Collective Intention Constitutes a Group

Barbara Dunin-Kęplicz[1] and **Rineke Verbrugge**[2]

[1] Institute of Informatics, Warsaw University,
Banacha 2, 02-097 Warsaw, Poland
`keplicz@mimuw.edu.pl`
[2] Artificial Intelligence, University of Groningen,
Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands
`rineke@ai.rug.nl`

Abstract. In this paper the notion of *collective intention* in groups of agents involved in *cooperative problem solving* (CPS) in *multiagent systems* (MAS) is investigated. Starting from individual *intentions*, *goals*, and *beliefs*, we arrive at an understanding of a collective intention in a group of agents. The presented definitions are rather strong, in particular a collective intention includes that all members *intend* for all others to share that intention. In this situation, a group of agents is created on the basis of a 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 a collective intention within a group lasts long enough.

Within the presented approach, two alternative versions of collective intentions are discussed.They are formalized in a multi-modal logical framework. Completeness of this logic with respect to an appropriate class of Kripke models is proved elsewhere. Presented definitions reflect only vital aspects of motivational attitudes, leaving room for case-specific extensions. This makes the framework flexible and not overloaded.

1 Introduction

Some MAS are referred to as *intentional systems* realizing 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 and desires as its *goals*, intentions and, next, commitments refer to its *motivational* attitudes.

The elements of individual and common activity, all together constituting *teamwork* or CPS, are reflected in an agent's practical reasoning. 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*. The key concept is the one of *intention*. We do not aim to present a psychologically sound theory of motivations driving human behaviour. We are interested in studying those aspects of motivations 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. On the other hand, we do not consider any specific notion of rationality, for example the economic one used in game theory. The only assumption made here is that agents are logical reasoners.

¿From the perspective of group activity, motivational attitudes are considered on three levels: *individual, social* (i.e. bilateral), and *collective*. Our goal is to arrive at an understanding of the notion of *intention in groups of agents*. We agree with [17] 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.

There is a common agreement that intentions, and then commitments, play a number of important roles in practical reasoning [2, 5]:

- 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.*

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, however, leads to the problem of balancing *pro-active*, (i.e. goal-directed) and *reactive* (i.e. event-driven) behaviour.

In our approach we try to maintain this balance very carefully on the three different levels of teamwork: individual, social and collective. On the individual and social level the problem of persistence of both intentions and then commitments is first expressed in agent's individual *intention* and *commitment strategies*, addressing the question: *when and how can an agent responsibly drop its intention or commitment ?* One answer to this question is discussed in [19]. The collective level is apparently much more complex; our approach to collective commitments can be found in [9, 10], while in [7, 11] the role of dialogue in teamwork is investigated. In our framework an agent's pro-activeness and reactiveness are implicitly or explicitly involved on consecutive stages of the reconfiguration algorithm [10]. The formal specification of these situations will be given in a forthcoming paper summing up our theory of CPS.

In our BDI framework, CPS is modeled explicitly. We base our analysis on the [22] model, containing the consecutive stages of *potential recognition*, *team formation*, *plan formation*, and *team action*. Wooldridge and Jennings presented the best recognized and most influential formal theory of CPS. However, they work within a different philosophical tradition and with another set of key notions than we do. As diversity is the core of multiagent systems, there is enough room for different approaches to cooperative problem solving.

In the present paper, the main subject is collective intention as a basis for the establishment of a team. Two different definitions are presented, together with examples of situations in which they apply. 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 in the full paper [12]. 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. However, for specific applications it is possible to reduce some of the infinitary character of collective beliefs and intentions to more manageable proportions.

The rest of this paper is structured in the following manner. Section 2 gives a short logical background. Section 3 is a reminder of the epistemic theory we use, and section 4 describes the logical theory of individual goal and intentions. The heart of the paper is formed by section 5, in which collective intentions are investigated. Sections 6 and 7 close off with a discussion of the results and some ideas for future research.

2 Preliminaries

We propose the use of multi-modal logics to formalize agents' informational and motivational attitudes. In our framework the axioms relating motivational attitudes of agents appear with respect to *propositions*, reflecting particular states of affairs that an agent aims for. Table 1 gives the formulas appearing in this paper, together with their intended meanings, where the symbol φ denotes a proposition. The individual motivational attitudes are primitive but are governed by axiom systems and corresponding semantics, while the collective motivational attitudes are defined by axioms in terms of the individual ones.

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

Table 1. Formulas and their intended meaning

2.1 Kripke models

Each Kripke model for a set of agents $\mathcal{A} = \{1, \ldots, n\}$ consists of a set of worlds, a set of accessibility relations between worlds, and a valuation of the propositional atoms, as follows.

Definition 1 (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 propositional formulas in states.

The truth definition for complex formulas works as usually in modal logics. Truth conditions for the epistemic operators are given in section 3, and those for motivational operators in sections 4 and 5. In general, the expression $M, s \models \varphi$ is read as "formula φ is satisfied by world s in structure M".

3 Beliefs

To represent beliefs, we adopt a standard $KD45_n$-system for n agents as explained in [13], where we take $\mathrm{BEL}(a, \varphi)$ to have as intended meaning "agent a believes proposition φ". $KD45_n$ consists of 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 $\neg\mathrm{BEL}(i, \bot)$ (Consistency)
R1 From φ and $\varphi \rightarrow \psi$ infer ψ (Modus Ponens)
R2 From φ infer $\mathrm{BEL}(i, \varphi)$ (Belief Generalization)

In the semantics, there are accessibility relations B_i that lead from worlds s to worlds that are consistent with agent i's beliefs in s. Thus, BEL is defined semantically as follows:

$$M, s \models \mathrm{BEL}(i, \varphi) \text{ iff } M, t \models \varphi \text{ for all } t \text{ such that } sB_it.$$

Note that the 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.

The property of negative introspection is controversial; we are agnostic about this and dropping **A5** will not have important consequences for the logical framework presented in this paper.

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", where $G \subseteq \mathcal{A}$. It is defined semantically as $M, s \models \text{E-BEL}_G(\varphi)$ iff for all $i \in G,\ M, s \models \text{BEL}(i, \varphi))$, which corresponds to the following axiom:

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

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. $\text{C-BEL}_G(\varphi)$ is meant to be true if everyone in G believes φ, everyone in G believes that everyone in G believes φ, etc. Let $\text{E-BEL}_G^1(\varphi)$ be an abbreviation for $\text{E-BEL}_G(\varphi)$, and let $\text{E-BEL}_G^{k+1}(\varphi)$ for $k \geq 1$ be an abbreviation of $\text{E-BEL}_G(\text{E-BEL}_G^k(\varphi))$. Thus we have $M, s \models \text{C-BEL}_G(\varphi)$ iff $M, s \models \text{E-BEL}_G^k(\varphi)$ for all $k \geq 1$. Note that even collective beliefs need not be true, so $M, s \models \text{C-BEL}_G(\varphi)$ need not imply $M, s \models \varphi$. 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 [13]):

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

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 $\varphi \to \text{E-BEL}_G(\psi \wedge \varphi)$ infer $\varphi \to \text{C-BEL}_G(\psi)$ (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 [13]. In the sequel, we will use the following standard properties of C-BEL_G (see for example [13, exercise 3.11]).

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

- $\text{C-BEL}_G(\varphi \wedge \psi) \leftrightarrow \text{C-BEL}_G(\varphi) \wedge \text{C-BEL}_G(\psi)$
- $\text{C-BEL}_G(\varphi) \to \text{C-BEL}_G(\text{C-BEL}_G(\varphi))$

A good reference to the problems concerning collective belief and to their possible solutions is [13, Chapter 11]. In any case, it is generally agreed that collective belief is a good abstraction tool to study teamwork.

4 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 section focuses on individual goals and intentions, and gives a short overview of our choice of axioms from [19], and corresponding semantic conditions.

As to Kripke semantics, the binary accessibility relations G_i and I_i correspond to each agent's goals (or desires) and intentions, all of which lead from a world to a world. Evaluation of formulas at worlds is defined in the obvious manner inspired by epistemic logic. For $i = 1, \dots, n$ we have:

$M, s \models \mathrm{GOAL}(i, \varphi)$ iff $\forall t$ with sG_it, $M, t \models \varphi$
$M, s \models \mathrm{INT}(i, \varphi)$ iff $\forall t$ with sI_it, $M, t \models \varphi$

For the motivational operators GOAL and INT the axioms include the system K_n. For $i = 1, \dots, n$, next to **A1** and **R1** as in section 3, the following axioms and rules are included:

A2$_G$ $\mathrm{GOAL}(i, \varphi) \wedge \mathrm{GOAL}(i, \varphi \to \psi) \to \mathrm{GOAL}(i, \psi)$ (Goal Distribution)
A2$_I$ $\mathrm{INT}(i, \varphi) \wedge \mathrm{INT}(i, \varphi \to \psi) \to \mathrm{INT}(i, \psi)$ (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. As the agent may have many different objectives, its goals need not be consistent with each other. Then, the agent chooses a limited number of its goals to be intentions. It is not the main focus of this paper to discuss how intentions are formed from a set of goals (but see [7, 6]). However, goals and their relation with intentions form an important part of BDI-theory, so goals are first-class citizens in our system. 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 $i = 1, \dots, n$ (Intention Consistency Axiom)

Rao and Georgeff also add an analogous axiom for the consistency of goals. However, it was argued above that an agent's goals are not necessarily consistent with each other. Thus, we adopt the basic system K_n for goals. 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.

4.1 Interdependencies between Attitudes

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

A7$_{GB}$ $\mathrm{GOAL}(i,\varphi) \to \mathrm{BEL}(i, \mathrm{GOAL}(i,\varphi))$ (Positive Introspection for Goals).
A7$_{IB}$ $\mathrm{INT}(i,\varphi) \to \mathrm{BEL}(i, \mathrm{INT}(i,\varphi))$ (Positive Introspection for Intentions).
A8$_{GB}$ $\neg\mathrm{GOAL}(i,\varphi) \to \mathrm{BEL}(i, \neg\mathrm{GOAL}(i,\varphi))$ (Negative Introspection Goals).
A8$_{IB}$ $\neg\mathrm{INT}(i,\varphi) \to \mathrm{BEL}(i, \neg\mathrm{INT}(i,\varphi))$ (Negative Introspection Intentions).

The semantic properties corresponding to **A7**$_{IB}$ is $\forall s,t,u((sB_it \wedge tI_iu) \to sI_iu)$, analogously for **A7**$_{GB}$. The property that corresponds to **A8**$_{IB}$ is is $\forall s,t,u((sI_it \wedge sB_iu) \to uI_it)$, analogously for **A8**$_{GB}$.

We assume that every intention corresponds to a goal:

A9$_{IG}$ $\mathrm{INT}(i,\varphi) \to \mathrm{GOAL}(i,\varphi)$ (Intention implies goal)

The corresponding semantic property is that $G_i \subseteq I_i$.

5 Collective Motivational Attitudes

Now, we move to the collective level of cooperation. In our approach, groups are created on the basis of *collective intentions*, which are defined in subsection 5.1, and exist as long as the collective intention between team members exists. However in this paper we abstract from the exact ways in which groups are formed, and refer the interested reader to [16, 22, 7].

As a reminder, in our approach collective intention is viewed as an inspiration for the creation of collective commitment leading directly to action execution. 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.

The reader will note that collective intention is not introduced as a primitive modality, with some restrictions on the semantic accessibility relations (as in e.g. [4]). We do give necessary and sufficient conditions for such a collective motivational attitude to be present. In this way, we hope to make the behavior of a team easier to predict.

In the philosophical and MAS literature there is an ongoing discussion whether collective intentions may be reduced to individual ones plus collective beliefs about them (see [21, 3, 15]). Even though our definitions seem to be reductive, they involve fixed-point constructions, and for this reason they are deeper than a simple compound built out of individual intentions and collective beliefs about them by propositional connectives only.

Let us stress again that we have tried to find *minimal* conditions for collective intentions to be present, and not to weigh down the definitions with all aspects that may play a part. Such elements as conventions, abilities, opportunities, power relations and social structure (see [21, 22] 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 in the form of additional axioms.

5.1 Collective Intentions: the Standard Case

In this paper, we focus on strictly cooperative groups, which makes the definition of collective intention rather strong. In such groups, a necessary condition for a collective intention is that all members of the group G have the associated individual intention $\mathrm{INT}(i, \varphi)$ towards the overall goal φ. In fact, this condition is taken to give the full definition of collective intention in [20] (see [22] for a similar definition of collective goal). However, this is certainly not sufficient. Imagine that two agents want to achieve the same goal but are in a competition, willing to achieve the goal exclusively. Therefore, to exclude cases of competition between individuals or subgroups, 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 $\text{M-INT}_G(\varphi)$. Furthermore, all members of the group are aware of this mutual intention, that is, they have a collective belief about this $(\text{C-BEL}_G(\text{M-INT}_G(\varphi))$.

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

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

The 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 we do not have infinite formulas to express this, let $\text{E-INT}^1_G(\varphi)$ be an abbreviation for $\text{E-INT}_G(\varphi)$, and let $\text{E-INT}^{k+1}_G(\varphi)$ for $k \geq 1$ be an abbreviation of $\text{E-INT}_G(\text{E-INT}^k_G(\varphi))$. Thus we have $M, s \models \text{M-INT}_G(\varphi)$ iff $M, s \models \text{E-INT}^k_G(\varphi)$ for all $k \geq 1$. Define world t to be G_I*-reachable* from world s if there is a path 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 section 3 and [13] for an analogous property for collective belief and common knowledge, respectively):

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

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

M2 $\text{M-INT}_G(\varphi) \leftrightarrow \text{E-INT}_G(\varphi \wedge \text{M-INT}_G(\varphi))$
RM1 From $\varphi \to \text{E-INT}_G(\psi \wedge \varphi)$ infer $\varphi \to \text{M-INT}_G(\psi)$ (Induction Rule)

The resulting system is called $KD_n^{\text{M-INT}_G}$, 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 [12].

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

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

Note that this definition is different from the one given in [8, 9].

Let us remark that, 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. Note that collective intentions as defined by **M3** are appropriate to model those situations in which communication, in particular announcements, work, especially if one initiator establishes the team. We have shown in detail in [7] 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 group or system behaviour. For example, protocols for emergency situations lead to such 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 depends on it!

The amount of necessary communication, as expressed by the definition's conjunct $\text{C-BEL}_G(\text{M-INT}_G(\varphi))$, 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 1

Lemma 2. *Let φ be a formula and $G \subseteq \{1, \ldots, n\}$. Then the following holds:* $\text{C-INT}_G(\varphi) \to \text{C-BEL}_G(\text{C-INT}_G(\varphi))$.

5.2 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 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, but perception is. Both have the individual intention to help, and, as two persons are needed for a successful rescue, in fact a mutual intention is already established. Both agents may even form an individual belief about the mutual intention being established, so at this point there may be $\text{M-INT}_G(\varphi) \wedge \text{E-BEL}_G(\text{M-INT}_G(\varphi))$, where $G = \{a, b\}$ and where φ stands for the person under the ice having been rescued. However, without any communication, the collective belief about the mutual intention (and thus a standard collective intention) cannot be established.

In order to build a proper collective commitment, leading to team action, from the present attitude, the collective belief is necessary. Both agents know this, so even if communication is impossible at present, when they try to establish a team together, they 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.; moreover, everyone intends that there be a collective belief in the group of this whole infinite conjunction. Taking the abbreviations $\text{E-INT}^k_G(\varphi)$ and E-BEL^k_G as before, we have the following:

$M, s \models \text{M-INT}'_G(\varphi)$ iff $M, s \models \text{E-INT}^k_G(\varphi) \wedge \text{E-BEL}^l_G(\text{E-INT}^m_G(\varphi))$ for all $k, l, m \geq 1$. Semantically, this boils down to the following definition; definitions for G_I-reachability and G_B-reachability can be found in sections 5.1 and 3, respectively. Further, t is $G_{I,B}$-reachable from s if there is a path in the Kripke model from s to t along accessibility arrows of the form I_i and/or B_i that are associated with members i of G.

Now define world t to be $G_{\text{M-INT}'_G}$*-reachable* from world s if either of the following two holds:

1. t is G_I-reachable from s; or
2. there are t' and t'' in the model and $i, j \in G$ such that sI_it', t'' is $G_{I,B}$-reachable from t', and $t''I_jt$.

Then the following property holds:

$M, s \models \text{M-INT}'_G(\varphi)$ iff $M, t \models \varphi$ for all t that are $G_{\text{M-INT}'_G}$-reachable from s.

Using this property, it can be shown (similarly to the case of the standard definition of mutual intention) that the following axiom and rule can be soundly added to the union of KD_n and **M1** (for intentions) and $KD45^C_n$ (for beliefs):

M2′ $\text{M-INT}'_G(\varphi) \leftrightarrow \text{E-INT}_G(\varphi \wedge \text{M-INT}'_G(\varphi) \wedge \text{C-BEL}_G(\text{M-INT}'_G(\varphi)))$
RM1′ From $\varphi \to \text{E-INT}_G(\psi \wedge \varphi \wedge \text{C-BEL}_G(\text{E-INT}_G(\psi \wedge \varphi)))$ infer $\varphi \to \text{M-INT}'_G(\psi)$ (Induction Rule)

The resulting system is called $KD_n^{\text{M-INT}'_G}$, and it is 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 (by a proof analogous to the one in [12]).

The notion of $\text{M-INT}'_G$ is appropriate for 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 group 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 (alternative) mutual intention are present. It is defined by the following axiom:

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

Note that $\text{M-INT}'_G$ includes intentions about awareness, whereas in the original definition of C-INT_G awareness exists, whether or not it is intended.

6 Discussion and Conclusions

The most influential theory of CPS is the one of Wooldridge and Jennings [22]. We agree with them that ([22]):

> "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."

As regards commitments, they state that ([22]):

> "Previous attemps to formalise commitment have not distinguished between the *commitment* that underpins an intention and the convention. We clearly distinguish the two concepts: a *commitment* is a pledge or a promise; a *convention* is a means of monitoring a commitment — it specifies under what circumstances a commitment can be abandoned and how an agent should behave when one of these conditions arises."

In our formalization we do not share this point of view. Inspired by Castelfranchi's pre-formal discussion [3], we assume that commitments are stronger than intentions, in the sense that it is commitments, not intentions, that trigger action execution. In other words, intentions create a sort of inspiration for a goal-directed group activity, expressed finally in (plan-based) commitments (cf. [8, 9]).

The emphasis on establishing appropriate collective attitudes for teamwork is shared with Grosz and Kraus [14]. 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 subactions 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 [10]).

Tuomela and Miller [21] consider another collective motivational attitude to be the prerequisite for collective activity, namely we-intentions. Their preformal definition is agreement-based, in contrast to our plan-based collective commitments (see [10]). In their view, this agreement need not be fully specified. Therefore, there is no concrete representation of collective plans leading to collective actions in their approach.

Balzer and Tuomela [1] take a technical approach using fixed points, inspired by the work on common knowledge in epistemic logic [18, 13]. 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 [7, 10, 11]). Also, we provide a completeness proof for the logic with respect to the intended semantics in [12].

Rao, Georgeff and Sonenberg [20] consider some related issues with an emphasis on the ontology and semantics of social agents carrying out social plans. They use a weaker definition of joint intention than ours.

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 groups, 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 the 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 [10], a generic reconfiguration algorithm is presented in abstract terms. In a forthcoming paper, we investigate the persistence of motivational attitudes during CPS from a theoretical viewpoint. In particular, the reconfiguration algorithm is proved to be correct. Future work will be to implement interesting instantiations of the reconfiguration algorithm.

8 Acknowledgements

This work is supported by the Polish KBN Grant 7T11C 006 20. We would also like to thank Alexandru Baltag for pointing out that a previous definition of collective intention covered too many cases.

References

1. W. Balzer and R. Tuomela. A fixed point approach to collective attitudes. In G. Holmstrom-Hintikka and R. Tuomela, editors, *Contemprorary Action Theory Volume 2: Social Action*, pages 115–142, Dordrecht, 1997. Kluwer.
2. M. Bratman. *Intension, Plans, and Practical Reason.* Harvard University Press, Cambridge (MA), 1987.
3. C. Castelfranchi. Commitments: From individual intentions to groups and organizations. In V. Lesser, editor, *Proceedings First International Conference on Multi-Agent Systems*, pages 41–48, San Francisco, 1995. AAAI-Press and MIT Press.
4. L. Cavedon, A. Rao, and G. Tidhar. Social and individual commitment (preliminary report). In L. Cavedon, A. Rao, and W. Wobcke, editors, *Intelligent Agent Systems: Theoretical and Practical Issues*, volume 1209 of *Lecture Notes in Artificial Intelligence*, pages 152–163. Springer Verlag, Berlin, 1997.
5. P.R. Cohen and H.J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
6. F. Dignum and R. Conte. Intentional agents and goal formation: Extended abstract. In M.P. Singh, A.S. Rao, and M.J. Wooldridge, editors, *Preproceedings Fourth International Workshop on Agent Theories, Architectures and Languages*, pages 219–231, Providence, Rhode Island, 1997.
7. F. Dignum, B. Dunin-Keplicz, and R. Verbrugge. Creating collective intention through dialogue. In J. Cunningham and D. Gabbay, editors, *Proceedings International Conference on Formal and Applied Practical Reasoning (FAPR 2000)*, pages 145–158, London, 2000. Imperial College.
8. B. Dunin-Kęplicz and R. Verbrugge. Collective commitments. In M. Tokoro, editor, *Proceedings Second International Conference on Multi-Agent Systems*, pages 56–63, Menlo Park (CA), 1996. AAAI-Press.
9. B. Dunin-Kęplicz and R. Verbrugge. Collective motivational attitudes in cooperative problem solving. In Gorodetsky et al. V, editor, *Proceedings of the First International Workshop of Eastern and Central Europe on Multi-agent Systems*, pages 22–41, St. Petersburg, 1999. CEEMAS'99.
10. B. Dunin-Kęplicz and R. Verbrugge. A reconfiguration algorithm for distributed problem solving. *Electronic Modeling*, 22:68 – 86, 2000.
11. B. Dunin-Kęplicz and R. Verbrugge. The role of dialogue in collective problem solving. In Mathias Petsch and Brian Lees, editors, *Proceedings of the Workshop on Intelligent Agents for Computer-supported Cooperative Work*, pages 7–14, Barcelona, 2000.

12. B. Dunin-Kęplicz and R. Verbrugge. Collective intentions. Technical Report TCW-2001, Dept. of Artificial Intelligence, University of Groningen, 2001. downloadable from http://www.ai.rug.nl/rineke/collint1202.ps.
13. R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, MA, 1995.
14. B.J. Grosz and S. Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.
15. A. Haddadi. *Communication and Cooperation in Agent Systems: A Pragmatic Theory*, volume 1056 of *Lecture Notes in Artificial Intelligence*. Springer Verlag, Berlin, 1995.
16. N. Jennings. Commitments and conventions: The foundation of coordination in multi-agent systems. *Knowledge Engineering Review*, 3:223–250, 1993.
17. H.J. Levesque, P.R. Cohen, and J.H.T. Nunes. On acting together. In *Proceedings Eighth National Conference on AI (AAAI90)*, pages 94–99, Menlo Park (CA), Cambridge (MA), 1990. AAAI-Press and MIT Press.
18. J.-J. Ch. Meyer and W. van der Hoek. *Epistemic Logic for AI and Theoretical Computer Science*. Cambridge University Press, Cambridge, 1995.
19. A. Rao and M. Georgeff. Modeling rational agents within a BDI-architecture. In R. Fikes and E. Sandewall, editors, *Proceedings of the Second Conference on Knowledge Representation and Reasoning*, pages 473–484. Morgan Kaufman, 1991.
20. A. Rao, M. Georgeff, and E. Sonenberg. Social plans: A preliminary report. In E. Werner and Y. Demazeau, editors, *Decentralized A.I.-3*, pages 57–76, Amsterdam, 1992. Elsevier.
21. R. Tuomela and K. Miller. We-intentions. *Philosophical Studies*, 53:367–390, 1988.
22. M. Wooldridge and N.R. Jennings. Cooperative problem solving. *Journal of Logic and Computation*, 9:563–592, 1999.