Fundamenta Informaticae 51 (2002) 271–295
IOS Press

Collective Intentions

Barbara Dunin-Kęplicz*
Institute of Informatics, Warsaw University
Banacha 2, 02-097 Warsaw, Poland
and
Institute of Computer Science, Polish Academy of Sciences
Ordona 21, 01-237 Warsaw, Poland
e-mail: keplicz@mimuw.edu.pl

Rineke Verbrugge†
Department of Artificial Intelligence, University of Groningen
Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands
e-mail: rineke@ai.rug.nl

Abstract. In this paper the notion of *collective intention* in teams of agents involved in *cooperative problem solving* (CPS) in *multiagent systems* (MAS) is investigated. Starting from individual *intentions*, *goals*, and *beliefs* defining agents' local asocial motivational and informational attitudes, we arrive at an understanding of collective intention in cooperative teams. The presented definitions are rather strong, in particular a collective intention implies that all members *intend* for all others to share that intention. Thus a team is created on the basis of collective intention, and exists as long as this attitude between team members exists, after which the group may disintegrate. For this reason it is crucial that collective intention lasts long enough.

Collective intentions are formalized in a multi-modal logical framework. Completeness of this logic with respect to an appropriate class of Kripke models is proved. Two versions of collective intentions are discussed in the context of different situations. It is assumed that these definitions reflect solely vital aspects of motivational attitudes, leaving room for case-specific extensions. This makes the framework flexible and not overloaded. Together with individual and collective knowledge and belief, collective intention constitutes a basis for preparing a plan, reflected in the strongest attitude, i.e., in collective commitment, defined and investigated in our other papers.

1. Introduction

In *multiagent systems* (MAS) one of the central issues is the study of how groups work, and how the technology enhancing group interaction can be implemented. From the distributed Artificial Intelligence perspective, multiagent systems are computational systems in which a collection of loosely-coupled

*Address for correspondence: Institute of Informatics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland
†Address for correspondence: Department of Artificial Intelligence, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands

autonomous agents interact in order to solve a given problem. As this problem is usually beyond the agents' individual capabilities, agents exploit their ability to *communicate*, *cooperate*, *coordinate*, and *negotiate* with one another. Apparently, the type of social interactions involved depends on circumstances and may vary from altruistic cooperation through to open conflict. A paradigmatic example of joint activity is *cooperative problem solving* (CPS) in which a group of autonomous agents choose to work together, both in advancement of their own goals as well as for the good of the system as a whole.

Some MAS are referred to as *intentional systems*. In such systems, in order to give a representation of the mental states and cognitive processes involved in a multiagent system, agents are represented as maintaining an intentional stance towards their environment. Such systems realize the *practical reasoning* paradigm ([2]) – the process of deciding, moment by moment, which action to perform in the furtherance of our goals. The best known and most influential are *belief-desire-intention systems*. BDI-agents are characterized by a "mental state" described in terms of *beliefs*, corresponding to the information the agent has about the environment; *desires*, representing options available to the agent, i.e. different states of affairs that the agent may choose to commit to; and *intentions* representing the chosen options. Ultimately, in our approach, intentions are viewed as an inspiration for a goal-directed activity, reflected in commitments, while beliefs are viewed as the agent's *informational* attitudes, desires or goals, intentions, and commitments refer to its *motivational* attitudes. When considering these collective notions, a concept of a group of agents is essential. In [35] a group is defined in the following way:

> A *group* or multiagent system is a system of agents that are somehow constrained in their mutual interactions. Typically, these constraints arise because the agents play different *roles* in the group, and their roles impose requirements on how they are to behave and interact with others.

This paper is concerned with a specific kind of group, namely a team, defined in [35] as follows:

> A *team* is a group in which the agents are restricted to having a common goal of some sort. Typically, team-members cooperate and assist each other in achieving their common goal.

Collective intention, as a specific joint mental attitude, is the central topic addressed in *teamwork*. We agree with [26] that:

> Joint intention by a team does not consist merely of simultaneous and coordinated individual actions; to act together, a team must be aware of and care about the status of the group effort as a whole.

In fact, we assume that a team is constituted as soon as a collective intention is present among the members, and stays together as long as the collective intention persists.

We agree with Bratman that in human practical reasoning, intentions are first class citizens, in the sense that they are not reducible to beliefs and desires [2]. They form a rather special consistent subset of an agent's goals, that the agent wants to focus on for the time being. Thus they create a screen of admissibility for the agent's further, possibly long-term, deliberation. In contrast to [7] we are interested in generic characteristics of intentions, resigning from classifying them further along different dimensions. The most crucial aspect of intentions is that they are considered as an inspiration for goal-directed

activity, reflected in the strongest motivational attitudes, that is in social (bilateral) and collective commitments. In our view, social commitments are related to individual actions, while collective commitments are related to plan-based team actions. The essential characteristics of both types of commitments is that they directly lead to action execution. As commitments are not the subject of this paper, we refer the reader to our other papers concerning commitments, e.g. [12].

In our work on collective intentions, the objective is two-fold. First, to formally characterize what it means for a team to have a collective intention towards a common goal (represented for example as a state of the world to be achieved). This characterization will be done using multi-modal logic, in the form of a *static* theory, comprising a descriptive view on collective intentions.

Secondly, let us remember that a team of agents in a multiagent system operates in a dynamic and often unpredictable environment. Such an environment poses the problem that team members may fail to bring their tasks to a good end or new opportunities may appear. This *reconfiguration problem*, was treated in [13], where collective intentions are maintained by properly adapting collective commitments to the changing circumstances. This contributes to the *dynamic*, more prescriptive theory of collective intentions. The formal specification of situations in which agents' attitudes change is shortly introduced in [15], and will be extensively discussed in a forthcoming paper summing up our theory of CPS.

The present paper falls squarely in the scope of the first objective to define a static theory of collective intentions. Two different definitions of collective intentions are presented, together with examples of situations in which they apply. Both notions are formalized in a multi-modal logical framework. Completeness of this logic with respect to an appropriate class of Kripke models is proved. Thus, a Computational Logic framework for specifying MAS involved in CPS is provided. The presented system is known to be EXPTIME-complete, so in general it is not feasible to give automated proofs of desired properties; at least there is no single algorithm that performs well on all inputs. As with other modal logics, the better option would be to develop a variety of different algorithms and heuristics, each performing well on a limited class of inputs. For example, it is known that restricting the number of propositional atoms to be used or the depth of modal nesting may reduce the complexity (cf. [22, 24, 18, 34]). Also, when considering specific applications it is possible to reduce some of the infinitary character of collective beliefs and intentions to more manageable proportions (cf. [16, Ch. 11]).

In this paper we leave out temporal considerations. Our full theory is, however, based on Kripke models including a temporal order. There are different possible choices of temporal ontology, for example between linear time, as in Cohen and Levesque's work [7], and branching time as in Rao and Georgeff's work [28] and in [15]. The definitions of collective intentions in terms of more basic attitudes, as presented in this paper, may be combined with either choice, depending on the application.

The rest of this paper is structured in the following manner. Section 2 provides some background about the role of intentions in practical reasoning. Section 3 gives a short logical background, a reminder of the epistemic theory we use, and a description of the logical theory of individual goals and intentions. The heart of the paper is formed by sections 4 and 5, in which collective intentions are investigated and provided with a completeness proof. Sections 6 and 7 close off with a discussion of the results and some ideas for future research.

2. The role of intentions in practical reasoning

Practical reasoning is the form of reasoning that is aimed at conduct rather than knowledge. The cycle of this reasoning involves:

1. repeatedly updating beliefs about the environment
2. deciding what options are available
3. "filtering" these options to determine new intentions
4. creating commitments on the basis of intentions
5. performing actions in accordance with commitments.

Practical reasoning involves two important processes: deciding *what* goals need to be achieved, and then *how* to achieve them. The former process is known as *deliberation*, the latter as *means-end-reasoning*. A key concept in the theory of practical reasoning is that of *intention*. In [17] it is characterized from the psychological viewpoint:

> "The concept of intention was (and still is) one of the most controversial in the history of psychology. Certain people – the eliminativists – purely and simply refuse to introduce this concept into their theories, claiming not only that it is useless, but also that it mindlessly confuses the issues. Others, in contrast, think of it as one of the essential concepts of psychology and that it should be given a central role, for it constitutes a keystone of the explanation of human behaviour in terms of mental states. Finally, the psycho-analytical school sees it as merely a vague concept which is handy in certain cases, but which should generally be replaced by desire and drives, which alone are capable of taking account of the overall behaviour of the human being in his or her aspirations and suffering."

We do not aim to present a psychologically sound theory of motivations driving human behaviour. We are interested in studying those motivational aspects that are involved in the *rational* decision making process. Thus, we disregard irrational drives and desires, that make human behaviour difficult to interpret and to predict. We do not consider any specific notion of *rationality*, in particular the economic one used in game theory. The only assumption made here is that agents are logical reasoners.

There is common agreement that intentions play a number of important roles in practical reasoning [2, 7]:

- I1 *Intentions drive means-end-reasoning.*
- I2 *Intentions constrain future deliberation.*
- I3 *Intentions persist.*
- I4 *Intentions influence beliefs upon which future practical reasoning is based.*

$\mathrm{BEL}(a, \varphi)$	agent a has the belief that φ
$\text{E-BEL}_G(\varphi)$	every agent in group G has the belief that φ
$\text{C-BEL}_G(\varphi)$	group G has the collective belief that φ
$\mathrm{GOAL}(a, \varphi)$	agent a has as a goal that φ be true
$\mathrm{INT}(a, \varphi)$	agent a has the intention to make φ true
$\text{E-INT}_G(\varphi)$	every agent in group G has the individual intention to make φ true
$\text{M-INT}_G(\varphi)$	group G has the mutual intention to make φ true
$\text{C-INT}_G(\varphi)$	group G has the collective intention to make φ true

Table 1. Formulas and their intended meaning

A key problem in the design of BDI-agents is how to achieve a good balance between these different concerns. It becomes especially important when an agent needs to drop some of its intentions. This happens for many different reasons: because the intentions will never be achieved, they are achieved already or there are no longer reasons supporting them. Thus, from time to time an agent's intentions should be reconsidered. This leads to the problem of balancing *pro-active*, (i.e. goal-directed) and *reactive* (i.e. event-driven) behaviour. We try to maintain this balance very carefully on both the individual and the collective level. The problem of persistence of intentions is expressed in an agent's *intention strategies*, addressing the question: *when and how can an agent responsibly drop its intentions?* One answer to this question is discussed in [28]. The collective level is apparently more complex. Collective intention helps the team to monitor its behaviour during teamwork: even if some members drop their individual intentions, the team replans, aiming that the collective intention is ultimately realized. The precise description of this reconfiguration process can be found in [12, 13], where an agent's pro-activeness and reactiveness are implicitly or explicitly involved in consecutive stages of the reconfiguration algorithm.

3. Preliminaries

As mentioned before, we propose the use of multi-modal logics to formalize agents' informational and motivational attitudes as well as actions they perform. In the present paper, where we restrict ourselves to the static aspects of the agents' mental states, we only present axioms relating attitudes of agents with respect to *propositions*, not actions. A proposition reflects a particular state of affairs.

Table 1 gives the formulas appearing in this paper, together with their intended meanings. The symbol φ denotes a proposition.

3.1. The logical language

Formulas are defined with respect to a fixed finite set of agents. The basis of the inductive definition is given in the following definition.

Definition 3.1. (Language)
The language is based on the following two sets:

- a numerable set $\mathcal{P}$ of *propositional symbols*;
- a finite set $\mathcal{A}$ of *agents*, denoted by numerals $1, 2, \ldots, n$.

Definition 3.2. (Formulas)
We inductively define a set $\mathcal{L}$ of formulas as follows.

F1 each atomic proposition $p \in \mathcal{P}$ is a formula;

F2 if φ and ψ are formulas, then so are $\neg\varphi$ and $\varphi \wedge \psi$;

F4 if φ is a formula, $i \in \mathcal{A}$, and $G \subseteq \mathcal{A}$, then the following are formulas:

epistemic modalities $\mathrm{BEL}(i,\varphi)$, $\text{E-BEL}_G(\varphi)$, $\text{C-BEL}_G(\varphi)$;

motivational modalities $\mathrm{GOAL}(i,\varphi)$, $\mathrm{INT}(i,\varphi)$, $\text{E-INT}_G(\varphi)$, $\text{M-INT}_G(\varphi)$, $\text{M-INT}'_G(\varphi)$, $\text{C-INT}_G(\varphi)$, $\text{C-INT}'_G(\varphi)$.

The constructs $\top$, $\bot$, $\vee$, $\rightarrow$ and $\leftrightarrow$ are defined in the usual way.

3.2. Kripke models

Each Kripke model for the language $\mathcal{L}$ consists of a set of worlds, a set of accessibility relations between worlds, and a valuation of the prepositional atoms, as follows.

Definition 3.3. (Kripke model)
A Kripke model is a tuple
$\mathcal{M} = (W, \{B_i : i \in \mathcal{A}\}, \{G_i : i \in \mathcal{A}\}, \{I_i : i \in \mathcal{A}\}, \mathit{Val})$, such that

1. W is a set of possible worlds, or states;
2. For all $i \in \mathcal{A}$, it holds that $B_i, G_i, I_i \subseteq W \times W$. They stand for the accessibility relations for each agent with respect to *beliefs*, *goals*, and *intentions*, respectively. For example, $(s,t) \in B_i$ means that t is an epistemic alternative for agent i in state s.
3. $\mathit{Val} : \mathcal{P} \times W \rightarrow \{0,1\}$ is the function that assigns the truth values to atomic propositions in states.

At this stage, it is possible to define the truth conditions pertaining to the language $\mathcal{L}$, as far as the propositional connectives and individual modal operators are concerned. The expression $\mathcal{M}, s \models \varphi$ is read as "formula φ is satisfied by world s in structure $\mathcal{M}$".

Definition 3.4. (Truth definition)

- $\mathcal{M}, s \models p \Leftrightarrow \mathit{Val}(p,s) = 1$;
- $\mathcal{M}, s \models \neg\varphi \Leftrightarrow \mathcal{M}, s \not\models \varphi$;
- $\mathcal{M}, s \models \varphi \wedge \psi \Leftrightarrow \mathcal{M}, s \models \varphi$ and $\mathcal{M}, s \models \psi$;
- $\mathcal{M}, s \models \mathrm{BEL}(i,\varphi)$ iff $\mathcal{M}, t \models \varphi$ for all t such that sB_it;
- $\mathcal{M}, s \models \mathrm{GOAL}(i,\varphi)$ iff $\mathcal{M}, t \models \varphi$ for all t such that sG_it;
- $\mathcal{M}, s \models \mathrm{INT}(i,\varphi)$ iff $\mathcal{M}, t \models \varphi$ for all t such that sI_it.

3.3. Axioms for beliefs

To represent beliefs, we adopt the standard $KD45_n$-system for n agents as explained in [16], containing the following axioms and rules for $i = 1, \ldots, n$:

A1 All instantiations of tautologies of the propositional calculus

A2 $\mathrm{BEL}(i,\varphi) \wedge \mathrm{BEL}(i,\varphi \rightarrow \psi) \rightarrow \mathrm{BEL}(i,\psi)$ (Belief Distribution)

A4 $\mathrm{BEL}(i,\varphi) \rightarrow \mathrm{BEL}(i,\mathrm{BEL}(i,\varphi))$ (Positive Introspection)

A5 $\neg\mathrm{BEL}(i,\varphi) \rightarrow \mathrm{BEL}(i,\neg\mathrm{BEL}(i,\varphi))$ (Negative Introspection)

A6 $\neg\mathrm{BEL}(i,\bot)$ (Belief Consistency)

R1 From φ and $\varphi \rightarrow \psi$ infer ψ (Modus Ponens)

R2 From φ infer $\mathrm{BEL}(i,\varphi)$ (Belief Generalization)

Note that, in the semantics, the accessibility relations B_i need not be reflexive, corresponding to the fact that an agent's beliefs need not be true. On the other hand, the accessibility relations B_i are transitive, euclidean and serial. These conditions correspond to the axioms of positive and negative introspection and to the fact the agent has no inconsistent beliefs, respectively. It has been proved that $KD45_n$ is sound and complete with respect to these semantics.

One can define modal operators for group belief. The formula $\text{E-BEL}_G(\varphi)$ is meant to stand for "every agent in group G believes φ". It is defined semantically as $\mathcal{M}, s \models \text{E-BEL}_G(\varphi)$ iff for all $i \in G$, $\mathcal{M}, s \models \mathrm{BEL}(i,\varphi)$), which corresponds to the following axiom:

C1 $\text{E-BEL}_G(\varphi) \leftrightarrow \bigwedge_{i \in G} \mathrm{BEL}(i,\varphi)$

A traditional way of lifting single-agent concepts to multiagent ones is through the use of *collective belief* $\text{C-BEL}_G(\varphi)$. This rather strong operator is similar to the more usual one of common knowledge, except that a collective belief among a group that φ need not imply that φ is true.

$\text{C-BEL}_G(\varphi)$ is meant to be true if everyone in G believes φ, everyone in G believes that everyone in G believes φ, etc. Let $\text{E-BEL}^1_G(\varphi)$ be an abbreviation for $\text{E-BEL}_G(\varphi)$, and let $\text{E-BEL}^{k+1}_G(\varphi)$ for $k \geq 1$ be an abbreviation of $\text{E-BEL}_G(\text{E-BEL}^k_G(\varphi))$. Thus we have $\mathcal{M}, s \models \text{C-BEL}_G(\varphi)$ iff $\mathcal{M}, s \models \text{E-BEL}^k_G(\varphi)$ for all $k \geq 1$. Define t to be G_B-*reachable* from s if there is a path in the Kripke model from s to t along accessibility arrows B_i that are associated with members i of G. Then the following property holds (see [16]):

$$\mathcal{M}, s \models \text{C-BEL}_G(\varphi) \text{ iff } \mathcal{M}, t \models \varphi \text{ for all } t \text{ that are } G_B\text{-reachable from } s.$$

Using this property, it can be shown that the following axiom and rule can be soundly added to the union of $KD45_n$ and **C1**:

C2 $\text{C-BEL}_G(\varphi) \leftrightarrow \text{E-BEL}_G(\varphi \wedge \text{C-BEL}_G(\varphi))$

RC1 From $\varphi \rightarrow \text{E-BEL}_G(\psi \wedge \varphi)$ infer $\varphi \rightarrow \text{C-BEL}_G(\psi)$ (Induction Rule)

The resulting system is called $KD45_n^C$, and it is sound and complete with respect to Kripke models where all n accessibility relations are transitive, serial and euclidean [16].

In the sequel, we will use the following standard properties of C-BEL$_G$ (see for example [16, exercise 3.11]).

Lemma 3.1. Let $G \subseteq \{1, \ldots, n\}$ be given. Then the following hold for all formulas φ, ψ:

- C-BEL$_G(\varphi \wedge \psi) \leftrightarrow$ C-BEL$_G(\varphi) \wedge$ C-BEL$_G(\psi)$
- C-BEL$_G(\varphi) \rightarrow$ C-BEL$_G$(C-BEL$_G(\varphi)$)

A problem with standard modal logics for beliefs and knowledge is that agents are formalized as being logically omniscient: they believe all theorems, as well as all logical consequences of their beliefs. Any modal logic with standard Kripke semantics in which belief is formalized as a necessity operator has this property. Logical omniscience definitely does not apply to human beings, who have only limited time available and have bounded rationality: it is unrealistic to assume that they believe every logical theorem, however complicated. There are several possible solutions to the problem, involving non-standard semantics or syntactic operators for awareness and explicit belief. Good reference to the logical omniscience problem and its possible solutions are [27, Chapter 2] and [16, Chapter 9].

Another problem with collective belief and common knowledge is that they are hard to attain in situations where the communication channel is not commonly known to be trustworthy. For example, in file transmission protocols at any time only a bounded level of belief E-BEL$_G^{k+1}(\varphi)$ (and knowledge as well) about the message is achieved [23, 31]. Good references to the difficulties concerning the attainment of collective belief, as well as to possible solutions, is [16, Chapter 11].

We acknowledge that these problems are important and should be adequately solved. In this paper, however, we focus on the formalization of collective motivational attitudes needed for teamwork, and for the time being we choose to base it on the relatively simple, even if not practically adequate, logic for collective belief defined above. Thus, we view collective belief as a good *abstraction tool* to study teamwork.

3.4. Axioms for individual motivational attitudes

Our framework to describe motivational attitudes and related aspects is minimal in the sense that we aim to deal with concise necessary and sufficient conditions. Additional aspects appearing on the stage in specific cases may be addressed by refining the system and adding new axioms. This subsection focuses on individual goals and intentions, and gives a short overview of our choice of axioms (adapted from [28]) and the corresponding semantic conditions. In this paper, we leave out of consideration the aspects of time and action in order to focus on the main problem, the definition of collective intentions in terms of more basic attitudes.

For the motivational operators GOAL and INT the axioms include the system K, which we adapt for n agents to K_n. For $i = 1, \ldots, n$ the following axioms and rules are included:

A1 All instantiations of tautologies of the propositional calculus

R1 From φ and $\varphi \rightarrow \psi$ infer ψ (Modus Ponens)

A2$_G$ $\mathrm{GOAL}(i,\varphi) \wedge \mathrm{GOAL}(i,\varphi \rightarrow \psi) \rightarrow \mathrm{GOAL}(i,\psi)$ (Goal Distribution)

A2$_I$ $\mathrm{INT}(i,\varphi) \wedge \mathrm{INT}(i,\varphi \rightarrow \psi) \rightarrow \mathrm{INT}(i,\psi)$ (Intention Distribution)

R2$_G$ From φ infer $\mathrm{GOAL}(i,\varphi)$ (Goal Generalization)

R2$_I$ From φ infer $\mathrm{INT}(i,\varphi)$ (Intention Generalization)

In a BDI system, an agent's activity starts from goals. In general, it may have many different objectives which will not all be pursued. As opposed to intentions, goals are not directly related to actions, so an agent can behave rationally, even though it has different inconsistent goals. Thus, in contrast to Rao and Georgeff we adopt the basic system K_n for goals. Then, the agent chooses a limited number of these goals to be intentions. In this paper we do not discuss how intentions are formed from a set of goals (but see [10, 8]). In any case, we assume that intentions are chosen in such a way that consistency is preserved. Thus for intentions we assume, as Rao and Georgeff do, that they should be consistent. This can be formulated as follows:

A6$_I$ $\neg\mathrm{INT}(i,\bot)$ for $i = 1,\ldots,n$ (Intention Consistency Axiom)

Nevertheless, in the presented approach other choices may be adopted without consequences for the rest of the definitions in this paper.

It is not hard to prove soundness and completeness of the basic axiom systems for goals and intentions with respect to suitable classes of models by a tableau method, and also give decidability results using a small model theorem.

3.5. Interdependencies between attitudes

Interdependencies between belief and individual motivational attitudes are expressed by the following axioms for $i = 1,\ldots,n$:

A7$_{GB}$ $\mathrm{GOAL}(i,\varphi) \rightarrow \mathrm{BEL}(i,\mathrm{GOAL}(i,\varphi))$ (Positive Introspection for Goals).

A7$_{IB}$ $\mathrm{INT}(i,\varphi) \rightarrow \mathrm{BEL}(i,\mathrm{INT}(i,\varphi))$ (Positive Introspection for Intentions).

A8$_{GB}$ $\neg\mathrm{GOAL}(i,\varphi) \rightarrow \mathrm{BEL}(i,\neg\mathrm{GOAL}(i,\varphi))$ (Negative Introspection for Goals).

A8$_{IB}$ $\neg\mathrm{INT}(i,\varphi) \rightarrow \mathrm{BEL}(i,\neg\mathrm{INT}(i,\varphi))$ (Negative Introspection for Intentions).

These four axioms express that agents are aware of the goals and intentions they have, as well as of the lack of those that they do not have. Notice that we do not add the axioms of *strong realism* that Rao and Georgeff adopt for a specific set of formulas φ, the so-called O-formulas: $\mathrm{GOAL}(i,\varphi) \rightarrow \mathrm{BEL}(i,\varphi)$ and $\mathrm{INT}(i,\varphi) \rightarrow \mathrm{BEL}(i,\varphi)$, corresponding to the fact that an agent believes that it can optionally achieve its goals and intentions by carefully choosing its actions. These axioms correspond to semantic restrictions on the branching time models considered in [28]. On the other hand, we do not adopt the converse axiom of *realism* advocated by Cohen and Levesque: $\mathrm{BEL}(i,\varphi) \rightarrow \mathrm{GOAL}(i,\varphi)$. In their formalism, where a possible world corresponds to a time line, the realism axiom expresses that agents

adopt as goals the inevitable facts about the world [7]. Both versions of realism are intimately connected to the choice of temporal structure, a question that we leave out of consideration here.

The semantic property corresponding to $\mathbf{A7}_{IB}$ is $\forall s,t,u((sB_it \wedge tI_iu) \rightarrow sI_iu)$, analogously for $\mathbf{A7}_{GB}$.

The property that corresponds to $\mathbf{A8}_{IB}$ is $\forall s,t,u((sI_it \wedge sB_iu) \rightarrow uI_it)$, analogously for $\mathbf{A8}_{GB}$.

Here follows the proof of the correspondence for $\mathbf{A8}_IB$. For the easy direction, suppose that $\forall s,t,u((sI_it \wedge sB_iu) \rightarrow uI_it)$ holds in a Kripke frame F. Now take any valuation *Val* on the set of worlds W, and let $\mathcal{M}$ be the Kripke model arising from F by adding *Val*. Now take any $s \in W$ with $\mathcal{M},s \models \neg\text{INT}(i,\varphi)$, then there is a $t \in W$ with sI_it and $\mathcal{M},t \not\models \varphi$. We will show that $\mathcal{M},s \models \text{BEL}(i,\neg\text{INT}(i,\varphi))$. So take any $u \in W$ such that sB_iu. By the condition on the frame, we have uI_it, so $\mathcal{M},u \models \neg\text{INT}(i,\varphi)$, and indeed $\mathcal{M},s \models \text{BEL}(i,\neg\text{INT}(i,\varphi))$. Therefore, $F \models \neg\text{INT}(i,\varphi) \rightarrow \text{BEL}(i,\neg\text{INT}(i,\varphi))$.

For the other direction, work by contraposition and suppose that the condition does not hold in a certain frame F. Then there are worlds s,t,u in the set of worlds W such that sI_it and sB_iu but *not* uI_it. Now the the valuation *Val* on F such that for all $v \in W$, $Val(p) = 1$ iff uI_iv, and let $\mathcal{M}$ be the Kripke model arising from F by adding *Val*. Then by definition $\mathcal{M},t \not\models p$, so $\mathcal{M},s \models \neg\text{INT}(i,p)$. On the other hand, $\mathcal{M},u \models \text{INT}(i,p)$, so $\mathcal{M},s \not\models \text{BEL}(i,\neg\text{INT}(i,p))$. We may conclude that $F \not\models \neg\text{INT}(i,p) \rightarrow \text{BEL}(i,\neg\text{INT}(i,p))$.

We assume that every intention corresponds to a goal:

$\mathbf{A9}_{IG}$ $\text{INT}(i,\varphi) \rightarrow \text{GOAL}(i,\varphi)$ (Intention implies goal)

This means that if an agent adopts a formula as an intention, it should have adopted that formula as a goal to achieve, which satisfies Bratman's notion that an agent's intentions form a specific subset of its goals [2]. Rao and Georgeff adopt this axiom as *goal-intention compatibility* for their class of O-formulas [28]. In our non-temporal context, the corresponding semantic property is that $G_i \subseteq I_i$.

Here follows the proof of the correspondence. For the easy direction, suppose that $G_i \subseteq I_i$ holds in a Kripke frame F. Now take any valuation *Val* on the set of worlds W, and let $\mathcal{M}$ be the Kripke model arising from F by adding *Val*. Now take any $s \in W$ with $\mathcal{M},s \models \text{INT}(i,\varphi)$, but suppose, in order to derive a contradiction, that $\mathcal{M},s \not\models \text{GOAL}(i,\varphi)$. Then there is a $t \in W$ with sG_it and $\mathcal{M},t \not\models \varphi$. But because $G_i \subseteq I_i$ we have sI_at as well, contradicting the assumption $\mathcal{M},s \models \text{INT}(i,\varphi)$. Therefore, $F \models \text{INT}(i,\varphi) \rightarrow \text{GOAL}(i,\varphi)$.

For the other direction, work by contraposition and suppose that $G_i \subseteq I_i$ does not hold in a certain frame F. Then there are worlds s,t in the set of worlds W such that sG_it but *not* sI_it. Now the the valuation *Val* on F such that for all $v \in W$, $Val(p) = 1$ iff sI_iv, and let $\mathcal{M}$ be the Kripke model arising from F by adding *Val*. Then by definition $\mathcal{M},s \models \text{INT}(i,p)$; but $\mathcal{M},t \not\models p$, so $\mathcal{M},u \not\models \text{GOAL}(i,p)$. We may conclude that $F \not\models \text{INT}(i,p) \rightarrow \text{GOAL}(i,p)$.

Note that in our system it fortunately does not follow that an agent intends all the consequences it believes its intentions to have, i.e. the side-effects of its intentions. There is weaker version that does hold, though, namely if $\models \varphi \rightarrow \psi$, then $\models \text{INT}(i,\varphi) \rightarrow \text{INT}(i,\psi)$. This is similar to the logical omniscience problem for logics of knowledge and belief discussed in subsection 3.3. For a discussion of the "side-effect problem" for intentions, see [2, 7, 28].

4. Collective motivational attitudes

In our approach, teams are created on the basis of *collective intentions*, which are defined in subsection 4. for the standard context and in subsection 4.2 for less ideal circumstances. A team exists as long as the collective intention between team members exists. However in this paper we abstract from the ways in which teams are formed, and refer the interested reader to [5, 9, a0, 25, 38].

As a reminder, in our approach collective intention is viewed as an inspiration for the creation of collective commitment leading directly to action execution. This is based on the linguistic tradition that intentions typically ultimately lead to actions, however, the immediate triggers of these actions are commitments. In the sequel we extend this view to the collective case. In this paper we concentrate on collective intentions, while collective commitments are treated in [aa, a2], and most extensively in a forthcoming paper.

We assume that the creation of a collective commitment is based on the corresponding collective intention and hinges on the allocation of actions according to an adopted plan. However, some agents in the team may not have delegated actions while still being involved in the collective intention and the collective commitment.

The reader will note that collective intentions are not introduced here as primitive modalities, with some restrictions on the semantic accessibility relations (as in e.g. [6]). We do give necessary and sufficient conditions for such collective motivational attitudes to be present. In this way, we hope to make the behavior of a team easier to predict. We have tried to find *minimal* conditions for collective intentions to be present, and not to weigh down the definitions with all aspects that play a part in the establishment of collective intentions. Such elements as conventions, abilities, opportunities, power relations and social structure (see [30, 32, 38] for a thorough discussion) certainly are important, and we leave open the possibility of defining and using them in specific cases where they play a crucial role. For example, abilities and opportunities are important in the dialogues leading up to the establishment of a collective intention and a team based on it [a0]. Power relations and social structure, on the other hand, are reflected in the definitions of collective commitments (see the forthcoming paper).

In the philosophical and MAS literature there is an ongoing discussion as to whether collective intentions may be reduced to individual ones plus collective beliefs about them (see [4, 2a, 33]). Even though our definition seems to be reductive, it involves nested intentions and collective epistemic operators, and for this reason is deeper than a simple compound built out of individual intentions and collective beliefs about them by propositional connectives only.

4.1. Collective intentions: the standard case

In this paper, we focus on strictly cooperative teams (see for example [3, 32] for good philosophical discussions of cooperation). This makes the definition of collective intention rather strong. In such teams, a necessary condition for a collective intention is that all members of the team G have the associated individual intention $\mathrm{INT}(i, \varphi)$ towards the state of the world represented by proposition φ. In fact, this condition is taken to give the full definition of collective intention in [29] (see [37] for a similar definition of collective goal). However, this is certainly not sufficient. Imagine that two agents want to achieve the same state of the world but are in a competition about this, willing to achieve it exclusively. Therefore, to exclude the case of competition, all agents should *intend* all members to have the associated individual intention, as well as the intention that all members have the individual intention, and so on;

we call such a mutual intention M-INT$_G(\varphi)$. Furthermore, all members of the team are aware of this mutual intention, that is, they have a collective belief about this (C-BEL$_G$(M-INT$_G(\varphi)$). Of course, team members remain autonomous maintaining their other motivational attitudes, and may even be in competition about other issues.

In order to formalize the above two conditions, E-INT$_G(\varphi)$ (standing for "everyone intends") is defined by the following axiom, corresponding to the semantic condition that $\mathcal{M}, s \models$ E-INT$_G(\varphi)$ iff for all $i \in G$, $\mathcal{M}, s \models$ INT(i, φ):

M1 E-INT$_G(\varphi) \leftrightarrow \bigwedge_{i \in G}$ INT(i, φ).

The mutual intention M-INT$_G(\varphi)$ is meant to be true if everyone in G intends φ, everyone in G intends that everyone in G intends φ, etc. As we do not have infinite formulas to express this, let E-INT$^1_G(\varphi)$ be an abbreviation for E-INT$_G(\varphi)$, and let E-INT$^{k+1}_G(\varphi)$ for $k \geq 1$ be an abbreviation of E-INT$_G$(E-INT$^k_G(\varphi)$). Thus we have $\mathcal{M}, s \models$ M-INT$_G(\varphi)$ iff $\mathcal{M}, s \models$ E-INT$^k_G(\varphi)$ for all $k \geq 1$. Define world t to be *G_I-reachable* from world s if there is a path of length ≥ 1 in the Kripke model from s to t along accessibility arrows I_i that are associated with members i of G. Then the following property holds (see subsection 3.3 and [16] for an analogous property for collective belief and common knowledge, respectively):

$$\mathcal{M}, s \models \text{M-INT}_G(\varphi) \text{ iff } \mathcal{M}, t \models \varphi \text{ for all } t \text{ that are } G_I\text{-reachable from } s.$$

Using this property, it can be shown that the following fixed-point axiom and rule can be soundly added to the union of KD_n and **M1**:

M2 M-INT$_G(\varphi) \leftrightarrow$ E-INT$_G(\varphi \wedge$ M-INT$_G(\varphi))$

RM1 From $\varphi \rightarrow$ E-INT$_G(\psi \wedge \varphi)$ infer $\varphi \rightarrow$ M-INT$_G(\psi)$ (Induction Rule)

The resulting system is called $KD^{\text{M-INT}_G}_n$, and it is sound and complete with respect to Kripke models where all n accessibility relations are serial. The completeness proof will be given in section 5. Now we will show the soundness of Rule **RM1** with respect to the given semantics. (The other axioms and rules are more intuitive, so we leave their soundness to the reader). Suppose that $\models \varphi \rightarrow$ E-INT$_G(\psi \wedge \varphi)$, meaning that $\varphi \rightarrow$ E-INT$_G(\psi \wedge \varphi)$ holds in all worlds of all Kripke models. We need to show that $\models \varphi \rightarrow$ M-INT$_G(\psi)$. So take any Kripke model $\mathcal{M} = (W, \{B_i : i \in \mathcal{A}\}, \{G_i : i \in \mathcal{A}\}, \{I_i : i \in \mathcal{A}\}, Val)$ with $\mathcal{A} = \{1, \ldots, n\}$, and any world $s \in W$ with $\mathcal{M}, s \models \varphi$. Now suppose that t is G_I-reachable from s in k steps along the path $w_0, \ldots w_k$ with $w_0 = s$ and $w_k = t$, by $k \geq 1$ relations of the form I_j ($j \in \{1, \ldots, n\}$). We need to show that $\mathcal{M}, t \models \psi$, for which we can show step by step that $\psi \wedge \varphi$ holds in all worlds w_i, $i \geq 1$, on the path from s to t. For the first step, e.g. sI_jw_1, we can use the fact that $\mathcal{M}, s \models \varphi \rightarrow$ E-INT$_G(\psi \wedge \varphi)$, and thus $\mathcal{M}, s \models$ E-INT$_G(\psi \wedge \varphi)$, to conclude that $\mathcal{M}, w_1 \models \psi \wedge \varphi$. Repeating this reasoning on the path to t, we conclude that for all i with $1 \leq i \leq k$, $\mathcal{M}, w_i \models \psi \wedge \varphi$, in particular $\mathcal{M}, t \models \psi$. We conclude $\mathcal{M}, s \models \varphi \rightarrow$ M-INT$_G(\psi)$, as desired.

Finally, the collective intention is defined by the following axiom:

M3 C-INT$_G(\varphi) \leftrightarrow$ M-INT$_G(\varphi) \wedge$ C-BEL$_G$(M-INT$_G(\varphi)$)

Note that this definition is different from the one given in [11, 12]. The definition would be even stronger if common knowledge instead of collective belief appeared in **M3**. However, because common knowledge is almost impossible to establish in multiagent systems due to (among other circumstances) the unreliability of communication media, we do not pursue this strengthening further.

The resulting system, which we call $KD_n^{\text{C-INT}_G}$, is the union of $KD_n^{\text{M-INT}_G}$ (for mutual intentions), $KD45_n^C$ (for collective beliefs) and axiom **M3**.

Let us give an informal **example** of the establishment of a collective intention. Two violinists, a and b, have studied together and have toyed with the idea of giving a concert together someday. Later this becomes more concrete: they both intend to perform together the solo parts of the Bach Double Concerto ($\text{INT}(a, \varphi)$ and $\text{INT}(b, \varphi)$, where φ stands for "a and b perform the solo parts of the Bach Double Concerto"). After communicating with each other about this, they start practising together. Clearly, a mutual intention as defined in **M2** is now in place (involving nested intentions like $\text{INT}(a, \text{INT}(b, \text{INT}(a, \varphi)))$ and so on). The communication established a collective belief $\text{C-BEL}_G(\varphi)$ (with $G = \{a, b\}$) about their mutual intention, according to **M3**. As sometimes happens in life, when people are ready, an opportunity appears: Carnegie Hall plans a concert for Christmas Eve, including the Bach Double Concerto. Now they refine their collective intention to a more concrete $\text{C-INT}_G(\psi)$ (where ψ stands for "a and b perform the solo parts of the Bach Double Concerto at the Christmas Eve concert in Carnegie Hall"). It happens that our two violinists are chosen from among a list of candidates to be the soloists, and both sign the appropriate contract. Because they do this together, we can speak about common knowledge, not merely collective belief, of their mutual intention: $\text{M-INT}_G(\psi) \wedge \text{C-KNOW}_G(\text{M-INT}_G(\psi))$.

One important difference between common knowledge and collective belief is that common knowledge can be justified if needed, and a commonly signed contract provides a perfect basis for this. It is clear that the two violinists have developed a very strong and concrete variant of collective intention due to their common knowledge of the mutual intention.

Even though $\text{C-INT}_G(\varphi)$ seems to be an infinite concept, collective intentions may be established in practice in a finite number of steps. As defined by **M3**, collective intentions are appropriate to model those situations in which communication, in particular announcements, work, especially if one initiator establishes the team. We have showed in detail in [9, 10] how team formation in such an ideal case may actually work in terms of the first two stages of collective problem solving, namely potential recognition and team formation, and how at these stages the proper attitudes are established through dialogues consisting of the appropriate speech acts.

The above definition is applicable also to cases that may be anticipated when designing a team or system behaviour. For example, emergency situations are such a class of cases. Given a specific application (e.g. a yacht on the sea), different emergency situations are often classified, and based on it, roles of team members are predefined, accordingly. In other words, in specific circumstances, team members know their roles in advance, and have individual intentions to fulfill them. They intend others to fulfill their intentions as well, etc. Thus, the mutual intention $\text{M-INT}_G(\varphi)$ is in place immediately, especially when saving lives depend on it!

The amount of necessary communication, as expressed by $\text{C-BEL}_G(\text{M-INT}_G(\varphi))$, clearly depends on circumstances, and varies from just recognizing the situation by perception when communication is difficult or impossible, through simply confirming what situation we deal with (then agents' roles are clear), to the more complex cases when, for example, some agents / roles are missing, so that more communication is needed.

The following lemma follows immediately from the definition of collective intention, using lemma 3.1.

Lemma 4.1. Let φ be a formula and $G \subseteq \{1, \ldots, n\}$. Then the following holds:

$$\text{C-INT}_G(\varphi) \rightarrow \text{C-BEL}_G(\text{C-INT}_G(\varphi)).$$

One question is if our definition of collective intention does not cover inappropriate cases where real teamwork is out of the question. Bratman [3, Ch. 5] characterizes shared cooperative activity. He presents some examples of situations where the agents share some attitudes, but shared cooperative activity is out of the question. It turns out that our definition excludes these cases, as well. For example,

> "Suppose that you and I each intend that we go to New York together, and this is known to both of us. However, I intend that we go together as a result of my kidnapping you and forcing you to join me. The expression of my intention, we might say, is the Mafia sense of "We're going to New York together". While I intend that we go to New York together, my intentions are clearly not cooperative in spirit."

Thus, taking φ ="a and b go to New York" with a for "you" b for "me", G for $\{a, b\}$, in the situation above $\text{E-INT}_G(\varphi)$ and possibly also $\text{C-BEL}_G(\text{E-INT}_G(\varphi))$ holds, but $\text{M-INT}_G(\varphi)$ and $\text{C-INT}_G(\varphi)$ do not. Specifically, it seems unlikely that $\text{INT}(b, \text{INT}(a, \varphi))$ holds for the Mafioso. Note that Rao, Georgeff and Sonenberg's definition of a *joint intention* among G to achieve φ is defined as $\text{E-INT}_G(\varphi) \wedge \text{C-BEL}_G(\text{E-INT}_G(\varphi))$ (translated to our notation), thus it erroneosly ascribes a joint intention to go to New York among the agents in the example. Incidentally, a similar one-level definition of mutual goals was also given in [37].

Comparison with the two-level definition

In previous work, we gave a somewhat weaker definition of collective intention than the one above (see [11, 12]). It consisted of two levels of reciprocal intentions in a team, and a collective belief about this; so it did not erroneously assign a collective intention to sets whose members are in individual competition, as a one-level definition does (such as it appears in e.g. [29]). Here follows the two-level definition:

$$\begin{aligned}\text{C-INT}_G(\varphi) \leftrightarrow\ & \text{E-INT}_G(\varphi) \wedge \text{C-BEL}_G(\text{E-INT}_G(\varphi)) \\ & \wedge \text{E-INT}_G(\text{E-INT}_G(\varphi)) \wedge \text{C-BEL}_G(\text{E-INT}_G(\text{E-INT}_G(\varphi)))\end{aligned}$$

However, the above definition did not preclude competition among more-person coalitions. Consider the following example. Three world-famous violinists A, B, and C are candidates to be one of the two lead players needed for the Bach Double Concerto, to be performed in Carnegie Hall on Christmas Eve. They are asked to decide among themselves who will be the two soloists. Imagine the situation where all three of them want to be one of the "chosen two", and they also want both other players to want this - as long as it is with them, not with the third player; e.g. A is against a coalition between B and C. Thus, for φ = "there will be a great performance of the Bach Double Concerto in Carnegie Hall on Christmas Eve", we have the two levels for reciprocal intention among $\{A, B, C\}$ (for example $\text{INT}(A, \text{INT}(B, \varphi))$, and even $\text{M-INT}_{\{A,B\}}$), but not a third one: A does not intend that B intends C to intend φ (so there is no $\text{M-INT}_{\{A,B,C\}}$). Thus one would hardly say that a collective intention among them is in place: they are not a team, but rather three competing coalitions of two violinists each.

If we adapt the definition above to make it consist of three levels of intention instead of two, the troublesome example would be solved. However, one may invent similar (admittedly artificial) examples for any k, using coalitions of k people from among a base set of at least $k+1$ agents. Thus, the infinitary mutual intention of the previous section was derived to avoid all such counterexamples. One can see, however, that in practical situations, for any fixed finite group, a finite number of levels of the mutual intention is sufficient to construct the collective intention among them.

4.2. Alternative definitions

In some situations, for example time-critical ones, we will now argue that teamwork may tentatively start even if the collective intention as defined above has not yet been established, especially in circumstances where it has not been possible (yet) to establish a collective belief among the team about their mutual intention. For such situations, as a start for teamwork we define a notion that is somewhat stronger than the mutual intention defined in axiom **M2**: in the axiom **M2′** below, even though a collective belief about the mutual intention has not been established in actual fact, all members of the group intend it to be established.

Consider, for **example**, a situation in which a person c has disappeared under the ice and two potential helpers a and b are in the neighbourhood; they do not know each other, and there is no clearly marked initiator among them. Suppose further that, at this point in time, communication among them is not possible, for example because of strong wind. Perception is possible in a limited way: they can see the other one move, but cannot distinguish facial expressions. Both have the individual intention to help (thus $\mathrm{INT}(a,\varphi)$ and $\mathrm{INT}(b,\varphi)$, i.e. $\text{E-BEL}_G(\varphi)$, where $G = a,b$ and φ stands for "c has been rescued"). Moreover, in general two persons are needed for a successful rescue, and this is a collectively believed fact ($\text{C-BEL}_G(\psi)$, where ψ stands for "at least two persons are needed to achieve φ"). As there are no other potential helpers around, a and b believe that they need to act together. Thus, we may expect that a mutual intention $\text{M-INT}_G(\varphi)$ is already established. Both agents may even form an individual belief about the mutual intention being established, so at this point there may be $\text{M-INT}_G(\varphi) \wedge \text{E-BEL}_G(\text{M-INT}_G(\varphi))$. However, communication being limited, the collective belief about the mutual intention ($\text{C-BEL}_G(\text{M-INT}_G(\varphi))$) cannot be established; for this reason, the standard collective intention $\text{C-INT}_G(\varphi)$ does not hold. On the other hand, time is critical, so *some* team-like attitude needs to be established. In this situation, it is justified that goal-directed activity may be based on a revised notion of mutual intention.

In order to build a proper collective commitment, leading to team action, from the present attitude $\text{M-INT}_G(\varphi)$, the collective belief is necessary. For example in the rescue situation, such a collective belief enables co-ordination needed for mouth-on-mouth breathing and heart massage. Both agents believe this: they believe that if φ is ever achieved, a collective intention $\text{C-INT}_G(\varphi)$ has been established before. Thus, even if communication is severely restricted at present, they still try to establish a team together, and do both *intend* that the collective belief about the mutual intention be established to make real teamwork possible.

Thus, the alternative mutual intention $\text{M-INT}'_G(\varphi)$ is meant to be true if everyone in G intends φ, everyone in G intends that everyone in G intends φ, etc. (as in $\text{M-INT}_G(\varphi)$); moreover, everyone intends that there be collective belief in the group of this whole infinite conjunction ($\text{E-INT}_G(\text{C-BEL}_G(\text{M-INT}_G(\varphi)))$). This is reflected by the following axiom, which can be soundly added to $KD_n^{\text{C-INT}_G}$ (for standard collective intentions):

M2′ $\text{M-INT}'_G(\varphi) \leftrightarrow \text{E-INT}_G(\varphi \wedge \text{C-INT}_G(\varphi))$

The resulting system is called $KD_n^{\text{M-INT}'_G}$, and it is easily seen to be sound and complete with respect to Kripke models where all n accessibility relations for both I and B are serial, while those for B are additionally transitive and euclidean.

The notion of $\text{M-INT}'_G$ is appropriate for unstable situations in which communication is hard or impossible and in which a team needs to be formed. From this perspective, $\text{M-INT}'_G$ may be called a "pre-collective intention", from which the team members will in a later stage hopefully establish a collective belief. This leads to an alternative definition of collective intention $\text{C-INT}'_G$ based on $\text{M-INT}'_G$. This notion is stronger than its standard counterpart C-INT_G: now, both intended and factual establishment of a collective belief about the mutual intention are present. It is defined by the following axiom:

M3′ $\text{C-INT}'_G(\varphi) \leftrightarrow \text{M-INT}'_G(\varphi) \wedge \text{C-BEL}_G(\text{M-INT}'_G(\varphi))$

The system resulting when adding **M3′** to $KD_n^{\text{M-INT}'_G}$ is called $KD_n^{\text{C-INT}'_G}$. Note that $\text{M-INT}'_G$ includes intentions about awareness, whereas in the original definition C-INT_G awareness exists, whether or not intended. Therefore $\text{M-INT}'_G$ is stronger that M-INT_G, but it is not comparable to C-INT_G. Furthermore, $\text{C-INT}'_G$ is a stronger notion than C-INT_G.

4.3. Comparing the strength of the different notions of mutual and collective intentions

The alternative definition of collective intention is the strongest notion; it implies both the alternative mutual intention and the standard collective intention. These two in turn are logically independent of each other, but both imply the standard mutual intention, Formally, the diamond of implications may be represented as follows:

$$\vdash \text{C-INT}'_G(\varphi) \rightarrow \text{M-INT}'_G(\varphi)$$

$$\vdash \text{C-INT}'_G(\varphi) \rightarrow \text{C-INT}_G(\varphi)$$

$$\vdash \text{C-INT}_G(\varphi) \rightarrow \text{M-INT}_G(\varphi)$$

$$\vdash \text{M-INT}'_G(\varphi) \rightarrow \text{M-INT}_G(\varphi)$$

$$\nvdash \text{M-INT}'_G(\varphi) \rightarrow \text{C-INT}_G(\varphi)$$

$$\nvdash \text{C-INT}_G(\varphi) \rightarrow \text{M-INT}'_G(\varphi)$$

The proofs are straightforward and are left to the reader.

5. Completeness proof for the logic of mutual intention $KD_m^{\text{M-INT}_G}$

In this section, a completeness proof is given for the logic of mutual intentions in the standard case (see subsection 4.1). Together with soundness of the system, which is immediate, the completeness proof enables the designer of a multiagent system to test the validity of various properties concerned with collective intentions, by checking models instead of constructing axiomatic proofs. In addition, the completeness proof gives an upper bound on the complexity of reasoning about the satisfiablity of such properties: by a "small model theorem" for an analogous system, the problem has been shown to be in EXPTIME (see [16]).

The method of proof is one used often in modal logic when proving completeness with respect to *finite* models, for example when one shows decidability of a system. The proof is inspired by the one for the logic of common knowledge in [16], which is in turn inspired by Parikh's completeness proof for propositional dynamic logic. In fact, the main difference consists in adapting their proof to our slightly different choice of axioms, and filling in some steps that were left to the reader in [16]. The full presentation of the proof is meant to suggest to the reader that the method may be adapted to prove completeness as well for the combined systems for individual and mutual / collective beliefs and intentions such as $\text{KD}_n^{\text{C-INT}_G}$, $\text{KD}_n^{\text{M-INT}'_G}$ and $\text{KD}_n^{\text{C-INT}'_G}$.

We have to prove that, supposing that $KD_m^{\text{M-INT}_G} \not\vdash \varphi$, there is a serial model $\mathcal{M}$ and a $w \in \mathcal{M}$ such that $\mathcal{M}, w \not\models \varphi$. There will be four steps:

1 A finite set of formulas Φ, the *closure* of φ, will be constructed that contains φ and all its subformulas, plus certain other formulas that are needed in step 4 below to show that an appropriate valuation falsifying φ at a certain world can be defined. The set Φ is also closed under single negations.

2 A "Lindenbaümchen" lemma will be proved: a consistent set of sentences from Φ can always be extended to a set that is maximally consistent in Φ.

3 These finitely many maximally consistent sets will correspond to the states in the Kripke countermodel against φ, and appropriate accessibility relations and a valuation will be defined on these states.

4 It will be shown, using induction on all formulas in Φ, that the model constructed in step 3 indeed contains a world in which φ is false. This is the most complex step in the proof.

Below, the closure of a sentence φ is defined. One can view it as the set of formulas that are *relevant* for making a countermodel against φ.

Definition 5.1. The *closure* of φ with respect to $KD_m^{\text{M-INT}_G}$ is the minimal set Φ of $KD_m^{\text{M-INT}_G}$-formulas such that for all $G \subseteq \{1, \ldots, m\}$ the following hold:

1. $\varphi \in \Phi$.
2. If $\psi \in \Phi$ and χ is a subformula of ψ, then $\chi \in \Phi$.
3. If $\psi \in \Phi$ and ψ itself is not a negation, then $\neg\psi \in \Phi$.
4. If $\text{M-INT}_G(\psi) \in \Phi$ then $\text{E-INT}_G(\psi \wedge \text{M-INT}_G(\psi)) \in \Phi$.

5. If E-INT$_G(\psi) \in \Phi$ then INT$(i, \psi) \in \Phi$ for all $i \in G$.

6. $\neg$INT$(i, \bot) \in \Phi$ for all $i \leq m$.

It is straightforward to prove that for every formula φ, the closure Φ of φ with respect to $KD_m^{\text{M-INT}_G}$ is a *finite* set of formulas.

This finishes step 1 of the completeness proof. The next definition leads up to the Lindenbaümchen Lemma, step 2 of the proof.

Definition 5.2. A finite set of formulas Γ such that $\Gamma \subseteq \Phi$ is *maximally* $KD_m^{\text{M-INT}_G}$ *-consistent in* Φ if and only if:

1. Γ is $KD_m^{\text{M-INT}_G}$-consistent, i.e. $KD_m^{\text{M-INT}_G} \not\vdash \neg(\bigwedge_{\psi \in \Gamma} \psi)$.

2. There is no $\Gamma' \subseteq \Phi$ such that $\Gamma \subset \Gamma'$ and Γ' is still $KD_m^{\text{M-INT}_G}$-consistent.

Lemma 5.1. (Lindenbaümchen Lemma)
Let Φ be the closure of φ with respect to $KD_m^{\text{M-INT}_G}$. If $\Gamma \subseteq \Phi$ is $KD_m^{\text{M-INT}_G}$-consistent, then there is a set $\Gamma' \supseteq \Gamma$ which is maximally $KD_m^{\text{M-INT}_G}$-consistent in Φ.

Proof By standard techniques of modal logic: enumerating all formulas and subsequently adding a formula or its negation depending on whether $KD_m^{\text{M-INT}_G}$-consistency is preserved or not.

Now we are ready to take step 3, namely to define the model that will turn out to contain a world where $\neg\varphi$ holds.

Definition 5.3. Let $\mathcal{M}_\varphi = < S_\varphi, \pi, I_1, \ldots, I_m >$ be a Kripke model defined as follows:

- As domain of states, one state s_Γ is defined for each maximally $KD_m^{\text{M-INT}_G}$-consistent $\Gamma \subseteq \Phi$. Note that, because Φ is finite, there are only finitely many states. Formally, we define $\text{CON}_\Phi = \{\Gamma \mid \Gamma \text{ is maximally } KD_m^{\text{M-INT}_G}\text{-consistent in } \Phi\}$ and $S_\varphi = \{s_\Gamma \mid \Gamma \in \text{CON}_\Phi\}$.

- To make a truth assignment π, we want to conform to the propositional atoms that are contained in the maximally consistent sets corresponding to each world. Thus, we define $\pi(s_\Gamma)(p) = 1$ if and only if $p \in \Gamma$. Note that this makes all propositional atoms that do not occur in φ false in every world of the model.

- The relations I_i are defined as follows: $I_i = \{(s_\Gamma, s_\Delta) \mid \psi \in \Delta \text{ for all } \psi \text{ such that INT}(i, \psi) \in \Gamma\}$.

It will turn out that using this definition, we not only have $\mathcal{M}_\varphi, s_\Gamma \models p$ iff $p \in \Gamma$ for propositional atoms p, but such an equivalence holds for all relevant formulas. This is proved in the Finite Truth Lemma, the main result of step 4.

In order to prove the Finite Truth Lemma, we need to prove some essential properties of maximally $KD_m^{\text{M-INT}_G}$-consistent sets in Φ, namely the Consequence Lemma and the Finite Valuation Lemma.

Lemma 5.2. (Consequence Lemma)
If $\Gamma \in \text{CON}_\Phi$, $\psi_1, \ldots, \psi_n \in \Gamma$, $\chi \in \Phi$ and $KD_m^{\text{M-INT}_G} \vdash \psi_1 \to (\psi_2 \to (\ldots(\psi_n \to \chi)\ldots))$, then $\chi \in \Gamma$.

Proof The proof is straightforward, by standard reasoning about maximal consistent

Lemma 5.3. (Finite Valuation Lemma)
If Γ is $KD_m^{\text{M-INT}_G}$-consistent in some closure Φ, then for all ψ, χ it holds that:

1. If $\neg\psi \in \Phi$, then $\neg\psi \in \Gamma$ iff $\psi \notin \Gamma$.
2. If $\psi \wedge \chi \in \Phi$, then $\psi \wedge \chi \in \Gamma$ iff $\psi \in \Gamma$ and $\chi \in \Gamma$.
3. If $\text{INT}(i,\psi) \in \Phi$, then $\text{INT}(i,\psi) \in \Gamma$ iff $\psi \in \Delta$ for all Δ with $(s_\Gamma, s_\Delta) \in I_i$.
4. If $\text{E-INT}_G(\psi) \in \Phi$, then $\text{E-INT}_G(\psi) \in \Gamma$ iff $\psi \in \Delta$ for all Δ and all $i \in G$ such that $(s_\Gamma, s_\Delta) \in I_i$.
5. If $\text{M-INT}_G(\psi) \in \Phi$, then $\text{M-INT}_G(\psi) \in \Gamma$ iff $\psi \in \Delta$ for all Δ that are G_I reachable from s_Γ.

Proof Items 1 and 2 are proved by standard modal logic techniques. We will now prove 3, 4, and 5.
3: **the** INT**-case** Suppose $\text{INT}(i,\psi) \in \Phi$.

$\Rightarrow$ Suppose $\text{INT}(i,\psi) \in \Gamma$, and suppose that $(s_\Gamma, s_\Delta) \in I_i$. Then by definition of I_i, we immediately have $\psi \in \Delta$, as desired.

$\Leftarrow$ Suppose, by contraposition, that $\text{INT}(i,\psi) \notin \Gamma$. We need to show that there is a Δ such that $(s_\Gamma, s_\Delta) \in I_i$ and $\psi \notin \Delta$. It suffices to show the following **Claim**: the set of formulas $\Delta' = \{\chi \mid \text{INT}(i,\chi) \in \Gamma\} \cup \{\neg\psi\}$ is $KD_m^{\text{M-INT}_G}$-consistent. For if the claim is true, then by the Lindenbaümchen Lemma there exists a maximally $KD_m^{\text{M-INT}_G}$-consistent $\Delta \supseteq \Delta'$ in Φ. By the definitions of Δ' and I_i we have $(s_\Gamma, s_\Delta) \in I_i$, and by 1, we have $\psi \notin \Delta$, as desired. So let us prove the claim. In order to derive a contradiction, suppose Δ' is not $KD_m^{\text{M-INT}_G}$-consistent. Because Δ' is finite, we may suppose that $\{\chi \mid \text{INT}(i,\chi) \in \Gamma\} = \{\chi_1, \ldots, \chi_n\}$. Then by definition of inconsistency,

$$KD_m^{\text{M-INT}_G} \vdash \neg(\chi_1 \wedge \ldots \wedge \chi_n \wedge \neg\psi).$$

By propositional reasoning, we get

$$KD_m^{\text{M-INT}_G} \vdash \chi_1 \rightarrow (\chi_2 \rightarrow (\ldots(\chi_n \rightarrow \psi)\ldots)).$$

Then by necessitation ($\mathbf{R2}_I$) plus a number of applications of ($\mathbf{A2}_I$) and more propositional reasoning, we derive

$$KD_m^{\text{M-INT}_G} \vdash \text{INT}(i,\chi_1) \rightarrow (\text{INT}(i,\chi_2) \rightarrow (\ldots(\text{INT}(i,\chi_n) \rightarrow \text{INT}(i,\psi))\ldots)).$$

However, we know that $\text{INT}(i,\chi_1), \ldots, \text{INT}(i,\chi_n) \in \Gamma$ and $\text{INT}(i,\psi) \in \Phi$, so by the Consequence Lemma, $\text{INT}(i,\psi) \in \Gamma$, contradicting our starting assumption.

4: **the** E-INT$_G$**-case** Suppose E-INT$_G(\psi) \in \Phi$; then by the construction of Φ also INT$(i,\psi) \in \Phi$ for all $i \in G$.

$\Rightarrow$ Suppose E-INT$_G(\psi) \in \Gamma$. Axiom (M1) and some easy propositional reasoning gives us $KD_m^{\text{M-INT}_G} \vdash$ E-INT$_G(\psi) \rightarrow$ INT(i,ψ) for all $i \in G$. Because INT$(i,\psi) \in \Phi$ we can use the Consequence Lemma and derive that INT$(i,\psi) \in \Gamma$ for all $i \in G$. Thus, by the $\Rightarrow$-step of the INT-case, we have $\psi \in \Delta$ for all Δ and all $i \in G$ such that $(s_\Gamma, s_\Delta) \in I_i$, as desired.

$\Leftarrow$ The proof is very similar to the $\Rightarrow$-step, this time using (**M1**) and the $\Leftarrow$-step of the INT-case.

5: **the** M-INT$_G$**-case** Let $s_\Gamma \longrightarrow^k s_\Delta$ stand for "s_Δ is G_I-reachable from s_Γ in k steps". Suppose M-INT$_G(\psi) \in \Phi$; then by the construction of Φ also E-INT$_G(\psi \wedge$ M-INT$_G(\psi)) \in \Phi$, as well as its subformulas.

$\Rightarrow$ Suppose M-INT$_G(\psi) \in \Gamma$. We will prove by induction that for all $k \geq 1$ and all Δ, if $s_\Gamma \longrightarrow^k s_\Delta$, then ψ, M-INT$_G(\psi) \in \Delta$. (Note that this is stronger than what is actually needed for the $\Rightarrow$-step; such a loaded induction hypothesis makes the proof easier.

k=1 Suppose that $s_\Gamma \longrightarrow^1 s_\Delta$; this means that $\Gamma I_i \Delta$ for some $i \in G$. By axiom **M2** we have $\$KD_m^{\text{M-INT}_G} \vdash$ M-INT$_G(\psi) \rightarrow$ E-INT$_G(\psi \wedge$ M-INT$_G(\psi))$.

So because M-INT$_G(\psi) \in \Gamma$ and E-INT$_G(\psi \wedge$ M-INT$_G(\psi)) \in \Phi$, the Consequence Lemma implies that E-INT$_G(\psi \wedge$ M-INT$_G(\psi)) \in \Gamma$. But then, by 4, the $\Rightarrow$-side of 3, and 2, we conclude that ψ, M-INT$_G(\psi) \in \Delta$, as desired.

k=n+1 Suppose that $s_\Gamma \longrightarrow^{n+1} s_\Delta$ for some $n \geq 1$, then there is a Δ' such that $s_\Gamma \longrightarrow^n s_{\Delta'}$ and $s_{\Delta'} \longrightarrow s_\Delta$. By the induction hypothesis, we have ψ, M-INT$_G(\psi) \in \Delta'$. Now, just as in the base case k=1, one can prove that the formulas ψ, M-INT$_G(\psi)$ are transferred from Δ' to the direct successor Δ.

$\Leftarrow$ This time we work directly, not by contraposition. So suppose $\psi \in \Delta$ for all Δ for which s_Δ is G_I-reachable from s_Γ. We have to prove that M-INT$_G(\psi) \in \Gamma$.

First a general remark. Because each s_Δ corresponds to a *finite* set of formulas Δ, each Δ can be represented as the finite conjunction of its formulas, denoted as φ_Δ. Note that it is crucial that we restricted ourselves to the finite closure Φ.

Now define W as $\{\Lambda \in \text{CON}_\Phi \mid \psi \in \Delta$ for all Δ for which s_Λ is G_I-reachable from $s_\Gamma\}$. So in particular, $\Gamma \in W$. Intuitively, W should become the set of worlds in which M-INT$_G(\psi)$ holds.

Now let $\varphi_W = \bigvee_{\Lambda \in W} \varphi_\Lambda$. This formula is the disjunction of the descriptions of all states corresponding to W. ¿From the finiteness of W, it follows that φ_W is a formula of the language. Similarly, define $\varphi_{\overline{W}} = \bigvee_{\Theta \in \overline{W}} \varphi_\Theta$, where $\overline{W} = \{\Theta \in \text{CON}_\Phi \mid \Theta \notin W\}$.

Our aim is to prove the following *Claim*:

$$KD_m^{\text{M-INT}_G} \vdash \varphi_W \rightarrow \text{E-INT}_G(\varphi_W).$$

First, let's show how this claim helps to prove the desired conclusion M-INT$_G(\psi) \in \Gamma$. Because $\psi \in \Lambda$ for all $\Lambda \in W$ and ψ occurs in all conjunctions φ_Λ for all $\Lambda \in W$, we have $KD_m^{\text{M-INT}_G} \vdash \varphi_W \to \psi$. Starting from this and the claim above, we may distribute E-INT$_G$ over the implication by a number of uses of (**R2**$_I$), (**A2**$_I$), (**M1**) and some propositional reasoning to derive $KD_m^{\text{M-INT}_G} \vdash \varphi_W \to \text{E-INT}_G(\psi \wedge \varphi_W)$. Rule (**RM1**) immediately gives $KD_m^{\text{M-INT}_G} \vdash \varphi_W \to \text{M-INT}_G(\psi)$. Now because φ_Γ is one of the disjuncts of φ_W, we have $KD_m^{\text{M-INT}_G} \vdash \varphi_\Gamma \to \text{M-INT}_G(\psi)$. Finally, using the Consequence Lemma and some more propositional reasoning, we conclude M-INT$_G(\psi) \in \Gamma$, as desired.

Thus, it remains to prove the claim $KD_m^{\text{M-INT}_G} \vdash \varphi_W \to \text{E-INT}_G(\varphi_W)$. We do this in the following five steps.

1. We first show that for all $i \in G$ and for all $\Lambda \in W$ and $\Theta \in \overline{W}$, $KD_m^{\text{M-INT}_G} \vdash \varphi_\Lambda \to \text{INT}(i, \neg\varphi_\Theta)$. So suppose $\Lambda \in W$ and $\Theta \in \overline{W}$.

   By definition of W and $\overline{W}$, we have $\psi \in \Delta$ for all Δ for which s_Δ is G_I-reachable from s_Λ, but there is a Δ' such that $s_{\Delta'}$ is G_I-reachable from s_Θ and $\psi \notin \Delta'$. Therefore, $(s_\Lambda, s_\Theta) \notin I_i$ for any $i \in G$. Choose an $i \in G$. By definition of I_i, there is a formula χ_i such that $\text{INT}(i, \chi_i) \in \Lambda$ while $\chi_i \notin \Theta$. As Θ is maximally $KD_m^{\text{M-INT}_G}$-consistent in Φ, we have $KD_m^{\text{M-INT}_G} \vdash \varphi_\Theta \to \neg\chi_i$, thus by contraposition $KD_m^{\text{M-INT}_G} \vdash \chi_i \to \neg\varphi_\Theta$. Using (**R2**$_I$) and (**A2**$_I$), we derive $KD_m^{\text{M-INT}_G} \vdash \text{INT}(i, \chi_i) \to \text{INT}(i, \neg\varphi_\Theta)$, and as $\text{INT}(i, \chi_i) \in \Lambda$, we have $KD_m^{\text{M-INT}_G} \vdash \varphi_\Lambda \to \text{INT}(i, \neg\varphi_\Theta)$.
2. $KD_m^{\text{M-INT}_G} \vdash \varphi_\Lambda \to \text{INT}(i, \bigwedge_{\Theta \in \overline{W}} \neg\varphi_\Theta)$. In fact, this follows from 1 by propositional logic and the well-known derived rule of standard modal logic that intention distributes over conjunctions.
3. Here we show that $KD_m^{\text{M-INT}_G} \vdash \bigvee_{\Delta \in \text{CON}_\Phi} \varphi_\Delta$.

   Proof Suppose on the contrary that the formula $\neg \bigvee_{\Delta \in \text{CON}_\Phi} \varphi_\Delta$, which is equivalent by De Morgan's laws to $\bigwedge_{\Delta \in \text{CON}_\Phi} \neg\varphi_\Delta$, is $KD_m^{\text{M-INT}_G}$-consistent.

   Then we can find for every $\Delta \in \text{CON}_\Phi$ a conjunct ψ_Δ of φ_Δ such that $\overline{\Delta} := \{\neg\psi_\Delta \mid \Delta \in \text{CON}_\Phi\}$ is $KD_m^{\text{M-INT}_G}$-consistent.

   Thus, by Lemma 5.1, there is a set of formulas $\Theta \supseteq \overline{\Delta}$ which is maximally $\overset{\text{M-INT}_G}{m}$-consistent in Φ. Now we come to the desired contradiction by diagonalization: Θ contains both ψ_Θ (which was defined as a conjunct of φ_Θ) and, because $\Theta \supseteq \overline{\Delta}$, also $\neg\psi_\Theta$.
4. $KD_m^{\text{M-INT}_G} \vdash \varphi_W \leftrightarrow (\bigwedge_{\Theta \in \overline{W}} \neg\varphi_\Theta)$. This follows almost immediately from 3.
5. Here we show the final claim that

$$KD_m^{\text{M-INT}_G} \vdash \varphi_W \to \text{E-INT}_G(\varphi_W).$$

   Proof: By 2 and 4 we have for all $i \in G$ that $KD_m^{\text{M-INT}_G} \vdash \varphi_\Gamma \to \text{INT}(i, \varphi_W)$, so by (**M1**) and some propositional reasoning, $KD_m^{\text{M-INT}_G} \vdash \varphi_\Gamma \to \text{E-INT}_G(\varphi_W)$; finally, because $\Gamma \in W$, our claim holds.

Lemma 5.4. (Finite Truth Lemma)
If $\Gamma \in \text{CON}_\Phi$, then for all $\psi \in \Phi$ it holds that $\mathcal{M}_\varphi, s_\Gamma \models \psi$ iff $\psi \in \Gamma$.

Proof Immediately from the Finite Valuation Lemma, by induction on the structure of ψ. Details are left to the reader.

Theorem 5.1. (completeness of $KD_m^{\text{M-INT}_G}$)
If $KD_m^{\text{M-INT}_G} \nvdash \varphi$, then there is a serial model $\mathcal{M}$ and a $w \in \mathcal{M}$ such that $\mathcal{M}, w \not\models \varphi$.

Proof Suppose $KD_m^{\text{M-INT}_G} \nvdash \varphi$. Take $\mathcal{M}_\varphi$ as defined in definition 5.3. Note that there is a formula χ logically equivalent to $\neg\varphi$ that is an element of Φ; if φ does not start with a negation, χ is the formula $\neg\varphi$ itself. Now, using Lemma 5.1, there is a maximally consistent $\Gamma \subseteq \Phi$ such that $\chi \in \Gamma$. By the Finite Truth Lemma, this implies that $\mathcal{M}_\varphi, s_\Gamma \models \chi$, thus $\mathcal{M}_\varphi, s_\Gamma \not\models \varphi$. The model is serial, because for all $s_\Delta \in S_\varphi$ we have by the finite valuation lemma that $M, s_\Delta \models \neg\text{INT}(i, \bot)$ for all $i \leq m$; so all worlds have I_i successors for all agents.

6. Discussion and conclusions

The most influential theory of CPS is the one of Wooldridge and Jennings [38]. We agree that ([38]):

> "The key mental states that control agent behaviour in our model are intentions and joint intentions — the former define local asocial behaviour, the latter control social behaviour. Intentions are so central because they provide both the stability and predictability that is necessary for social interaction, and the flexibility and reactivity that is necessary to cope with the changing environment."

The actual formal frameworks of their papers are quite different from ours, however. Wooldridge and Jennings define joint commitment towards φ in a more dynamic way than we define collective intentions: initially, the agents do not believe φ is satisfied, and subsequently have φ as a goal until the termination condition is satisfied, including (as conventions) conditions on the agents to make their eventual beliefs that termination is warranted into mutual beliefs. Subsequently, they define having a joint intention to do α as "having a joint commitment that α will happen next, and then α happens next". In contrast, agreeing with [4], we view collective commitments as stronger than collective intentions, and base the collective commitment on a specific social plan meant to realize the collective intention. Our ideas on collective commitments are presented in [11, 12] and in a forthcoming paper, which discusses the dynamic aspects.

The emphasis on establishing appropriate collective attitudes for teamwork is shared with Grosz and Kraus [19, 20]. Nevertheless, the intentional component in their definition of collective plans is much weaker than our collective intention: Grosz and Kraus' agents involved in a collective plan have individual intentions towards the overall goal and a collective belief about these intentions; intentions with respect to the other agents play a part only at the level of individual sub-actions of the collective plan. We stress, however, that team members' intentions about their colleagues' motivation to achieve the overall goal play an important role in keeping the team on track even if their plan has to be changed radically due to a changing environment (see also [13]).

Balzer and Tuomela [1] take a technical approach using fixed points, inspired by the work on common knowledge in epistemic logic [27, 16]. They define we-attitudes such as collective goals and intentions

using fixed-point definitions. Our definitions use fixed-point constructions as well, but interpret collective intentions a bit differently. In Balzer's and Tuomela's view, abilities and opportunities play a part during the construction of a collective intention (the stage of team formation). In our approach, on the other hand, abilities are mainly important at the two surrounding stages, namely during potential recognition (before the stage of team formation) and during plan formation, where a collective commitment is established on the basis of a collective intention and a social plan (see [10, 13, 14]). Also, we provide a completeness proof for the logic with respect to the intended semantics.

Rao, Georgeff and Sonenberg [29] consider some related issues with an emphasis on the ontology and semantics of social agents carrying out social plans. They use a much weaker definition of joint intention than ours: it is only one-level, being defined as "everyone has the individual intention, and there is a collective belief about this". Thus, their definition does not preclude cases of coercion and competition.

Haddadi [21] gives an internal or prescriptive approach that characterizes the stages of CPS in a manner similar to [38], but is based on the semantics of [28] instead of [26]. She introduces the notions of pre-commitments and commitments between pairs of agents and presents an extensive and well-founded discussion of their properties, including important aspects like communication. However, in contrast to our approach, she does not go beyond the level of pairwise commitments and is not explicit about their contribution to collective behavior in a bigger team.

7. Further research

On the basis of individual characteristics of particular agents, their mutual dependencies and other possibly complex criteria one can classify and investigate different types of teams, various types of cooperation, communication, negotiation etc. An interesting extension towards other than strictly cooperative groups will be a subject of our future research.

Although we view collective intention as a central concept during the whole process of CPS, in the present paper we focus on its static aspects during planning. The proper treatment of collective intentions, as well as commitments, in a dynamically changing environment entails the maintenance of all individual, social and collective motivational attitudes involved throughout the whole process. One of the main aspects in CPS is that, due to dynamic and possibly unpredictable environment, team members may fail their tasks or be presented with new opportunities. Thus, it is necessary that team members monitor their performance and re-plan based on the present situation. This leads to the reconfiguration problem. In [13], a generic reconfiguration algorithm for BDI systems is presented by us. In a forthcoming paper, we investigate the persistence and evolution of motivational attitudes during CPS. In addition, in [9, 10, 14] we characterize the role of dialogue in CPS.

8. Acknowledgements

This work is partially supported by the Polish KBN Grant 7T11C 006 20 and by the ALFEBIITE project. We would also like to thank Alexandru Baltag and Michael Luck for their comments.

References

[1] Balzer, W., Tuomela, R.: A Fixed Point Approach to Collective Attitudes, *Contemprorary Action Theory Volume 2: Social Action* (G. Holmstrom-Hintikka, R. Tuomela, Eds.), Kluwer, Dordrecht, 1997.

[2] Bratman, M.: *Intention, Plans, and Practical Reason*, Harvard University Press, Cambridge (MA), 1987.

[3] Bratman, M.: *Faces of Intention*, Cambrridge University Press, Cambridge, 1999.

[4] Castelfranchi, C.: Commitments: From Individual Intentions to Groups and Organizations, *Proceedings First International Conference on Multi-Agent Systems* (V. Lesser, Ed.), AAAI-Press and MIT Press, San Francisco, 1995.

[5] Castelfranchi, C., Miceli, M., Cesta, A.: Dependence Relations Among Autonomous Agents, in: Werner and Demazeau [36].

[6] Cavedon, L., Rao, A., Tidhar, G.: Social and Individual Commitment (Preliminary Report), in: *Intelligent Agent Systems: Theoretical and Practical Issues* (L. Cavedon, A. Rao, W. Wobcke, Eds.), vol. 1209 of *LNAI*, Springer Verlag, Berlin, 1997, 152–163.

[7] Cohen, P., Levesque, H.: Intention is Choice with Commitment, *Artificial Intelligence*, **42**, 1990, 213–261.

[8] Dignum, F., Conte, R.: Intentional Agents and Goal Formation: Extended Abstract, *Preproceedings Fourth International Workshop on Agent Theories, Architectures and Languages* (M. Singh, A. Rao, M. Wooldridge, Eds.), Providence, Rhode Island, 1997.

[9] Dignum, F., Dunin-Kęplicz, B., Verbrugge, R.: Agent Theory for Team Formation by Dialogue, *Proceedings Agent Theories, Architectures and Languages (ATAL 2000)* (C. Castelfranchi, Y. Lesperance, Eds.), Boston, 2000.

[10] Dignum, F., Dunin-Kęplicz, B., Verbrugge, R.: Creating Collective Intention through Dialogue, *Logic Journal of the IGPL*, **9**, 2001, 145–158.

[11] Dunin-Kęplicz, B., Verbrugge, R.: Collective Commitments, *Proceedings Second International Conference on Multi-Agent Systems* (M. Tokoro, Ed.), AAAI-Press, Menlo Park (CA), 1996.

[12] Dunin-Kęplicz, B., Verbrugge, R.: Collective motivational attitudes in cooperative problem solving, *Proceedings of the First International Workshop of Eastern and Central Europe on Multi-agent Systems (CEEMAS'99)* (V. Gorodetsky, Ed.), St. Petersburg, 1999.

[13] Dunin-Kęplicz, B., Verbrugge, R.: A Reconfiguration Algorithm for Distributed Problem Solving, *Electronic Modeling*, **22**, 2000, 68 – 86.

[14] Dunin-Kęplicz, B., Verbrugge, R.: The Role of Dialogue in Collective Problem Solving, *Proceedings of the Fifth International Symposium on the Logical Formalization of Commonsense Reasoning (Commonsense 2001)* (E. Davis, J. McCarthy, L. Morgenstern, R. Reiter, Eds.), New York, 2001.

[15] Dunin-Kęplicz, B., Verbrugge, R.: Evolution of Collective Commitment During Reconfiguration, *Proceedings of the First Conference on Autonomous Agents and Multiagent Systems (AAMAS02)*, 2002.

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

[17] Ferber, J.: *Multi-agent Systems: An Introduction to Distributed Artificial Intelligence*, Addison Wesley, Reading (MA), 1999.

[18] Graedel, E.: Why is Modal Logic so Robustly Decidable?, *Bulletin of the EATCS*, **68**, 1999, 90–103.

[19] Grosz, B., Kraus, S.: Collaborative Plans for Complex Group Action, *Artificial Intelligence*, **86(2)**, 1996, 269–357.

[20] Grosz, B., Kraus, S.: The Evolution of SharedPlans, in: *Foundations of Rational Agency* (A. Rao, M. Wooldridge, Eds.), Kluwer, Dordrecht, 1999, 227–262.

[21] Haddadi, A.: *Communication and Cooperation in Agent Systems: A Pragmatic Theory*, vol. 1056 of *LNAI*, Springer Verlag, Berlin, 1995.

[22] Halpern, J.: The Effect of Bounding the Number of Primitive Propositions and the Depth of Nesting on the Complexity of Modal Logic, *Artificial Intelligence*, **75**, 1995, 361–372.

[23] Halpern, J., Zuck, L.: A Little Knowledge Goes a Long Way: Simple Knowledge-Based Derivations and Correctness Proofs for a Family of Protocols, *6th ACM Symposium on Principles of Distributed Computing*, 1987.

[24] Hustadt, U., Schmidt, R.: On Evaluating Decision Procedures for Modal Logics, *Proceedings IJCAI'97* (M. Pollack, Ed.), Morgan Kauffman, Los Angeles (CA), 1997.

[25] Jennings, N.: Commitments and Conventions: The Foundation of Coordination in Multi-agent Systems, *Knowledge Engineering Review*, **3**, 1993, 223–250.

[26] Levesque, H., Cohen, P., Nunes, J.: On acting together, *Proceedings Eighth National Conference on AI (AAAI90)*, AAAI-Press and MIT Press, Menlo Park (CA), Cambridge (MA), 1990.

[27] Meyer, J.-J. C., van der Hoek, W.: *Epistemic Logic for AI and Theoretical Computer Science*, Cambridge University Press, Cambridge, 1995.

[28] Rao, A., Georgeff, M.: Modeling Rational Agents within a BDI-architecture, *Proceedings of the Second Conference on Knowledge Representation and Reasoning* (R. Fikes, E. Sandewall, Eds.), Morgan Kaufman, 1991.

[29] Rao, A., Georgeff, M., Sonenberg, E.: Social Plans: A Preliminary Report, in: Werner and Demazeau [36], 57–76.

[30] Singh, M.: Commitments among Autonomous Agents in Information-rich Environments, in: *Multi-Agent Rationality (Proceedings of MAAMAW'97)* (M. Boman, W. V. de Velde, Eds.), vol. 1237 of *LNAI*, Springer Verlag, Berlin, 1997, 141–155.

[31] Stulp, F., Verbrugge, R.: A knowledge-based Algorithm for the Internet Protocol TCP, *Bulletin of Economic Research*, **54**(1), 2002, 69–94.

[32] Tuomela, R.: *The Importance of Us: A Philosophical Study of Basic Social Notions*, Stanford Series in Philosophy, Stanford University Press, Stanford (CA), 1995.

[33] Tuomela, R., Miller, K.: We-intentions, *Philosophical Studies*, **53**, 1988, 367–390.

[34] Vardi, M.: Why is Modal Logic so Robustly Decidable?, *DIMACS Series on Discrete Mathematics and Theoretical Computer Science*, **31**, 1997, 149–184.

[35] Weiss, G., Ed.: *Multiagent Systems*, MIT Press, Cambridge, MA, 1999.

[36] Werner, E., Demazeau, Y., Eds.: *Decentralized A.I.-3*, Elsevier, Amsterdam, 1992.

[37] Wooldridge, M., Jennings, N.: Towards a Theory of Collective Problem Solving, in: *Distributed Software Agents and Applications* (J. Perram, J. Muller, Eds.), vol. 1069 of *LNAI*, Springer Verlag, Berlin, 1996, 40–53.

[38] Wooldridge, M., Jennings, N.: Cooperative Problem Solving, *Journal of Logic and Computation*, **9**, 1999, 563–592.