, Occurs, Agt, In, B, D, G, I, Comm, P_V, v, \Phi\rangle$ written $M, w, g \models \varphi$, as follows:

$M, w, g \models$ 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	$V_g(u_1) = V_g(u_2)$
$M, w, g \models (a \in gr)(i)$	iff	$V_g(a) \in V_g(gr)$
$M, w, g \models Occurs(e)(i)$	iff	$(V_g(e), V_g(i), w) \in Occurs$
$M, w, g \models Agts(gr, e)(i)$	iff	$agents(e) = V_g(gr)$
$M, w, g \models In(a, r)(i)$	iff	$(V_g(a), V_g(r), V_g(i), w) \in In$
$M, w, g \models Bel(a, \varphi)(i)$	iff	$\alpha = (V_g(a), V_g(i), w)$, $B\alpha = (W_{(Bel,\alpha)}, R_{(Bel,\alpha)})$ and $M, w', g \models \varphi$ for all $(w, w') \in R_{(Bel,\alpha)}$
$M, w, g \models Des(a, \varphi)(i)$	iff	$\alpha = (V_g(a), V_g(i), w)$, $D\alpha = (W_{(Des,\alpha)}, R_{(Des,\alpha)})$ and $M, w', g \models \varphi$ for all $(w, w') \in R_{(Des,\alpha)}$
$M, w, g \models Goal(a, \varphi)(i)$	iff	$\alpha = (V_g(a), V_g(i), w)$, $G\alpha = (W_{(Goal,\alpha)}, R_{(Goal,\alpha)})$ and $M, w', g \models \varphi$ for all $(w, w') \in R_{(Goal,\alpha)}$
$M, w, g \models Int(a, \varphi)(i)$	iff	$\alpha = (V_g(a), V_g(i), w)$, $I\alpha = (W_{(Int,\alpha)}, R_{(Int,\alpha)})$ and $M, w', g \models \varphi$ for all $(w, w') \in R_{(Int,\alpha)}$
$M, w, g \models Comm(gr, gr', e)(i)$	iff	$(V_g(gr), V_g(gr'), V_g(e), V_g(i), w) \in Comm$
$M, w, g \models Pref(a, \varphi, \psi)(i)$	iff	$P_V(V_g(a), V_g(i), cw(w, [[\varphi \wedge \neg\psi]]_g^M)) > P_V(V_g(a), V_g(i), cw(w, [[\psi \wedge \neg\varphi]]_g^M)$
$M, w, g \models \neg\psi$	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.

Received 25 September 1999