Fundamenta Informaticae 63 (2004) 1–25
IOS Press

A Tuning Machine for Cooperative Problem Solving

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
keplicz@mimuw.edu.pl

Rineke Verbrugge
Institute of Artificial Intelligence
University of Groningen
Grote Kruisstraat 2/1
9712 TS Groningen, The Netherlands
rineke@ai.rug.nl

Abstract. In this paper we aim to formally model individual, social and collective motivational attitudes in teams of agents involved in Cooperative Problem Solving. Particular attention is given to the strongest motivational attitude, collective commitment, which leads to team action. First, building on our previous work, a logical framework is sketched in which social commitments and collective intentions are formalized. Then, different versions of collective commitments are given, reflecting different aspects of Cooperative Problem Solving, and applicable in different situations. The definitions differ with respect to the *aspects* of teamwork of which the agents involved are aware, and the *kind* of awareness present within a team. In this way a kind of tuning mechanism is provided for the system developer to tune a version of collective commitment fitting the circumstances. Finally, we focus attention on a few exemplar versions of collective commitment resulting from instantiating the general tuning scheme, and sketch for which kinds of organization and application domains they are appropriate.

*Address for correspondence: Institute of Informatics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland

1. Introduction

Variety is the core of multiagent systems. This simple sentence expresses the many dimensions on which the field of multiagent systems (henceforth MAS) is distinguished from distributed AI. The basic assumption underlying MAS is relaxing the constraints that were fixed before, in order to meet the needs of goal-directed behaviour in a dynamic and unpredictable environment. This is reflected in complex and possibly flexible patterns of interaction in MAS. Together with autonomy of agents and social structure of cooperative groups this determines the novelty of the agent-based approach.

Variety is the core of multiagent systems also because of important links with other disciplines, as witnessed by the following quote from [30]:

> A number of areas of philosophy have been influential in agent theory and design. The philosophy of beliefs and intentions, for example, led directly to the BDI model of rational agency, used to represent the internal states of an autonomous agent. Speech act theory, a branch of the philosophy of language, has been used to give a semantics to the agent communication language of FIPA. Similarly, argumentation theory — the philosophy of argument and debate, which dates from the work of Aristotle — is now being used by the designers of agent interaction protocols for the design of richer languages, able to support argument and non-deductive reasoning. Issues of trust and obligations in multi-agent systems have drawn on philosophical theories of delegation and norms.
>
> Social sciences: Although perhaps less developed than for economics, various links between agent technologies and the social sciences have emerged. Because multi-agent systems are comprised of interacting, autonomous entities, issues of organisational design and political theory become important in their design and evaluation. Because prediction of other agents' actions may be important to an agent, sociological and legal theories of norms and group behaviour are relevant, along with psychological theories of trust and persuasion. Moreover for agents acting on behalf of others (whether human or not), preference elicitation is an important issue, and so there are emerging links with marketing theory where this subject has been studied for several decades.

Some multiagent systems may be viewed as intentional systems implementing practical reasoning — the everyday process of deciding, step by step, which action to perform next. This model of agency originates from Michael Bratman's theory of human rational choice and action [3]. His theory is based on a complex interplay of informational and motivational aspects, constituting together a belief-desire-intention (BDI) model of rational agency. Intuitively, an agent's *beliefs* correspond to information the agent has about the environment, including other agents. An agent's *desires* represent state of affairs (options) that the agent would choose. Finally, an agent's *intentions* represent a special subset of its desires, namely the options that it has indeed chosen to achieve. The decision process of a BDI agent, based on the interaction of the above-mentioned attitudes, leads to the construction of the agent's *commitment*, leading directly to action execution.

Bratman's theory focuses on the motivational attitudes, in particular on the role that intentions play in practical reasoning. In this research we go one step further investigating the role of the strongest motivational attitude, namely an agent's commitments. After implementing a first step which was a formal specification of a BDI agent as an *individual, autonomous* entity (see section 3 for our choice of logic for this part), the long term goal is to organize *cooperation* of a group of agents in a way allowing to achieve a common, sometimes rather complex, goal when maintaining (at least partial) autonomy of agents involved.

The BDI model of agency comprises beliefs referring to agent's informational attitudes, intentions and then commitments referring to its motivational attitudes. The theory of informational attitudes has been formalized in terms of epistemic logic by [22, 31]. As regards motivational attitudes, the situation is much more complex. In Cooperative Problem Solving (henceforth CPS), a group as a whole needs to act in a coherent pre-planned way, presenting a unified *collective* motivational attitude. This attitude, while staying in accordance with individual attitudes of a group members, should have a higher priority than individual ones. Thus, from the perspective of Cooperative Problem Solving these attitudes are considered on three levels: individual, social (bilateral), and collective.

A coherent and conceptually unified theory of motivational attitudes was missing in the MAS literature. One of the reasons was that attitudes on the bilateral and collective level are not a straightforward extensions or simple sums of individual ones. In order to characterize them, additional subtle and diverse aspects of CPS need to be isolated and then appropriately defined. While this process is far from being trivial, the research presented here brings new results in this respect. A large share of this novelty pertains to the interplay between environmental and social aspects, which may become rather complex nowadays due to the increasing complexity of MAS.

Let us turn for a moment to CPS and teamwork in their commonsense meaning. It is clear that there are different *gradations* of being a team. Take, as **Example 1**, teamwork in a group of researchers who jointly plan their research and divide roles, and who reciprocally keep a check on how the others are doing and help their colleagues when needed in furtherance of their collective intention to prove a theorem. All aspects of teamwork are openly discussed in the team, and members keep each other informed about relevant changes in the plan. Contrast this kind of non-hierarchical teamwork with **Example 2**: a group of spies who all work for the same goal, say to locate Mr. X. In their case a plan is designed by one mastermind, who divides the roles and divulges to each participant *only* the information that is absolutely necessary for him to do his own part. Thus, members may not know what the main goal is, nor even which other agents are included in the group. In the latter example, even though the connection between members is much looser than in the first one, we would still like to speak about CPS, albeit a non-typical case.

In the two examples above, individual and collective awareness about the ingredients of CPS (like the main goal and the plan to achieve it) ranges from very high in the first example to very low in the second. Very informally, collective commitment is the motivational group attitude that provides the glue needed in a team in order to lead it from a still rather abstract collective intention to concrete team action, and it includes the team members' beliefs about each other and the plan they will follow.

Thus, we claim that the two examples above cannot be covered by *one* generic type of collective commitment. Thus far in the MAS literature, when collective attitudes such as collective or joint intentions and collective or joint commitments were characterized, authors provided just one definition geared towards a typical, ideal type of teamwork [41, 15, 35, 24]. These definitions of collective attitudes were independent of organisational structures and communication possibilities. In contrast, the present paper will provide a full range of types of collective commitments and weaker group attitudes that play a similar cohesive role, covering the range from proper teams to more loosely connected groups involved in CPS. We also claim that it is important for system developers to make appropriate decisions about the type or gradation of teamwork needed for a given goal in given circumstances, and to have a mechanism that helps them to choose the corresponding type of group commitment to be created. The material in this paper will help them do just that.

When asking what it means for a group of agents to be *collectively committed* to do something, both the circumstances in which the group is acting and properties of the organization it is part of, have to be taken into account. This implies the importance of differentiating the scope and strength of the notion of collective commitment. The resulting characteristics may differ significantly, and even become logically incomparable.

The aim of this research is to formally model the motivational stance towards Cooperative Problem Solving. Particularly we will investigate the strongest collective attitude, namely *collective commitment*. When addressing the problem in question, different aspects of collective behaviour in strictly cooperative teams of agents have to be modelled. (The case of competition is not included.) Among these, agents' *awareness* about the situation they are involved in seems to be a central one. The notion of awareness that is applied in agent systems may be viewed as a reduction of a general sense of "consciousness" to a state of an agent's beliefs about itself, about other agents, and finally about the state of an environment. Thus, in order to express different scopes and degrees of awareness of cooperating agents, their awareness of a relevant aspect may be characterised in terms of (rather strong) common belief or in weaker forms, like when everybody believes it, or even weaker, when only some agents believe it, depending on the needs and circumstances.

In this context, the idea of *a dial* to be used to tune the nature of the commitment to the particular purpose seems to be both technically interesting and intuitively appealing. We intend to provide a sort of *tuning mechanism* which enables the system developer to *calibrate* a type of collective commitment fitting the circumstances, analogously to adjusting dials on a sound system. The appropriate dials, characterised in the sequel, belong to a device representing a general schema of collective commitment. In order to illustrate the expressive power of such a sort of *tuning machine*, five definitions of commitments corresponding to different teamwork types occurring in practice are presented and discussed. Apparently, the entire spectrum of possibilities is much wider, due to the number of possibly independent choices to be made.

In agent technology, as well as in Computer Science in general, formal logics, particularly modal and temporal logics, are extensively applied. Thus, in the process of modelling BDI systems the key decision is a choice of an appropriate multimodal logic including elements of *intentional logic* to model motivational attitudes, *dynamic logic* to model preconditions of collective action as well as their effects, *epistemic logic* to model informational attitudes, and, possibly, *temporal logic* to model what possibly or necessarily will happen in future. In the presented approach, collective intention and collective commitments are not viewed as primitive modalities: they are defined in terms of individual goals, intentions and beliefs, as well as bilateral commitments. For the resulting multimodal logic the standard Kripke models were chosen, despite their well-known drawbacks (e.g. the logical omniscience problem). The problem of perfecting the logical apparatus is put off as the focus of this paper is on formalization of motivational aspects of CPS. The resulting notion of (group) commitment, described in a chosen multimodal logic, may then be naturally implemented in a specific multiagent system. This way the tuning mechanism may be viewed as a bridge between theory and practice.

This work fits into a research program developed for a couple of years already (compare [15, 11, 19, 17, 18, 20]). It is based on results of previous research, extending and enhancing these results. When investigating group activity during teamwork, crucial ingredients of collective commitment are first isolated and then discussed in depth. This process resulted in a formal theory of collective motivational attitudes. Starting from individual intentions, the first collective notion is the one of *collective intention* for a team, defined and extensively discussed in [19]. Our definitions are stronger than the ones introduced in [35], in particular a collective intention includes that members *intend* for all others to share that intention. Together with individual and collective knowledge and belief, a collective intention constitutes a basis for preparing a plan (or a set of plans). Based on this plan, we characterize the strongest motivational attitude, which is collective commitment of a team. We assume that bilateral aspects of a plan — mutual obligations between agents — are reflected in *social commitments*. Thus collective commitment is defined on the basis of collective intention and social commitments. In other words, our approach to collective commitment is plan-based: the ongoing collective intention is split up into subtasks, according to a given social plan, and then allocated to the team members, as reflected in social commitments between pairs of agents.

Note that in this paper we do not consider the dynamics of individual intention and social commitment adoption, nor the cognitive and social processes involved (but see [9, 11, 5]). Also we forego the important aspects of causality and obligation here. We instead aim to define complex social and collective motivational and epistemic attitudes in terms of simpler individual ones.

The rest of the paper is structured in the following way. In sections 2 and 3, a short reminder is given of the logical framework from our previous work [19]. Individual and collective beliefs are shortly treated, as well as individual motivational attitudes, social commitments and collective intentions. The central section 4 explores different dimensions along which collective commitments may be tuned to fit both the organization and the environment. A general scheme is presented in a multi-modal language, and five different notions of collective commitment fitting to concrete organizational structures are presented in section 5. Finally, section 6 provides some possible generalizations as options for further research, while section 7 focuses on discussion and conclusions. The reader may skip sections 2 and 3 at first reading, and instead start reading from section 4, only jumping back when needing more background about the building blocks of collective commitment. This paper is a revised and expanded version of [21].

2. The language and Kripke semantics

We propose the use of multi-modal logics to formalize agents' informational and motivational attitudes as well as actions they perform and their effects. In CPS, both motivational and informational attitudes are considered on the following three levels: individual, social and collective.

2.1. Language

Individual actions and formulas are defined inductively, both with respect to a fixed finite set of agents. The basis of the induction is given in the following definition.

Definition 2.1. (Basic elements of the language)

The language is based on the following three sets:

- a denumerable set $\mathcal{P}$ of *propositional symbols*;
- a finite set $\mathcal{A}$ of *agents*, denoted by numerals $1, 2, \ldots, n$;
- a finite set $\mathcal{A}t$ of *atomic actions*, denoted by a or b.

In our framework most modalities relating agents' motivational attitudes appear in two forms: with respect to *propositions*, or with respect to *actions*. These actions are interpreted in a generic way — we abstract from any particular form of actions: they may be complex or primitive, viewed traditionally with certain effects or with default effects [12, 13, 14], etc.

A proposition reflects a particular state of affairs. The transition from a proposition that an agent aims for to an action realizing this, is achieved by means-end-analysis. The set of formulas is defined in a double induction, together with the sets of individual actions and social plan expressions (see definitions 2.3 and 2.4). It is extended with other needed modalities. These are all explained later in the paper. See subsection 3.2 about epistemic modalities, and subsections 3.4, 3.5 and 3.6 about individual, social and collective motivational modalities.

Definition 2.2. (Formulas)

We inductively define a set of formulas L 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 $\varphi \wedge \psi$;

F3 if φ is a formula, α is an individual action, $i, j \in \mathcal{A}$, $G \subseteq \mathcal{A}$, and P a social plan expression, then the following are formulas:

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

motivational modalities $\mathrm{GOAL}(i, \varphi)$, $\mathrm{GOAL}(i, \alpha)$, $\mathrm{INT}(i, \varphi)$, $\mathrm{INT}(i, \alpha)$, $\mathrm{COMM}(i, j, \varphi)$, $\mathrm{COMM}(i, j, \alpha)$, $\text{E-INT}_G(\varphi)$, $\text{E-INT}_G(\alpha)$, $\text{M-INT}_G(\varphi)$, $\text{M-INT}_G(\alpha)$, $\text{C-INT}_G(\varphi)$, $\text{C-INT}_G(\alpha)$, $\text{R-COMM}_{G,P}(\varphi)$, $\text{R-COMM}_{G,P}(\alpha)$, $\text{S-COMM}_{G,P}(\varphi)$, $\text{S-COMM}_{G,P}(\alpha)$, $\text{W-COMM}_{G,P}(\varphi)$, $\text{W-COMM}_{G,P}(\alpha)$, $\text{T-COMM}_{G,P}(\varphi)$, $\text{T-COMM}_{G,P}(\alpha)$, $\text{D-COMM}_{G,P}(\varphi)$, $\text{D-COMM}_{G,P}(\alpha)$;

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

Next, we will subsequently describe the class of individual actions $\mathcal{A}c$, and the class of social plan expressions $\mathcal{S}p$. The class $\mathcal{A}c$ is meant to refer to agents' individual actions; they are usually represented without naming the agents.

The individual actions may be combined into group actions by the social plan expressions defined below.

Below, we give a particular choice of operators to be used when defining individual actions and social plan expressions. However, as actions and social plans are not the main subjects of this paper, in the sequel we hardly come into detail as to how particular individual actions and social plans are built up. Thus, another definition (e.g. without the iteration operation or without non-deterministic choice) may be used if more appropriate in a particular context.

Definition 2.3. (Individual actions)

The class $\mathcal{A}c$ of individual actions is defined inductively as follows:

AC1 each atomic action $a \in \mathcal{A}t$ is an individual action;

AC2 if $\varphi \in \mathcal{L}$, then $\texttt{confirm}\,\varphi$ is an individual action; (confirmation)

AC3 if $\alpha_1, \alpha_2 \in \mathcal{A}c$, then $\alpha_1; \alpha_2$ is an individual action; (sequential composition)

AC4 if $\alpha_1, \alpha_2 \in \mathcal{A}c$, then $\alpha_1 \cup \alpha_2$ is an individual action; (non-deterministic choice)

AC5 if $\alpha \in \mathcal{A}c$, then α^* is an individual action; (iteration)

AC6 if $\varphi \in \mathcal{L}$, then $\texttt{stit}(\varphi)$ is an individual action;

Here, in addition to the standard dynamic operators of [**AC1**] to [**AC5**], the operator `stit` of [**AC6**] stands for "sees to it that" or "brings it about that", and has been extensively treated in [37].

Definition 2.4. (Social plan expressions)

The class $\mathcal{S}p$ of social plan expressions is defined inductively as follows:

SP1 If $\alpha \in \mathcal{A}c$ and $i \in \mathcal{A}$, then $\langle \alpha, i \rangle$ is a well-formed social plan expression;

SP2 If α and β are social plan expressions, then $\langle \alpha;\beta \rangle$ (sequential composition) and $\langle \alpha \parallel \beta \rangle$ (parallelism) are social plan expressions.

A concrete example of a social plan expression will be given in subsection 3.1.

2.2. Kripke models

Each Kripke model for the language defined in the previous section consists of a set of worlds, a set of accessibility relations between worlds, and a valuation of the propositional atoms, as follows.

Definition 2.5. (Kripke model)

A Kripke model is a tuple
$\mathcal{M} = (W, \{B_i : i \in \mathcal{A}\}, \{G_i : i \in \mathcal{A}\}, \{I_i : i \in \mathcal{A}\}, Val)$, 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 w.r.t. beliefs, goals, and intentions, respectively. For example, $(w_1, w_2) \in B_i$ means that w_2 is an epistemic alternative for agent i in state w_1.
3. $Val : \mathcal{P} \times W \to \{0, 1\}$ is the function that assigns the truth values to propositional formulas in states.

The truth conditions for the propositional part of the language are all standard. Those for the modal operators are treated in section 3.

3. Building blocks of collective commitments

A collective commitment is built up from a number of building blocks, among them the following:

- the group's attitude toward the main goal: often this is a *collective intention* (see subsection 3.6);
- a *social plan* (see subsection 3.1) meant to achieve the main goal;

- agents' beliefs (see subsection 3.2) with respect to the effectiveness of the social plan;
- agents' *social commitments* (see subsectionSocialcomm) with respect to their own parts of the group activity towards the main goal, as given by the social plan;
- agents' beliefs about the existence of appropriate social commitments.

We propose the use of multi-modal logics to formalize agents' informational and motivational attitudes as well as actions they perform. In CPS, both motivational and informational attitudes are considered on the following three levels: individual, social and collective. In this section, we repeat some notions from our earlier work as they are important for the subject of the present paper. For example, individual motivational attitudes (subsection 3.4) are in turn the building blocks of social commitments and collective intentions, the main ingredients of collective commitments.

3.1. Social plans

Collective commitment are plan-based: they are defined with respect to a given *social plan*. Individual actions (from $\mathcal{A}c$, see section 2) may be combined into group actions by *social plan expressions*, as in definition 2.4 of section 2. Let us give a simple social plan, based on Example 1 of section 1. Consider a team consisting of three agents t (the theorem prover), l (the lemma prover) and c (the proof checker) who have as collective intention to prove a new mathematical theorem. In joint deliberation, they have divided their roles according to their abilities and preferences. Suppose during planning they define two lemmas, which also still need to be proved, and the following complex individual actions: *proveL1*, *proveL2* (to prove lemma 1, respectively 2), *checkL1*,*checkL2* (to check a proof of lemma 1, respectively 2), *proveTh* (prove the theorem from the conjunction of lemmas 1 and 2), *checkTh* (to check the proof of the theorem from the lemmas). One possible social plan they can come up with is the following. First, the lemma prover, who proves lemmas 1 and 2 in succession, and the theorem prover, who proves the theorem from the two lemmas, work in parallel, and subsequently the proof checker checks their proofs in a fixed order, formally:

$$P = \langle\langle\langle\langle proveL1, l\rangle; \langle proveL2, l\rangle\rangle \parallel \langle proveTh, t\rangle\rangle;$$
$$\langle\langle\langle checkL1, c\rangle; \langle checkL2, c\rangle\rangle; \langle checkTh, c\rangle\rangle\rangle$$

The social plan should be effective, as reflected in the predicate $constitute(\varphi, P)$. This states that successful realization of the plan P should lead to the achievement of the main goal φ of the system. The way the predicate $constitute(\varphi, P)$ is constructed and its properties have been discussed in [20].

3.2. Individual and collective beliefs

Some important building blocks of collective commitments concern the group members' *awareness* of aspects like the effectiveness of the social plan and the other agents' role in the group. Such awareness is typically formalized as different strengths of belief, from individual to collective.

To represent beliefs, we adopt a standard $KD45_n$-system for n agents as explained in [22, 31], where we take $\mathrm{BEL}(i, \varphi)$ to have as intended meaning "agent i believes proposition φ". $KD45_n$ consists of the following axioms and rules for $i = 1, \ldots, n$:

A1 All instantiations of propositional tautologies

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 w to worlds that are consistent with agent i's beliefs in w. Thus, BEL is defined semantically as follows:

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

One can define modal operators for group beliefs. The formula $\text{E-BEL}_G(\varphi)$ is meant to stand for "every agent in group G believes φ".

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

A traditional way of lifting single-agent concepts to multi-agent ones is through the use of *collective belief* (or common 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.

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

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 [22].

In the sequel, we will use the following standard properties of C-BEL_G (see for example [22, exercise 3.11]).

Lemma 3.1.

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

3.2.1. Degrees of belief in a group

It is well-known that for teamwork, as well as coordination, it often does not suffice that a group of agents all believe a certain proposition ($\text{E-BEL}_G(\psi)$), but they should collectively believe it ($\text{C-BEL}_G(\psi)$). An example is formed by collective actions where the success of each individual agent is vital to the result, for example, lifting a heavy object together or coordinated attack. It has been proved that for such an attack to be guaranteed to succeed, the starting time of the attack must be a collective belief (even common knowledge) for the generals involved [22].

$\mathrm{COMM}(i, j, \varphi)$	agent i commits to agent j to make φ true
$\mathrm{GOAL}(i, \varphi)$	agent i has as a goal that φ be true
$\mathrm{INT}(i, \varphi)$	agent i has the intention to make φ true
$\text{E-INT}_G(\varphi)$	every agent in 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
$\text{R-COMM}_{G,P}(\varphi)$	group G has robust collective commitment to achieve φ by plan P
$\text{S-COMM}_{G,P}(\varphi)$	group G has strong collective commitment to achieve φ by plan P
$\text{W-COMM}_{G,P}(\varphi)$	group G has weak collective commitment to achieve φ by plan P
$\text{T-COMM}_{G,P}(\varphi)$	group G has team commitment to achieve φ by plan P
$\text{D-COMM}_{G,P}(\varphi)$	group G has distributed commitment to achieve φ by plan P

Table 1. Formulas and their intended meaning

Parikh has introduced a hierarchy of levels of knowledge between individual knowledge and common knowledge and, together with Krasucki, proved a number of interesting mathematical properties. It turns out that, due to the lack of the truth axiom, the similarly defined hierarchy between individual belief and collective belief is structurally different from the knowledge hierarchy [32].

One positive feature of collective belief is that if C-BEL_G holds for ψ, then C-BEL_G also holds for all logical consequences of ψ. The same is true for common knowledge. Thus, agents reason in a similar way from ψ and collectively believe in this similar reasoning and the final conclusions.

In cases in which only $\text{E-BEL}_G(\psi)$ has been established, it is much more difficult for agents to maintain a model of the other team members with respect to ψ and its consequences. However, establishing $\text{E-BEL}_G(\psi)$ places much less constraints on the communication medium than $\text{C-BEL}_G(\psi)$ does. In short, one could say that common knowledge and collective belief are hard to achieve, but easy to understand. Thus, the system developer's decision about the level k of group belief ($\text{E-BEL}^k_G(\psi)$) to be established, hinges on determining a good balance between communication and reasoning for a particular application.

3.3. Notation for individual, social and collective motivational attitudes

Table 1 gives a number of formulas concerning motivational attitudes appearing in this paper, with their intended meanings. The symbol φ denotes a proposition, but all notions also exist with respect to an action α. Even though it may seem from the table as if the formulas have only an informal meaning (perhaps derived from folk psychology), this is actually not the case. In fact, the individual motivational attitudes are primitive but are governed by axiom systems and corresponding semantics, while the social and collective motivational attitudes are defined by axioms in terms of the individual ones.

3.4. Individual motivational attitudes

The theory of collective commitments has as essential basis a theory for practical reasoning for individual agents, covering such attitudes as individual goals and intentions. The key concept in the theory of practical reasoning is the one of *intention*. Intentions 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. However, from time to time an agent's intentions should be reconsidered, for example because they will never be achieved, they are achieved already, or there are no longer reasons supporting them. This leads to the problem of balancing *pro-active*, (i.e. goal-directed) and *reactive* (i.e. event-driven) behaviour.

For the motivational operators GOAL and INT the axioms include the basic modal system K_n. 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 [11, 9]). In any case, we assume that intentions are chosen in such a way that consistency is preserved. Thus for intentions (but not for goals) we assume that they should be consistent:

A6$_I$ $\neg\mathrm{INT}(i, \perp)$ 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.

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 for Goals).

A8$_{IB}$ $\neg\mathrm{INT}(i, \varphi) \to \mathrm{BEL}(i, \neg\mathrm{INT}(i, \varphi))$ (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.

The semantic property corresponding to **A7**$_{IB}$ is $\forall s, t, u((sB_i t \wedge tI_i u) \to sI_i u)$, analogously for **A7**$_{GB}$. The property that corresponds to **A8**$_{IB}$ is $\forall s, t, u((sI_i t \wedge sB_i u) \to uI_i t)$, analogously for **A8**$_{GB}$. The correspondence proofs are given in the Appendix.

In our system, we also 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 $G_i \subseteq I_i$. Again, this is proved in the Appendix.

3.5. Social commitments

In the course of creating a collective commitment, group members *socially commit* (or promise) to take action on their part of the social plan, so that the group may achieve its main goal.

A social commitment between two agents is stronger than an individual intention of one agent. If an agent *socially commits* to a second agent to do something, then the first agent should have the *intention* to do that. Moreover, the first agent commits to the second one only if the second one is *interested* in the first one fulfilling its intention. These two conditions are inspired by [4], but we find that for a social commitment to arise, a third condition is necessary, namely that the agents are aware about the situation, i.e. about their individual attitudes (cf. also [36] for an early discussion about the properties of promises). Such awareness, expressed in terms of collective belief, is generally achieved by communication.

Here follows the defining axiom for social commitments with respect to propositions:

SC1

$$\mathrm{COMM}(i, j, \varphi) \leftrightarrow \mathrm{INT}(i, \varphi) \wedge \mathrm{GOAL}(j, \mathtt{stit}(i, \varphi)) \wedge$$

$$\text{C-BEL}_{\{i,j\}}(\mathrm{INT}(i, \varphi) \wedge \mathrm{GOAL}(j, \mathtt{stit}(i, \varphi)))$$

where $\mathtt{stit}(i, \varphi)$ means that agent i sees to it (takes care) that φ becomes true (see [37]).

Social commitments with respect to actions are defined by the axiom:

SC2

$$\mathrm{COMM}(i, j, \alpha) \leftrightarrow \mathrm{INT}(i, \alpha) \wedge \mathrm{GOAL}(j, \mathit{done}(i, \alpha)) \wedge$$

$$\text{C-BEL}_{\{i,j\}}(\mathrm{INT}(i, \alpha) \wedge \mathrm{GOAL}(j, \mathit{done}(i, \alpha)))$$

Here, $\mathit{done}(i, \alpha)$ means that agent i has just executed action α. The above definition reflects only the ingredients of social commitment that may be expressed in the language of motivational and doxastic (belief) attitudes. As social commitments are not the subject of this paper, we thus forego the important concept of obligation. Also, in social commitment adoption, usually agent i takes on a social commitment $\mathrm{COMM}(i, j, \alpha)$ *because* the other agent is interested in its executing α; here, such causality is not reflected in the definition above (see [5] for a recent discussion).

Social commitment obeys positive introspection, i.e.

$$\mathrm{COMM}(i, j, \varphi) \rightarrow \mathrm{BEL}(i, \mathrm{COMM}(i, j, \varphi)).$$

This follows from the awareness condition included in the defining axiom itself. It is not possible to derive negative introspection, because agents are in general not aware of the absence of collective beliefs (i.e. $\neg\text{C-BEL}_G(\varphi) \rightarrow \mathrm{BEL}(i, \neg\text{C-BEL}_G(\varphi))$ is not provable for $i \in G$).

3.6. Collective intentions

In our approach, teams are created on the basis of *collective intentions*, and exist as long as the collective intention between team members exists. A collective intention may be viewed as an inspiration for team activity. Collective intention and collective commitment are not introduced as primitive modalities, with some restrictions on the semantic accessibility relations (as in e.g. [7]). We do give necessary and sufficient, but still minimal, conditions for such collective motivational attitudes to be present. In this way, we hope to make the behavior of a team easier to predict.

In this paper, we focus on strictly cooperative teams, which makes the definition of collective intention rather strong. In such teams, a necessary condition for a collective intention $\text{C-INT}_G(\varphi)$ is that all members of the team G have the associated individual intention $\mathrm{INT}(i, \varphi)$ towards the goal φ. However, to exclude the case of competition, all agents should also *intend* all members to have the associated individual intention, and the intention that all members have the individual intention, and so on; we call such a mutual intention $\text{M-INT}_G(\varphi)$. Thus, $\text{M-INT}_G(\varphi)$ is meant to be true if everyone in G intends φ ($\text{E-INT}_G(\varphi)$), everyone in G intends that everyone in G intends φ ($\text{E-INT}_G(\text{E-INT}_G(\varphi))$), etc.

The distinguishing features of collective intentions ($\text{C-INT}_G(\varphi)$) over and above mutual ones, is that all members of the team are aware of the mutual intention, that is, they have a collective belief about this ($\text{C-BEL}_G(\text{M-INT}_G(\varphi))$). The above conditions are captured by the following axioms:

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

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

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

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

Note that this definition of collective intention **M3** is stronger than the one given in our older work [15, 16]. In [19], we extensively discuss the axioms above, provide them with a Kripke semantics and a completeness proof, and compare them with alternatives such as joint intention theory and SharedPlans theory [29, 23, 40]. 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: an initiator persuades all potential team members to adopt a mutual intention, and, if successful, announces that the mutual intention is established [10, 11].

It is easy to see that once a collective intention is established, agents are aware of it:

Lemma 3.2.

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

3.6.1. Different kinds of collective intention

In addition to collective intention as defined above, one could define collective intentions that are suitable for specific environments. In circumstances where communication is hampered but cooperative action is vital, agents must sometimes make do with a less strong version of collective intention, which does not include collective belief about the mutual intention, but instead only a mutual intention to establish it. In [19] we gave an alternative definition of mutual intention $\text{M-INT}'_G$ with a corresponding definition of collective intention appropriate for such cases in which communicative possibilities are severely limited. Of course these two definitions do not exhaust the range of possibilities and one could design a tuning machine for collective intentions just as we do in section 4 for collective commitments, but in this paper we do not delve further into this subject.

4. Tuning machine for collective commitment

After a group is constituted, another stage of CPS is started, namely plan formation, leading ultimately to a *collective commitment* between the team members. While a collective intention may be viewed as an inspiration for team activity, the collective commitment reflects the concrete manner of achieving the intended goal by the team. This concrete manner is provided by planning, and hinges on the allocation of actions according to an adopted plan. This allocation is concluded when agents accept pairwise (i.e. social) commitments to realize their individual actions. This way, our approach to collective commitments is plan-based.

While investigating calibration of commitment to the particular purpose and the specific circumstances, we isolated and separately characterized the following ingredients of collective commitments:

- collective intention on which the team is built,
- degrees of belief in a team,
- different aspects of team awareness.

They may be viewed as three types of 'dials' that are separately tuned in order to obtain a situation-sensitive notion of collective commitment of a desired strength. Before treating these 'dials' separately, we give a general schema for defining collective commitments. This generic schema together with a tuning mechanism may be viewed as a sort of *tuning machine* for creating collective commitments.

4.1. General schema of collective commitment

In our generic description we will solely define the basic ingredients constituting collective commitments, leaving room for case-specific extensions. The obligatory ingredients are related to different aspects of teamwork:

1. Mutual intention $\text{M-INT}_G(\varphi)$ between a group of agents, allowing them to act as a team. (See subsection 3.6 for a formal definition and discussion.)

   Let us stress the crucial role of mutual intention when creating a group: the team is *based* on this attitude, and exists as long as the mutual intention between team members exists. Thus, no teamwork is considered without a mutual intention among team members.

2. Social plan P on which a collective commitment will be based. (See subsection 3.1 for an example.)

   The social plan provides a concrete manner for the team to collectively achieve the overall goal of the system, the object of their mutual intention.

3. Pairwise social commitments $\text{COMM}(i, j, \alpha)$ for actions occurring in the social plan. (For a definition of social commitments, see subsection 3.5.)

   The group splits the tasks according to their social plan, and each agent takes on responsibility to do its part by accepting relevant social commitments.

Next to the above ingredients, different degrees of awareness about them may be present in a team. This may vary from the lack of any awareness to collective belief about the given aspect, as was discussed in subsection 3.2. Let us write $awareness_G(\psi)$ for "group G is aware that ψ". Thus, a general schema covering different types of collective commitment is the following, where the conjuncts between curly brackets may be present or not, according to the position of the awareness 'dial' :

$$\begin{aligned}
&\text{C-COMM}_{G,P}(\varphi) \leftrightarrow \\
&\quad \text{M-INT}_G(\varphi) \ \wedge \{\ awareness_G(\text{M-INT}_G(\varphi))\} \wedge \\
&\quad constitute(\varphi, P) \wedge \{awareness_G(constitute(\varphi, P))\} \wedge \\
&\quad \bigwedge_{\alpha\in P} \bigvee_{i,j\in G} \text{COMM}(i, j, \alpha) \ \wedge\ \{awareness_G(\bigwedge_{\alpha\in P} \bigvee_{i,j\in G} \text{COMM}(i, j, \alpha))\}
\end{aligned}$$

In words, group G has a collective commitment to achieve overall goal φ based on social plan P ($\text{C-COMM}_{G,P}(\varphi)$) iff all of the following hold. The group mutually intends φ (with or without being aware); moreover, successful execution of social plan P leads to φ (with or without the group being aware of this); and finally, for every one of the actions α that occur in social plan P, there should be one agent in the group who is socially committed to at least one (mostly other) agent in the group to fulfil the action (with or without the group being aware of this).

Instantiating the above schema corresponds to tuning the $awareness_G$-dials from $\emptyset$, through individual beliefs and different degrees of E-BEL_G^k, to collective belief, and analogously for degrees of knowledge. These degrees have been discussed in subsection 3.2. Now, we turn to a more detailed description of the *aspects* of teamwork that a group is aware of, as given by the three conjuncts in curly brackets presented in the general schema above.

4.2. Different aspects of agents' awareness

The notion of collective commitment, whichever strength of it is considered, combines essentially different aspects of teamwork: strictly technical ones related to social plans, as well as those related to agents' intentional stance. The latter concern different aspects of awareness that appear in a group of agents in the course of CPS. The degree of this awareness, characterized in terms of different types of beliefs, may be different. Below, only the strongest version is considered, namely collective belief about the relevant aspect of CPS. Thus, the $awareness_G$-dial is set to C-BEL_G in all bracketed conjuncts of the general schema of section 4.1. For this reason it is justified to speak about *collective awareness* in this context. In other circumstances, the degree of awareness can be weakened by using E-BEL_G (or another E-BEL_G^k) instead of C-BEL_G. Let us discuss the relevant aspects in detail.

1. Collective intention is the attitude constituting the group as a whole. Thus, it introduces (rather strong) collective awareness of the group as a cooperative team of agents. Formally this is expressed as a conjunct in the definition of collective intention:

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

2. When a team of agents exists, the next step is plan generation or adoption. Regardless of the method of arriving at this point, the type of awareness connected with this is collective awareness of the correctness of the plan with respect to the overall goal. Formally:

   $\text{C-BEL}_G(constitute(\varphi, P))$

3. When a plan as a recipe is in place, then the particular actions from it need to be allocated to particular team members in order to create pairwise social commitments between them. This way a social structure is built within a team, and the plan acquires the property of being social. The type of awareness connected with this phase may be twofold. We make an even more subtle distinction here than in the general schema, because we believe that it corresponds to an important difference between two types of teams that could not be distinguished by their motivational attitudes before:

   (a) The first one is a collective awareness of the social structure in a team with respect to a given plan. This includes a *detailed* awareness of each social commitment involved. Formally:
   $\bigwedge_{\alpha\in P} \bigvee_{i,j\in G} \text{C-BEL}_G(\text{COMM}(i,j,\alpha))$
   This corresponds to the interpretation *de re*.

(b) The second one refers to the detailed collective awareness about the plan, but a more *global* collective awareness of the social structure within the team, namely of the bare existence of social commitments with respect to the social plan. Formally:
$\text{C-BEL}_G(\bigwedge_{\alpha \in P} \bigvee_{i,j \in G} \text{COMM}(i, j, \alpha))$
This corresponds to the interpretation *de dicto*.

The distinction de re / de dicto stems from the philosophy of language [33]. A sentence of the form $\exists x \text{BEL}(j, A(x))$ is a *de re* belief attribution which relates agent j to a *res*, an individual that the belief is about. On the other hand, $\text{BEL}(j, \exists x A(x))$ is a *de dictum* belief attribution, relating agent j to a *dictum*, namely the proposition $\exists x A(x)$. This distinction is also fruitful for complex epistemic operators such as collective belief. Note that C-BEL_G in (a) and (b) distributes over conjunction ($\bigwedge_{\alpha \in P}$), so that only the position of C-BEL_G with respect to $\bigvee_{i,j \in G}$ matters. We give a small lemma about the relation between the two types of awareness:

Lemma Detailed awareness ($\bigwedge_{\alpha \in P} \bigvee_{i,j \in G} \text{C-BEL}_G(\text{COMM}(i, j, \alpha))$) implies global awareness ($\text{C-BEL}_G(\bigwedge_{\alpha \in P} \bigvee_{i,j \in G} \text{COMM}(i, j, \alpha))$), but not vice versa.

Proof We work in a system that includes $KD45_n^C$ for individual and collective belief. Let us reason semantically.

Suppose

$$\mathcal{M}, w \models \bigwedge_{\alpha \in P} \bigvee_{i,j \in G} \text{C-BEL}_G(\text{COMM}(i, j, \alpha)).$$

Now take any $\alpha \in P$, then there is a pair $i, j \in G$ such that

$$\mathcal{M}, w \models \text{C-BEL}_G(\text{COMM}(i, j, \alpha)),$$

so a fortiori, by propositional logic and collective belief distribution,

$$\mathcal{M}, w \models \text{C-BEL}_G(\bigvee_{i,j \in G} \text{COMM}(i, j, \alpha)).$$

As $\alpha \in P$ was arbitrary, we may conclude

$$\mathcal{M}, w \models \bigwedge_{\alpha \in P} \text{C-BEL}_G(\bigvee_{i,j \in G} (\text{COMM}(i, j, \alpha))),$$

which is equivalent to

$$\mathcal{M}, w \models \text{C-BEL}_G(\bigwedge_{\alpha \in P} \bigvee_{i,j \in G} (\text{COMM}(i, j, \alpha))),$$

because in general

$$KD45_n^C \vdash \text{C-BEL}_G(\psi_1 \wedge \ldots \wedge \psi_n) \leftrightarrow \text{C-BEL}_G(\psi_1) \wedge \ldots \wedge \text{C-BEL}_G(\psi_n).$$

The converse does not hold. Take e.g. group $G = \{1, 2, 3\}$ and plan $P = \langle\langle\langle 1, a\rangle; \langle 2, b\rangle\rangle; \langle 3, c\rangle\rangle$ where each social commitment is collectively believed only by its two participants but not by the group as a whole, even if there is collective awareness that for each action, some social commitment is in place. In such a case, there is global awareness without local awareness about the three social commitments:

$$\begin{aligned}\mathcal{M}, w \models\ & \mathrm{COMM}(1,2,a) \wedge \mathrm{COMM}(2,3,b) \wedge \mathrm{COMM}(3,1,c) \wedge \\ & \neg\text{C-BEL}_G(\mathrm{COMM}(1,2,a)) \wedge \neg\text{C-BEL}_G(\mathrm{COMM}(2,3,b)) \wedge \\ & \neg\text{C-BEL}_G(\mathrm{COMM}(3,1,c)) \wedge \text{C-BEL}_G(\bigwedge_{\alpha \in P} \bigvee_{i,j \in G} (\mathrm{COMM}(i,j,\alpha)))\end{aligned}$$

The above aspects of awareness will be viewed as building blocks when distinguishing different strengths of collective commitments.

5. Different notions of collective commitment

In order to make the theory of collective commitments more concrete, we will now instantiate the general schema (and its more refined variant using detailed and global awareness as defined above) in five different ways. All of these lead to types of group commitments actually occurring in different organization types in practice, as we will illustrate by example organizations.

The following exemplar definitions are produced by keeping the $awareness_G$-dial fixed to a choice between $\emptyset$ and collective belief, and the dial for 'kind of mutual intention' fixed as the standard definition of subsection 3.6. We will start from the strongest form of collective commitment: its expressive power fully reflects the collective aspects of CPS. Subsequently, some of the underlying assumptions will be relaxed, leading ultimately to weaker notions of team and distributed commitment.

5.1. Robust collective commitment

Our discussion on different types on collective commitments will start from the two strongest cases based on collective planning, including negotiating and persuading each other who will do what.

Robust collective commitment is the strongest type that one can make on the basis of collective beliefs as awareness type. When instantiating the general schema of subsection 4.1, all bracketed conjuncts are instantiated by putting the $awareness_G$-dial to C-BEL$_G$, and in addition, for the last conjunct, version 3a from subsection 4.2 is chosen, namely detailed (or de re) collective awareness.

This means intuitively that, in addition to collective planning, for every one of the actions α that occur in social plan P, there should be one agent in the group who is socially committed to at least one (mostly other) agent in the group to fulfil the action. Moreover, the team as a whole is aware of every single social commitment ($\mathrm{COMM}(i,j,\alpha)$) that has been established about particular actions from the social plan. All these characteristics lead to the following definition of *robust collective commitment* (R-COMM$_{G,P}$):

$$\begin{aligned}& \text{R-COMM}_{G,P}(\varphi) \leftrightarrow \text{C-INT}_G(\varphi) \wedge \\ & \quad constitute(\varphi, P) \wedge \text{C-BEL}_G(constitute(\varphi, P)) \wedge \\ & \quad \bigwedge_{\alpha \in P} \bigvee_{i,j \in G} \mathrm{COMM}(i,j,\alpha) \wedge \bigwedge_{\alpha \in P} \bigvee_{i,j \in G} \text{C-BEL}_G(\mathrm{COMM}(i,j,\alpha))\end{aligned}$$

By the last conjunct, everybody's responsibility is public. The aspect of sharing responsibility is of crucial importance here. Among others it implies that there is no need for an initiator in such a team. There is detailed (vs. global) collective awareness of social commitments.

Example Robust collective commitment may be applicable in (small) companies where all team members involved are share-holders. Typically, planning is done collectively, whether from first principles or choosing from a plan library. Everybody's responsibility is public, because the social commitments are established publicly. In particular, when any form of revision is needed due to dynamic circumstances, the entire team may be collectively involved. This type of collective commitment is also the one most suited for self-leading teams, which are not directly led by a manager. Instead the team is responsible for achieving some high-level goals, and is entirely free to divide roles, devise a plan, etc. [2]. The non-hierarchical team of researchers introduced by Example 1 in the introduction and discussed further in subsection 3.1 is a typical example of such a self-leading team establishing a robust collective commitment.

5.2. Strong collective commitment

Just as in the case of robust collective commitment, when instantiating the general schema of subsection 4.1 for *strong collective commitment*, all bracketed conjuncts are instantiated by putting the $awareness_G$-dial to C-BEL_G, however, for the last conjunct, version 3b from subsection 4.2 is chosen, namely global (or de dicto) collective awareness. This makes it somewhat weaker than robust collective commitment.

Thus, in contrast to robust collective commitment, in the case of strong collective commitment ($\text{S-COMM}_{G,P}$), there is no detailed public awareness about particular social commitments, but the group as a whole believes that things are under control, i.e., that every part of the plan is within somebody's responsibility:

$$\begin{aligned}&\text{S-COMM}_{G,P}(\varphi) \leftrightarrow \text{C-INT}_G(\varphi) \wedge \\ &\quad constitute(\varphi, P) \wedge \text{C-BEL}_G(constitute(\varphi, P)) \wedge \\ &\quad \bigwedge_{\alpha \in P} \bigvee_{i,j \in G} \text{COMM}(i, j, \alpha) \wedge \text{C-BEL}_G(\bigwedge_{\alpha \in P} \bigvee_{i,j \in G} \text{COMM}(i, j, \alpha))\end{aligned}$$

As the responsibility is not shared due to the lack of detailed awareness in the last conjunct, the case of a team leader or initiator fits here. Also, as pair-wise social commitments are not collectively known, they cannot be collectively revised when such a need appears.

Example Strong collective commitment may be applicable in companies with one or more leaders and rather separate sub-teams. Typically, planning is done collectively. However, establishing bilateral commitments is not done publicly in the whole team, but in subgroups, for example, members might promise the leader of their sub-team that they will do their own part. Sometimes this global awareness of social commitments suffices, and this may be preferable in order not to waste energy or communication resources.

5.3. Weak collective commitment

In a somewhat weaker case of collective commitment, the degree of team awareness is even more limited. When the plan as a whole is not known to the team and no collective decision making is assumed, there is no awareness in the team that the plan leads to proper realization of the goal (C-BEL$_G$ $constitute(\varphi, P)$ is not in place). We deal with a *weak collective commitment* (W-COMM$_{G,P}$). Formally, weak collective commitments are distinguished from strong ones by instantiating the second bracketed conjunct of the general schema with the $awareness_G$-dial set at $\emptyset$:

$$\text{W-COMM}_{G,P}(\varphi) \leftrightarrow \text{C-INT}_G(\varphi) \wedge constitute(\varphi, P) \wedge$$
$$\bigwedge_{\alpha \in P} \bigvee_{i,j \in G} \text{COMM}(i, j, \alpha) \wedge \text{C-BEL}_G(\bigwedge_{\alpha \in P} \bigvee_{i,j \in G} \text{COMM}(i, j, \alpha))$$

In this case, the team knows the overall goal, but does not know details of the plan: there is no collective awareness of the plan's correctness. Apparently, also in this case no collective revision of social commitments may take place.

Example Weak collective commitment may be applicable in companies with a dedicated planner or planning department. Typically, the planner individually believes the plan to be correct $constitute(\varphi, P)$, and this may suffice. Concrete examples of such companies are large multi-nationals with extensive planning departments.

Remark about robust, strong and weak commitments The above three versions of collective commitment implicitly reflect different social structures in which a group is involved. Social structure originates from different types of hierarchies, based on power relations and dependency relations [6]. These relations are implicitly reflected in the set of social commitments included in the collective commitment by the conjunct $\bigwedge_{\alpha \in P} \bigvee_{i,j \in G} \text{COMM}(i, j, \alpha)$. For example, in a hierarchical tree structure, social commitments are made to the agent that is the direct ancestor in the tree.

In the case of weak collective commitments, it is not collectively believed that the social plan realizes the overall goal, though the members are aware about their share in it and the fact that all actions are taken on by committed members; in the strong and robust cases, the team is also aware that the plan as a whole assures proper realization of the goal φ.

Apparently, there are also other possibilities of group involvement in CPS, which we define below. Note that agents' limited orientation in the task and action distribution may be done on purpose, even though the overall goal is known to everybody, according to the definition of collective intention.

5.4. Team commitment

In the case of *team commitment* (T-COMM$_{G,P}$) agents remain aware solely about their piece of work, without any orientation about involvement of others, except regarding their collective intention to achieve the main goal. In this situation, there is no collective belief that all actions have been adopted by other committed members, but a team as a structure still exists. Thus, formally, in team commitments only the $awareness_G$-dial for the first bracketed conjunct of the general schema of subsection 4.1 is set to C-BEL$_G$, while both others are set to $\emptyset$:

$$\text{T-COMM}_{G,P}(\varphi) \leftrightarrow \text{C-INT}_G(\varphi) \wedge constitute(\varphi, P) \wedge$$

$$\bigwedge_{\alpha \in P} \bigvee_{i,j \in G} \text{COMM}(i, j, \alpha)$$

Because of the presence of collective intention, the overall goal and composition of the team are collectively believed. Planning is not at all collective: it may be that even task division is not public; this is often done on purpose. Thus, distribution of social commitments cannot be public either.

Example Team commitment may be applicable in companies assigning limited trust to their employees. Information about the precise involvement of colleagues and other aspects of the plan may be confidential.

5.5. Distributed commitment

The last case distinguished here is *distributed commitment* ($\text{D-COMM}_{G,P}$). It deals with the situation when agents' awareness is even more restricted: they may not even know the overall goal, only their share in an 'undefined' project. Formally, the $awareness_G$-dials for all three bracketed conjuncts of the general schema are set to $\emptyset$:

$$\text{D-COMM}_{G,P}(\varphi) \leftrightarrow \mathit{constitute}(\varphi, P) \wedge \bigwedge_{\alpha \in P} \bigvee_{i,j \in G} \text{COMM}(i, j, \alpha).$$

This means that no 'real' team of cooperating agents is created, so that no collective intention C-INT_G is in place. Instead, a rather loosely coupled group of agents works in a distributed manner without autonomous involvement in the project to be realized.

Example Distributed commitment may be applicable in companies contracting out some labour to outsiders. The overall goal and the group of agents involved may be classified information, for example in order to avoid competition.

Another typical case of distributed commitment is displayed by groups of spies as introduced by Example 2 in the introduction. In their case, lack of information about the tasks or even the identity of other group members may be beneficial to everybody's safety. They work with an inflexible plan set in advance by one mastermind, so their autonomy and flexibility are severely curtailed. On the positive side, the need for communication before and during group action is limited as well.

General remarks about the five types of group commitment The weaker notions such as team commitment and distributed commitment (when a team does not exist as a whole) are especially suited to model hierarchically organized teams, where power relations between team members play a role. The simplest case of agents' organization is teamwork completely controlled by an initiator agent. Though we refrain from introducing the power aspect explicitly in the definitions, their different strengths may be useful in various situations (see [6]), especially when maintaining a balance between the centralized power and the spread of knowledge.

The stronger notions like robust collective commitment and strong collective commitment are well-suited to model so-called self-leading teams which are currently studied in the organizational science literature [2]. In the strongest versions of the definition, all agents involved are collectively aware of the situation as a whole, as is reflected in the following theorems:

Theorem 5.1. (awareness of robust collective commitment)

$$\text{R-COMM}_{G,P}(\varphi) \rightarrow \text{C-BEL}_G(\text{R-COMM}_{G,P}(\varphi)).$$

Theorem 5.2. (awareness of strong collective commitment)

$$\text{S-COMM}_{G,P}(\varphi) \to \text{C-BEL}_G(\text{S-COMM}_{G,P}(\varphi)).$$

The proofs are immediate from the definitions and lemma 3.2. Note that the theorem does not hold for weaker forms of group commitments. In these cases agents are not necessarily aware of the strength (kind) of group commitment between them. Note also, that some intermediate levels of commitment may be characterized by replacing C-BEL_G by E-BEL_G in some contexts. The motivation behind this is rather clear: in some cases only E-BEL_G can be achieved. These situations, however, will not be discussed in this paper.

6. Possible generalizations

The presented definitions of (collective) commitments enable to organize teams or larger organizational structures according to a specific (chosen) type of commitment. However, in real applications much more complex and/or distributed structures may be considered: sub-teams of agents, created on the basis of various commitments, may be combined into larger structures. Thus, heterogeneity of these structures is achieved. In order to cover the variety of possibilities, potential 'ties' in these complex organizations may be implemented in many different ways. One of them would be introducing an organization's social structure *explicitly*, for example by a labeled tree, in contrast to the *implicit* form adopted in the definitions of collective commitments presented here. An explicit social structure has many advantages, and may be applicable in a variety of situations. Most essentially, it gives an opportunity to appropriately organize various substructures within a complex framework. Thus, *scalability* of these organizations comes to the fore, although this problem itself has not yet been addressed in the research on motivational attitudes. Nevertheless, when using such an explicit framework, specification of truly large organizations is possible, making them easier to predict. We plan such an investigation as our next research subject.

7. Discussion and conclusions

We have incrementally built a static theory of CPS, starting from individual intentions, through social commitments, leading ultimately to collective intentions and collective commitments. All these notions are defined in multi-modal logics with clear semantics (cf. [19]), comprising a descriptive view on collective commitments. In contrast to [15], we do not give one iron-clad definition of collective commitment here. Instead, we provide a sort of tuning mechanism for the system developer to calibrate an appropriate type of collective commitment, taking into account both the circumstances in which a group is acting, for example possibilities of communication, as well as organizational structure. The multi-modal logic framework allows to express subtle aspects of CPS, modeling different situations occurring in practical domains.

The presented system, containing logics of collective beliefs and collective intentions, is known to be EXPTIME-hard. Therefore, 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. [25, 26, 38]). 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. [22, 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 [8], branching time as in Rao and Georgeff's work [34], and alternating-time temporal logic (ATL) as in [27]. The definitions of collective commitments in terms of more basic attitudes, as presented in this paper, may be combined with either choice, depending on the application.

The definitions of collective commitments are not overloaded, and therefore easy to understand and to use. Some other approaches to collective commitments (see e.g. [29, 41]) introduce other aspects of collective attitudes, not treated here. For example, Wooldridge and Jennings consider triggers for commitment adoption formulated as preconditions [41]. As another example, Aldewereld, Van der Hoek and Meyer add constraints about goal adoption and achievement to their definitions of joint motivational attitudes [1]. If needed, these extensions may be incorporated into our framework as well by adding extra axioms. Note that in contrast to other approaches ([41],[29]), the collective commitment is not iron-clad: it may vary in order to adapt to changing circumstances, in such a way that the collective intention on which it is based can still be reached.

In the present paper, we do not describe how collective intentions, and then collective commitments, are actually established in a group. This important aspect has been extensively treated in [11, 18]. There, the whole process of dialogue among computational agents involved in CPS is made transparent, including the effects of utterances on agents' individual mental states and on their collective attitudes.

Our approach is especially strong when re-planning is needed. In contrast to [41], using our definitions of collective commitment it is often sufficient to revise some of the pair-wise social commitments, instead of involving the entire team in the re-planning process (in the strong versions of the definition). This is a consequence of basing collective commitment on an explicitly represented plan, and of building it from pair-wise social commitments. In effect, if the new plan resulting from the analysis of the current situation within the team and the environment is as close as possible to the original one, the process of re-planning is maximally efficient. This reconfiguration problem was treated extensively in [17], where an abstract reconfiguration algorithm was presented. In current work, we formally specify situations in which agents' collective attitudes change [20]. This contributes to the *dynamic*, more prescriptive theory of collective intentions. Combining the static and dynamic aspects, the full theory may serve a system designer as a specification to create a correct system, as well as to verify it.

Acknowledgments

We would like to thank Cristiano Castelfranchi, Keith Clark, Rino Falcone, and Andrew Jones for fruitful discussions about this work. We also thank Mike Luck, whose remarks about a previous version of this paper we found very useful. This work is supported by the Polish KBN Grant 7T11C 006 20 and by the EU funded ALFEBIITE++ project.

8. Appendix: proofs of correspondences

The interdependency axioms in subsection 3.4 correspond to semantic properties of the underlying Kripke frames as follows.

Fact 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 $\mathbf{A8}_{GB}$.

Proof 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 $\mathcal{M}, s \models \neg\mathrm{INT}(i, \varphi)$, then there is a $t \in W$ with sI_it and $\mathcal{M}, t \not\models \varphi$. We will show that $\mathcal{M}, s \models \mathrm{BEL}(i, \neg\mathrm{INT}(i, \varphi))$. So take any $u \in W$ such that sB_iu. By the condition on the frame, we have uI_it, so $\mathcal{M}, u \models \neg\mathrm{INT}(i, \varphi)$, and indeed $\mathcal{M}, s \models \mathrm{BEL}(i, \neg\mathrm{INT}(i, \varphi))$. Therefore, $F \models \neg\mathrm{INT}(i, \varphi) \rightarrow \mathrm{BEL}(i, \neg\mathrm{INT}(i, \varphi))$.

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 $\mathcal{M}, s \models \neg\mathrm{INT}(i, p)$. On the other hand, $\mathcal{M}, u \models \mathrm{INT}(i, p)$, so $\mathcal{M}, s \not\models \mathrm{BEL}(i, \neg\mathrm{INT}(i, p))$. We may conclude that $F \not\models \neg\mathrm{INT}(i, p) \rightarrow \mathrm{BEL}(i, \neg\mathrm{INT}(i, p))$.

Fact The corresponding semantic property corresponding to $\mathbf{A9}_{IG}$ is that $G_i \subseteq I_i$.

Proof 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 M be the Kripke model arising from F by adding *Val*. Now take any $s \in W$ with $M, s \models \mathrm{INT}(i, \varphi)$, but suppose, in order to derive a contradiction, that $M, s \not\models \mathrm{GOAL}(i, \varphi)$. Then there is a $t \in W$ with sG_it and $M, t \not\models \varphi$. But because $G_i \subseteq I_i$ we have sI_at as well, contradicting the assumption $M, s \models \mathrm{INT}(i, \varphi)$. Therefore, $F \models \mathrm{INT}(i, \varphi) \rightarrow \mathrm{GOAL}(i, \varphi)$.

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 M be the Kripke model arising from F by adding *Val*. Then by definition $M, s \models \mathrm{INT}(i, p)$; but $M, t \not\models p$, so $M, u \not\models \mathrm{GOAL}(i, p)$. We may conclude that $F \not\models \mathrm{INT}(i, p) \rightarrow \mathrm{GOAL}(i, p)$.

References

[1] Aldewereld, H., van der Hoek, W., Meyer, J.-J.: Rational Teams: Logical Aspects of Multi-Agent Systems, *Fundamenta Informaticae*, **this issue**, 2004.

[2] Beyerlin et al., M., Ed.: *Theories of Self-managing Work Teams*, JAI Press, Greenwich (CN), 1994.

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

[4] Castelfranchi, C.: Commitments: From Individual Intentions to Groups and Organizations, in: Lesser [28], 41–48.

[5] Castelfranchi, C.: *Grounding We-Intentions in Individual Social Attitudes: On Social Commitment Again*, Technical report, CNR, Institute of Psychology, 1999, Manuscript.

[6] Castelfranchi, C., Miceli, M., Cesta, A.: Dependence Relations Among Autonomous Agents, in: Werner and Demazeau [39].

[7] 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.

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

[9] 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.

[10] Dignum, F., Dunin-Kęplicz, B., Verbrugge, R.: Agent Theory for Team Formation by Dialogue, *Intelligent Agents VII: Agent Theories, Architectures and Languages* (C. Castelfranchi, Y. Lesperance, Eds.), 1986, Springer Verlag, Berlin, 2001.

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

[12] Dunin-Kęplicz, B., Radzikowska, A.: Actions with Typical Effects: Epistemic Characterization of Scenarios, in: Lesser [28], page 445.

[13] Dunin-Kęplicz, B., Radzikowska, A.: Epistemic Approach to Actions with Typical Effects, *Proceedings ECSQARU'95*, Fribourg, 1995.

[14] Dunin-Kęplicz, B., Radzikowska, A.: Modelling Nondeterminstic Actions with Typical Effects, *Proceedings DIMAS'95*, Cracow, 1995.

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

[16] 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.

[17] Dunin-Kęplicz, B., Verbrugge, R.: A Reconfiguration Algorithm for Distributed Problem Solving, *Engineering Simulation*, **18**, 2001, 227 – 246.

[18] 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.

[19] Dunin-Kęplicz, B., Verbrugge, R.: Collective Intentions, *Fundamenta Informaticae*, **51(3)**, 2002, 271–295.

[20] Dunin-Kęplicz, B., Verbrugge, R.: Evolution of Collective Commitments During Teamwork, *Fundamenta Informaticae*, **56**, 2003, 329–371.

[21] Dunin-Kęplicz, B., Verbrugge, R.: A tuning machine for collective commitments, *Proceedings of The First International Workshop on Formal Approaches to Multi-Agent Systems* (B. Dunin-Kęplicz, R. Verbrugge, Eds.), Warsaw, 2003.

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

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

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

[25] 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.

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

[27] Jamroga, W., van der Hoek, W.: Agents that Know how to Play, *Fundamenta Informaticae*, **this issue**, 2004.

[28] Lesser, V., Ed.: *Proceedings First International Conference on Multi-Agent Systems*, AAAI-Press and MIT Press, San Francisco, 1995.

[29] 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.

[30] Luck, M., McBurney, P., Preist, C.: *Agent Technology: Enabling Next Generation Computing: A Roadmap for Agent Based Computing*, Agentlink, 2003.

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

[32] Parikh, R., Krasucki, P.: Levels of Knowledge in Distributed Computing, *Sadhana: Proceedings of the Indian Academy of Sciences*, **17**, 1992, 167–191.

[33] Quine, W.: Quantifiers and Propositional Attitudes, *Journal of Philosophy*, **53**, 1956, 177–187.

[34] 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.

[35] Rao, A., Georgeff, M., Sonenberg, E.: Social Plans: A Preliminary Report, in: Werner and Demazeau [39], 57–76.

[36] Searle, J. R.: *Speech Acts*, Cambridge University Press, Cambridge, 1969.

[37] Segerberg, K.: Bringing it About, *Journal of Philosophical Logic*, **18**, 1989, 327–347.

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

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

[40] Wooldridge, M.: *Reasoning About Rational Agents*, MIT Press, Cambridge, MA, 2000.

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