# A logical view on teamwork

Barbara Dunin-Kęplicz[1,2] and Rineke Verbrugge[3]

[1] Institute of Informatics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland
[2] Institute of Computer Science, Polish Academy of Sciences, Ordona 21, 01-237 Warsaw, Poland, keplicz@mimuw.edu.pl
[3] Department of Artificial Intelligence, Faculty of Mathematics and Natural Sciences, University of Groningen, PO Box 407, 9700 AK Groningen, The Netherlands, rineke@ai.rug.nl

**Abstract.** This chapter presents the non-dynamic part of a formal framework for teamwork in multi-agent systems. The framework consists of both *a static part*, defining collective motivational attitudes in such a way that the system developer can adapt them to the circumstances, and *a dynamic part* monitoring the changes in team attitudes during the course of cooperative problem solving (CPS).
In the first part of this chapter, the notion of *collective intention* in teams of agents is investigated. Starting from individual *intentions*, *goals*, and *beliefs* defining agents' local attitudes, we arrive at an understanding of collective intention in cooperative teams as a rather strong concept: it implies that all members *intend* for all others to share that intention. This way a team is glued together by collective intention, and exists as long as this attitude holds, after which it may disintegrate.
Collective intentions are formalized in a multi-modal logical framework. Together with individual and common knowledge and/or belief, collective intention constitutes a basis for preparing a plan, reflected in the strongest attitude, i.e., in *collective commitment*, defined and investigated in the next part. Distinct versions of collective commitments that are applicable in various situations, differ with respect to the *aspects* of teamwork of which the agents involved are aware, and the *kind* of awareness present within a team. This way a kind of tuning mechanism is provided for the system developer to tune a version of collective commitment fitting the circumstances. Finally, a few exemplar versions of collective commitment resulting from instantiating the general tuning scheme are presented.

## 1 Teamwork: from biology to organizational psychology

Teamwork can be seen as the pinnacle of cooperative activity: a group works together to achieve a collaborative goal that is difficult or impossible to attain individually. In the furtherance of this goal, agents need to work together in dynamic and flexible ways that also call for explicit communication and reasoning about the team's aims, plans, and progress. It seems that, in contrast to the animal kingdom, this ability of explicit reasoning about teamwork enables humans

to achieve much more flexible kinds of cooperative problem solving than that expressed by, for example, animals hunting in groups [1]. Teamwork has arisen during evolution through a number of precursors, weaker forms of cooperation that each place their own cognitive and communicative demands on the participants. An interesting overview of this evolution of increasingly intricate forms of cooperation is given in the chapter by Gärdenfors [2].

Human teamwork has become more and more complex, especially over the last few decades. Pressures like increasing competition and need for fast innovation call for teams with high levels of diverse skills, an ability for rapid response, and adaptability. Organizational psychologists have made many empirical investigations into what is a team, which factors make teams successful, and which interventions help to make teams more effective [3–6]. In Kozlowski and Ilgen's meta-review of teamwork research in organizational science, they provide the following definition of a *team* [6]:

> A team can be defined as (a) two or more individuals who (b) socially interact (face-to-face or, increasingly, virtually); (c) possess one or more common goals; (d) are brought together to perform organizationally relevant tasks; (e) exhibit interdependencies with respect to workflow, goals, and outcomes; (f) have different roles and responsibilities; and (g) are together embedded in an encompassing organizational system, with boundaries and linkages to the broader system context and task environment.

One of the determinants of good teams that Kozlowski and Ilgen [6] discuss is the way in which information flows in a team. Depending on the context, it may be appropriate for agents to have fully *shared team mental models* (reminiscent of common knowledge); in other teams with a more complex workflow and more specialization, a more effective way to approach information may include so-called *transactive memory* (which is analogous to distributed knowledge, with the addition that team members should keep track of one another's domains of expertise)[4]. They name *cohesiveness* as another important element of team success, causing the group to remain united to reach their common goal.

Kozlowski and Ilgen note that many of their colleagues in organizational psychology blur the individual perspectives and the collective ones when studying team notions. We feel that a more formal model of teamwork, in which the individual and collective levels are clearly delineated and explicitly related to one another, would be helpful for organizational psychologists. Thus, the precise notion of collective intention discussed in this chapter may help in investigations of team cohesiveness. An example of such a formal model of teamwork is presented in this chapter. It has been devised in the context of multi-agent systems, with computational team members in mind, but we such a formal analysis might also help to analyze and improve human teamwork.

[4] Similar group notions of knowledge are discussed in [7]; there as well as in [6], we think that an explicit link to multi-agent epistemic logic would enable an even more perspicuous analysis of 'group cognition'.

## 2 Teamwork in multi-agent systems: chapter overview

In *multi-agent systems* (MAS) one of the central issues is the study of how groups work, and how the technology enhancing group interaction can be implemented. From the distributed Artificial Intelligence perspective, multi-agent systems are computational systems in which a collection of loosely-coupled autonomous agents interact in order to solve a given problem. As this problem is usually beyond the agents' individual capabilities, agents exploit their ability to *communicate*, *cooperate*, *coordinate*, and *negotiate* with one another. Apparently, the type of social interactions involved depends on circumstances and may vary from altruistic cooperation through to open conflict. A paradigmatic example of joint activity is *cooperative problem solving* (CPS) in which a group of autonomous agents choose to work together, both in advancement of their own goals as well as for the good of the system as a whole.

### 2.1 Intentional agent systems

Some MAS are referred to as *intentional systems*. In such systems, in order to give a representation of the mental states and cognitive processes, agents are represented as maintaining an intentional stance towards their environment ([8]). Such systems realize the *practical reasoning* paradigm ([9]) – 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* [10]. BDI-agents are characterized by a "mental state" described in terms of *beliefs*, corresponding to the information the agent has about the environment; *desires* or *goals*, representing options available to the agent, i.e. different states of affairs that the agent may choose to commit to; and *intentions*, representing the chosen options. Ultimately, in our approach, intentions are viewed as an inspiration for a goal-directed activity, reflected in commitments. While beliefs are viewed as the agent's *informational* attitudes, desires or goals, intentions, and commitments refer to its *motivational* attitudes. These collective notions are considered in the context of teams of agents, as defined in [11]:

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

*Collective intention*, as a specific joint mental attitude, is the central topic addressed in *teamwork*. We agree with [12] that:

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

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

### 2.2 From intentions to actions

We agree with Bratman [9] that in human practical reasoning, intentions are first class citizens, in the sense that they are not reducible to beliefs and desires. They form a rather special consistent subset of an agent's goals, that the agent wants to focus on for the time being. Thus they create a screen of admissibility for the agent's further, possibly long-term, deliberation. In contrast to [13], we are interested in generic characteristics of intentions, resigning from classifying them further along different dimensions. The most crucial aspect of intentions is that they are considered as an inspiration for goal-directed activity, reflected in the strongest motivational attitudes, that is in social (bilateral) and collective commitments. In our view, social commitments are related to individual actions, while collective commitments are related to plan-based team actions. The essential characteristics of both types of commitments is that they directly lead to action execution. Commitments will be treated in depth in Section 5.

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

Secondly, let us remember that a team of agents in a multi-agent system operates in a dynamic and often unpredictable environment. Such an environment poses the problem that team members may fail to bring their tasks to a good end or new opportunities may appear. We treated this *reconfiguration problem* in [14], where collective intentions are maintained by properly adapting collective commitments to the changing circumstances. This contributes to the *dynamic*, more prescriptive theory of collective intentions. The formal specification of situations in which agents' attitudes change is extensively discussed in [15] and [16, Ch. 6].

Section 4 falls squarely in the scope of the first objective to define a static theory of collective intentions. A definition of collective intention is presented, together with examples of situations in which it applies. The notion is formalized in a multi-modal logical framework. Completeness of this logic with respect to an appropriate class of Kripke models has been proved in [17]. Thus, a Computational Logic framework for specifying MAS involved in teamwork has been provided. It is assumed that the presented definitions reflect solely vital aspects of motivational attitudes, leaving room for case-specific extensions. This makes the framework flexible and not overloaded.

### 2.3 The role of intentions in practical reasoning

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

1. repeatedly updating beliefs about the environment;
2. deciding what options are available;

3. "filtering" these options to determine new intentions;
4. creating commitments on the basis of intentions;
5. performing actions in accordance with commitments.

Practical reasoning involves two important processes: deciding *what* goals need to be achieved, and then *how* to achieve them. The former process is known as *deliberation*, the latter as *means-end-reasoning*. A key concept in the theory of practical reasoning is that of *intention*. In [18] it is characterized from the psychological viewpoint:

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

We do not aim to present a psychologically sound theory of motivations driving human behaviour. We are interested in studying those motivational aspects that are involved in the *rational* decision making process. Thus, we disregard irrational drives and desires, which make human behaviour difficult to interpret and to predict. We do not consider any specific notion of *rationality*, in particular the economic one used in game theory [19, 20]. The only assumption made here is that agents are logical reasoners.

There is common agreement that intentions play a number of important roles in practical reasoning [9, 13]:

- 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 leads to the problem of balancing *pro-active*, (i.e. goal-directed) and *reactive* (i.e. event-driven) behaviour. We try to maintain this balance very carefully on both the individual and the collective level. The problem of persistence of intentions is expressed in an agent's *intention strategies*, addressing the question: *when and how can an agent responsibly drop its*

*intentions?* One answer to this question has been discussed for individual agents in [10]. The collective level is apparently more complex. Collective intention helps the team to monitor its behaviour during teamwork: even if some members drop their individual intentions, the team replans, aiming that the collective intention is ultimately realized. The precise description of this reconfiguration process can be found in [14, 15] and most notably in [16, Chapters 5 and 6], where an agent's pro-activeness and reactiveness are implicitly or explicitly involved in consecutive stages of the reconfiguration algorithm.

### 2.4 Calibrating team attitudes to the circumstances

Variety is the core of multi-agent systems. This statement expresses the many dimensions on which 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.

In this chapter we focus on collective motivational attitudes, i.e. intentions and commitments in cooperative problem solving. What characterizes collective notions is an interplay between environmental and social aspects, which may become rather complex nowadays due to the increasing complexity of MAS. For example, 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 Section 5 is to formally model different aspects of collective commitments, including different scopes and degrees of awareness of cooperating agents. (The case of competition is not included.) 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, characterized in the sequel, belong to the device representing a general schema of collective commitment. The resulting notion of (group) commitment, described in multi-modal logics, may then be naturally implemented in a multi-agent system. This way, the tuning mechanism may be viewed as a bridge between theory and practice.

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 in Subsection 5.3. Apparently, the entire spectrum of possibilities is much wider, due to the number of possibly independent choices to be made.

## 3 The logical background

As mentioned before, we propose the use of multi-modal logics to formalize agents' informational and motivational attitudes as well as actions they perform. Because in this chapter we restrict ourselves mostly to the static aspects of the agents' mental states, we present solely axioms relating agents' attitudes with respect to *propositions*, reflecting a particular state of affairs.

Table 1 gives some important formulas appearing in this chapter, together with their intended meanings. The symbol $\varphi$ denotes a proposition.

| | |
|---|---|
| $\text{BEL}(a, \varphi)$ | agent $a$ has the belief that $\varphi$ |
| $\text{E-BEL}_G(\varphi)$ | every agent in group $G$ has the belief that $\varphi$ |
| $\text{C-BEL}_G(\varphi)$ | group $G$ has the common belief that $\varphi$ |
| $\text{GOAL}(a, \varphi)$ | agent $a$ has the goal to achieve $\varphi$ |
| $\text{INT}(a, \varphi)$ | agent $a$ has the intention to achieve $\varphi$ |
| $\text{E-INT}_G(\varphi)$ | every agent in group $G$ has the intention to achieve $\varphi$ |
| $\text{M-INT}_G(\varphi)$ | group $G$ has the mutual intention to achieve $\varphi$ |
| $\text{C-INT}_G(\varphi)$ | group $G$ has the collective intention to achieve $\varphi$ |
| $\text{C-COMM}_{G,P}(\varphi)$ | group $G$ has a collective commitment to achieve $\varphi$ by plan $P$ |
| $\text{R-COMM}_{G,P}(\varphi)$ | group $G$ has a robust collective commitment to achieve $\varphi$ by plan $P$ |
| $\text{S-COMM}_{G,P}(\varphi)$ | group $G$ has a strong collective commitment to achieve $\varphi$ by plan $P$ |
| $\text{W-COMM}_{G,P}(\varphi)$ | group $G$ has a weak collective commitment to achieve $\varphi$ by plan $P$ |
| $\text{T-COMM}_{G,P}(\varphi)$ | group $G$ has a team commitment to achieve $\varphi$ by plan $P$ |
| $\text{D-COMM}_{G,P}(\varphi)$ | group $G$ has a distributed commitment to achieve $\varphi$ by plan $P$ |

**Table 1.** Formulas and their intended meaning

### 3.1 The logical language and semantics

The language $\mathcal{L}$ is based on a denumerable set $\mathcal{P}$ of *propositional symbols*, and a finite set $\mathcal{A}$ of *agents*, denoted by numerals $1, 2, \ldots, n$. It is geared towards expressing the dynamics of attitudes in cooperative problem solving and does not include temporal operators, for example.

**Definition 1 (Formulas).**
We inductively define a set of formulas $\mathcal{L}$ as follows.

**F1** each atomic proposition $p \in \mathcal{P}$ is a formula;
**F2** if $\varphi$ and $\psi$ are formulas, then so are $\neg\varphi$ and $\varphi \wedge \psi$;
**F3** if $\varphi$ is a formula, $\alpha$ is an individual action, $i, j \in \mathcal{A}$, $G \subseteq \mathcal{A}$, $\sigma$ a finite sequence of formulas, $\tau$ a finite sequence of individual actions, and $P$ a social plan expression, then the following are formulas:
**epistemic** $\text{BEL}(i, \varphi)$, $\text{E-BEL}_G(\varphi)$, $\text{C-BEL}_G(\varphi)$;
**motivational** $\text{GOAL}(i, \varphi)$, $\text{INT}(i, \varphi)$, $\text{C-INT}_G(\varphi)$, $\text{COMM}(i, j, \alpha)$, $\text{C-COMM}_{G,P}(\varphi)$, $\text{R-COMM}_{G,P}(\varphi)$, $\text{S-COMM}_{G,P}(\varphi)$, $\text{W-COMM}_{G,P}(\varphi)$, $\text{T-COMM}_{G,P}(\varphi)$; $\text{D-COMM}_{G,P}(\varphi)$

**temporal action and dynamic** $done(P)$, $succ(P)$, $failed(P)$, $do(P)$, $[P]\varphi$.

Social plans combine individual actions into a group action, and are represented by social plan expressions. The class of social plan expressions $\mathcal{S}p$ is based on a set of individual actions $\mathcal{A}c$, which are application-dependent and will not be defined here. An example is given below the definition.

**Definition 2 (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 social plan expression;
**SP2** if $\varphi \in \mathcal{L}$ and $G \subseteq \mathcal{A}$, then $\mathtt{conf}_G\, \varphi$ is social plan expression (of which the subscript $G$ is often left out);
**SP3** If $\alpha$ and $\beta$ are social plan expressions, then $(\alpha;\beta)$ (sequential composition) and $(\alpha \parallel \beta)$ (parallellism) are social plan expressions.

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

Let us give a simple **example** of a social plan expression. Consider a team consisting of three agents $t$ (theorem prover), $l$ (lemma prover) and $c$ (proof checker) who have as collective intention to prove a new mathematical theorem. Suppose during planning they define two lemmas, which also still need to be proved. This leads to the following complex individual actions: *prL1*, *prL2* (to prove lemma 1, respectively 2), *chL1*, *chL2* (to check a proof of lemma 1, respectively 2), *prT* (to prove the theorem from the conjunction of lemmas 1 and 2), *chT* (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:
$P = \langle\langle\langle prL1, l\rangle; \langle prL2, l\rangle\rangle \parallel \langle prT, t\rangle\rangle; \langle\langle chL1, c\rangle; \langle chL2, c\rangle; \langle chT, c\rangle\rangle\rangle$.

**Definition 3 (Kripke model).**
A Kripke model is a tuple
$\mathcal{M} = (W, \{B_i : i \in \mathcal{A}\}, \{G_i : i \in \mathcal{A}\}, \{I_i : i \in \mathcal{A}\}, \{R_P : P \in \mathcal{S}p\}, Val, perf, agents)$
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, $(w_1, w_2) \in B_i$ means that $w_2$ is an epistemic alternative for agent $i$ in state $w_1$;
3. For all $P \in \mathcal{S}p$, $R_P \subseteq W \times W$. They stand for the dynamic accessibility relations, e.g. $(w_1, w_2) \in R_{(i,\alpha)}$ means that $w_2$ is a possible resulting state from $w_1$ by agent $i$ executing action $\alpha$; for example, $R_{\mathtt{conf}_G(\varphi)} = \{(w, w) \in W \mid \mathcal{M}, w \models \varphi\}$ (see Definition 4 for $\mathcal{M}, w \models \varphi$);

4. $Val : \mathcal{P}\times W \rightarrow \{0,1\}$ is the function that assigns truth values to propositional atoms in states;
5. *ag*: $\mathcal{S}p\times W \rightarrow 2^{\mathcal{A}}$ is the agents function that indicates for each social plan in each world which set of agents is involved in it, e.g., $ag(\langle\alpha, i\rangle, w) = \{i\}$;
6. *perf*: $2^{\mathcal{A}}\times\mathcal{S}p \rightarrow (W\rightarrow\{0,1,2\})$ is the social plan performance function such that $perf(G,P)(w)$ indicates the result in world $w$ of the performance of social plan $P$ by a group of agents $G$ (here 0 stands for failure, 1 for success, and 2 stands for "undefined", e.g. for states $w$ that are not an endpoint of accessibility relation $R_P$);
7. *next*: $2^{\mathcal{A}}\times\mathcal{S}p \rightarrow (W\rightarrow\{0,1\})$ is the next moment social plan performance function such that $next(G,P)(w) = 1$ indicates that in world $w$ the group of agents $G$ will next start performing social plan $P$.

The accessibility relations for the epistemic and motivational operators may obey some restrictions corresponding to appropriate axioms (see [21–23, 17]). The accessibility relations $R_P$ are built up from accessibility relations for individual actions in an appropriate way (see [24]).

**Definition 4 (Semantics).**
Below, we give non-standard parts of the truth definition.

- $\mathcal{M}, v \models succ(P) \Leftrightarrow perf(ag(P,v),P)(v) = 1$;
- $\mathcal{M}, v \models failed(P) \Leftrightarrow perf(ag(P,v),P)(v) = 0$;
- $\mathcal{M}, v \models done(P) \Leftrightarrow perf(ag(P,v),P)(v) \in \{0,1\}$;
- $\mathcal{M}, v \models do(P) \Leftrightarrow next(ag(P,v),P)(v) = 1$;
- $\mathcal{M}, v \models [P]\varphi \Leftrightarrow$ for all $w((v,w) \in R_P \Rightarrow \mathcal{M}, w \models \varphi)$.

Note that $done(P)$ does not need its own truth conditions, for it can also be defined from the other basic constructions, namely as $succ(P) \vee failed(P)$.

For the dynamic operator, $[P]\varphi$ stands for "whenever $P$ terminates, it must do so in a state satisfying $\varphi$". We define the dual construct by $\langle P\rangle\varphi = \neg[P]\neg\varphi$. Because actions and their effects are not the main subject, we have "hard-wired" their performance and effects in the functions *perf* and *next*, so at each world it is determined whether a certain social plan $P$ has just been carried out (*done*), and if so, whether it was successful or not (*succ* respectively *failed*). Also one can express that a social plan $P$ will be carried out next (*do*)[5].

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

[5] One may formulate general principles about reasonable behavior of these constructs, such as $do(P) \rightarrow \langle P\rangle\top$, corresponding to the semantic property that for all $v \in W$, if $next(ag(P,v),P)(v) = 1$, then there exists a $t \in W$ such that $(v,t) \in R_P$. Also one may posit dynamic logic axioms for plans like $\langle\alpha, i\rangle$, analogous to those for propositional dynamic logic [24]. However, this is outside the scope of this chapter. See [16, Chapter 6] for details of the dynamic part of our logic for teamwork, TeamLog$^{dyn}$.

**Definition 5 (Truth definition).**

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

### 3.2 Axioms for beliefs

To represent beliefs, we adopt the standard $KD45_n$-system for $n$ agents as explained in [21], containing the following axioms and rules for $i = 1, \ldots, n$ :

**A1** All instantiations of tautologies of the propositional calculus
**A2** $\mathrm{BEL}(i, \varphi) \wedge \mathrm{BEL}(i, \varphi \rightarrow \psi) \rightarrow \mathrm{BEL}(i, \psi)$ (Belief Distribution)
**A4** $\mathrm{BEL}(i, \varphi) \rightarrow \mathrm{BEL}(i, \mathrm{BEL}(i, \varphi))$ (Positive Introspection)
**A5** $\neg\mathrm{BEL}(i, \varphi) \rightarrow \mathrm{BEL}(i, \neg\mathrm{BEL}(i, \varphi))$ (Negative Introspection)
**D** $\neg\mathrm{BEL}(i, \bot)$ (Belief Consistency)
**R1** From $\varphi$ and $\varphi \rightarrow \psi$ infer $\psi$ (Modus Ponens)
**R2** From $\varphi$ infer $\mathrm{BEL}(i, \varphi)$ (Belief Generalization)

Note that, in the semantics, the accessibility relations $B_i$ need not be reflexive, corresponding to the fact that an agent's beliefs need not be true. On the other hand, the accessibility relations $B_i$ are transitive, euclidean and serial. These conditions correspond to the axioms of positive and negative introspection and to the fact the agent has no inconsistent beliefs, respectively. It has been proved that $KD45_n$ is sound and complete with respect to these semantics [25, 21].

One can define modal operators for group belief. Let $G \subseteq \{1, \ldots, n\}$ be a group. The formula $\text{E-BEL}_G(\varphi)$ is meant to stand for "every agent in group $G$ believes $\varphi$". It is defined semantically as $\mathcal{M}, s \models \text{E-BEL}_G(\varphi)$ iff for all $i \in G$, $\mathcal{M}, s \models \mathrm{BEL}(i, \varphi)$, which corresponds to the following axiom:

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

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

$\text{C-BEL}_G(\varphi)$ is meant to be true if everyone in $G$ believes $\varphi$, everyone in $G$ believes that everyone in $G$ believes $\varphi$, etc. Let $\text{E-BEL}^1_G(\varphi)$ be an abbreviation for $\text{E-BEL}_G(\varphi)$, and let $\text{E-BEL}^{k+1}_G(\varphi)$ for $k \geq 1$ be an abbreviation of $\text{E-BEL}_G(\text{E-BEL}^k_G(\varphi))$. Thus we have $\mathcal{M}, s \models \text{C-BEL}_G(\varphi)$ iff $\mathcal{M}, s \models \text{E-BEL}^k_G(\varphi)$ for all $k \geq 1$. Define $t$ to be *$G_B$-reachable* from $s$ if there is a path of length $\geq 1$ 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 [21]):

$\mathcal{M}, s \models \text{C-BEL}_G(\varphi)$ iff $\mathcal{M}, t \models \varphi$ for all $t$ that are $G_B$-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** C-BEL$_G(\varphi) \leftrightarrow$ E-BEL$_G(\varphi \wedge$ C-BEL$_G(\varphi))$
**RC1** From $\varphi \rightarrow$ E-BEL$_G(\psi \wedge \varphi)$ infer $\varphi \rightarrow$ 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 [21].

In the sequel, we will use the following standard properties of C-BEL$_G$ (see for example [21, exercise 3.11]).

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

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

**Remark** Note that also the converse of the last property of Lemma 1 holds, namely:

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

To prove this, we use contraposition. So suppose that $\mathcal{M}, s \not\models$ C-BEL$_G(\varphi)$, then for some $i \in \{1, \ldots, n\}$ and for some $t \in W$, we have $(s,t) \in B_i$ and $\mathcal{M}, t \not\models \varphi$. Now we crucially apply the fact that we are working in $KD45_n$ frames, so that the $B_i$ are euclidean relations, meaning that for all $s, t, u$, if $(s,t) \in B_i$ and $(s,u) \in B_i$, then $(t,u) \in B_i$. This implies in particular in our example that $(t,t) \in B_i$, and as $M, t \not\models \varphi$, this means that $M, t \not\models$ C-BEL$_G(\varphi)$. But then, as $(s,t) \in B_i$, this immediately gives the desired conclusion $\mathcal{M}, s \not\models$ C-BEL$_G($C-BEL$_G(\varphi))$. We will show later that an analogous property does *not* hold for mutual intentions, even though these are at first sight similar to common beliefs.

**Logical omniscience** A problem with standard modal logics for beliefs and knowledge is that agents are formalized as being logically omniscient: they believe all theorems, as well as all logical consequences of their beliefs. Any modal logic with standard Kripke semantics in which belief is formalized as a necessity operator has this property. Logical omniscience definitely does not apply to human beings, who have only limited time available and have bounded rationality: it is unrealistic to assume that they believe every logical theorem, however complicated. There are several possible solutions to the problem, involving non-standard semantics or syntactic operators for awareness and explicit belief. Good references to the logical omniscience problem and its possible solutions are [25, Chapter 2] and [21, Chapter 9].

Another problem with common belief and common knowledge is that they are hard to attain in situations where the communication channel is not commonly known to be trustworthy. For example, in file transmission protocols at any time

only a bounded level of belief E-BEL$_G^{k+1}(\varphi)$ (and knowledge as well) about the message is achieved [26, 27]. A good reference to the difficulties concerning the attainment of common belief, as well as to possible solutions, is [21, Chapter 11].

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

**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 (E-BEL$_G(\psi)$), but they should commonly believe it (C-BEL$_G(\psi)$).

One advantage of common belief is that if C-BEL$_G$ holds for $\psi$, then C-BEL$_G$ also holds for all logical consequences of $\psi$. Thus, agents reason in a similar way from $\psi$ and commonly believe in this similar reasoning and the final conclusions. In short, one could say that common belief is hard to achieve, but easy to understand.

In cases in which only 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 $\psi$ and its consequences. However, establishing E-BEL$_G(\psi)$ places much less constraints on the communication medium than C-BEL$_G(\psi)$ does. Thus, the system developer's decision about the level $k$ of group belief (E-BEL$_G^k(\psi)$) to be established, hinges on determining a good balance between communication and reasoning, taking into account a particular application.

### 3.3 Axioms for individual and social motivational attitudes

Our framework to describe motivational attitudes and related aspects is minimal in the sense that we aim to deal with concise necessary and sufficient conditions. Additional aspects appearing on the stage in specific cases may be addressed by refining the system and adding new axioms. This subsection focuses on individual goals and intentions, and gives a short overview of our choice of axioms (adapted from [10]) and the corresponding semantic conditions. In this chapter, we leave out of consideration the aspects of time and action in order to focus on the main problem, the definition of collective intentions and commitments in terms of more basic attitudes.

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

**A1** All instantiations of tautologies of the propositional calculus
**R1** From $\varphi$ and $\varphi \rightarrow \psi$ infer $\psi$ (Modus Ponens)
**A2**$_G$ GOAL$(i, \varphi) \wedge$ GOAL$(i, \varphi \rightarrow \psi) \rightarrow$ GOAL$(i, \psi)$ (Goal Distribution)
**A2**$_I$ INT$(i, \varphi) \wedge$ INT$(i, \varphi \rightarrow \psi) \rightarrow$ INT$(i, \psi)$ (Intention Distribution)
**R2**$_G$ From $\varphi$ infer GOAL$(i, \varphi)$ (Goal Generalization)

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

In a BDI system, an agent's activity starts from goals. In general, it may have many different objectives which will not all be pursued. As opposed to intentions, goals are not directly related to actions, so an agent can behave rationally, even though it has different inconsistent goals. Thus, in contrast to Rao and Georgeff we adopt the basic system $K_n$ for goals[6]. Then, the agent chooses a limited number of these goals to be intentions. Here we do not discuss how intentions are formed from a set of goals (but see [30, 31]). 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:

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

Nevertheless, in our approach other choices may be adopted without consequences for the presented definitions.

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 [32].

### 3.4 Interdependencies between attitudes

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

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

These four axioms express that agents are aware of the goals and intentions they have, as well as of the lack of those that they do not have. Notice that we do not add the axioms of *strong realism* that Rao and Georgeff adopt for a specific set of formulas $\varphi$, the so-called O-formulas: $\mathrm{GOAL}(i, \varphi) \rightarrow \mathrm{BEL}(i, \varphi)$ and $\mathrm{INT}(i, \varphi) \rightarrow \mathrm{BEL}(i, \varphi)$, corresponding to the fact that an agent believes that it can optionally achieve its goals and intentions by carefully choosing its actions. These axioms correspond to semantic restrictions on the branching time models considered in [10]. On the other hand, we do not adopt the converse axiom of *realism* advocated by Cohen and Levesque: $\mathrm{BEL}(i, \varphi) \rightarrow \mathrm{GOAL}(i, \varphi)$. In their

[6] Similarly to the logical omniscience problem discussed in Subsection 3.2, the choice of a normal modal logic for goals also presents unrealistic consequences. Because of the validity of $(\mathrm{GOAL}(i, \varphi) \wedge \mathrm{GOAL}(i, \neg\varphi)) \leftrightarrow \mathrm{GOAL}(i, \varphi \wedge \neg\varphi)$, having inconsistent goals would lead an agent $i$ to have *any* goal $\psi$. Possible solutions for such disadvantages of normal modal logics have been sought in reverting to a non-monotonic logic (see [28]) and in using non-normal modal logics with variations of neighborhood models (see [29]).

formalism, where a possible world corresponds to a time line, the realism axiom expresses that agents adopt as goals the inevitable facts about the world [13]. Both versions of realism are intimately connected to the choice of temporal structure, a theme that we leave out of consideration here.

The semantic property corresponding to $\mathbf{A7}_{IB}$ is $\forall s,t,u((sB_it \wedge tI_iu) \rightarrow sI_iu)$, analogously for $\mathbf{A7}_{GB}$. The property that corresponds to $\mathbf{A8}_{IB}$ is $\forall s,t,u((sI_it \wedge sB_iu) \rightarrow uI_it)$, analogously for $\mathbf{A8}_{GB}$. For a proof, see [33] or [16, Ch. 3].

We assume that every intention corresponds to a goal:

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

This means that if an agent adopts a formula as an intention, it should have adopted that formula as a goal to achieve, which satisfies Bratman's notion that an agent's intentions form a specific subset of its goals [9]. Rao and Georgeff adopt this axiom as *goal-intention compatibility* for their class of O-formulas [10]. In our non-temporal context, the corresponding semantic property is that $G_i \subseteq I_i$. For a proof, see [33] or [16, Ch. 3].

Note that in our system it fortunately does not follow that an agent intends all the consequences it believes its intentions to have, i.e. the side-effects of its intentions. There is weaker version that does hold, though, namely if $\models \varphi \rightarrow \psi$, then $\models \mathrm{INT}(i,\varphi) \rightarrow \mathrm{INT}(i,\psi)$. This is similar to the logical omniscience problem for logics of knowledge and belief discussed in Subsection 3.2. For a discussion of the "side-effect problem" for intentions, see [9, 13, 10].

### 3.5 Social commitments

As Castelfranchi showed, it is important to distinguish between individual intentions, bilateral commitments, and collective motivational attitudes [34]. A social (bilateral) commitment is not as strong as a collective one, but stronger than an individual intention. If an agent commits to a second agent to do something, then the first agent should have the *intention* to do that. Moreover, the second one should be *interested* in the first one fulfilling its intention. These two conditions (inspired by [34]), need to be enhanced by the condition expressing the agents' awareness about the situation, i.e. about their individual attitudes. Such awareness is usually expressed in terms of common belief. Here the defining axiom for social commitments is formulated with respect to actions[7]:

[7] It is possible to define social commitments with respect to propositions as well, as is done in [16, Chapter 4]:

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

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

where $\mathtt{stit}(i,\varphi)$ means that agent $i$ sees to it (brings it about) that $\varphi$ becomes true (see [35–37]). However, we only need the "action" form of social commitments in the rest of this chapter.

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

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

where $\mathit{done}(i, \alpha)$ means that agent $i$ has just executed action $\alpha$ (see Subsection 3.1).

## 4 Defining collective intention

In our approach, teams are created on the basis of *collective intentions*, which are defined in Subsection 4.1 for the standard context and in [17], [16, Ch. 3] for less ideal circumstances. A team exists as long as the collective intention between team members exists. Here we abstract from the ways in which teams are formed, and refer the interested reader to [38, 30, 39, 40] and [16, Ch. 8].

In the sequel, collective intention is viewed as an inspiration for the creation of collective commitment leading directly to action execution. This is based on the linguistic tradition that intentions typically ultimately lead to actions, however, the immediate triggers of these actions are commitments.

In the philosophical and MAS literature there is an ongoing discussion as to whether collective intentions may be reduced to individual ones plus common beliefs about them (see [34, 41–43]). Even though our definition seems to be reductive, it involves nested intentions and collective epistemic operators, and for this reason is deeper than a simple compound built out of individual intentions and common beliefs about them by propositional connectives only.

### 4.1 Collective intentions: the standard case

In this chapter we focus on strictly cooperative teams (see for example [44, 45, 43] for good philosophical discussions of cooperation). This makes the definition of collective intention rather strong. In such teams, a necessary condition for a collective intention is that all members of the team $G$ have the associated individual intention $\mathrm{INT}(i, \varphi)$ towards the state of the world represented by proposition $\varphi$. In fact, this condition is taken to give the full definition of collective intention in [46] (see [47] for a similar definition of collective goal). However, this is certainly not sufficient. Imagine that two agents want to achieve the same state of the world but are in a competition about this, willing to achieve it exclusively. Therefore, to exclude the case of competition, all agents should *intend* all members to have the associated individual intention, as well as the intention that all members have the individual intention, and so on; we call such a mutual intention $\text{M-INT}_G(\varphi)$. Furthermore, all members of the team are aware of this mutual intention, that is, they have a common belief about this: $\text{C-BEL}_G(\text{M-INT}_G(\varphi))$. Of course, team members remain autonomous in maintaining their other motivational attitudes, and may even be in competition about other issues.

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

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

The mutual intention M-INT$_G(\varphi)$ is meant to be true if everyone in $G$ intends $\varphi$, everyone in $G$ intends that everyone in $G$ intends $\varphi$, etc. As we do not have infinite formulas to express this, let E-INT$^1_G(\varphi)$ be an abbreviation for E-INT$_G(\varphi)$, and let E-INT$^{k+1}_G(\varphi)$ for $k \geq 1$ be an abbreviation of E-INT$_G$(E-INT$^k_G(\varphi)$). Thus we have $\mathcal{M}, s \models$ M-INT$_G(\varphi)$ iff $\mathcal{M}, s \models$ E-INT$^k_G(\varphi)$ for all $k \geq 1$. Define world $t$ to be $G_I$*-reachable* from world $s$ if there is a path of length $\geq 1$ in the Kripke model from $s$ to $t$ along accessibility arrows $I_i$ that are associated with members $i$ of $G$. Then the following property holds (see Subsection 3.2 and [21] for an analogous property for common belief and common knowledge, respectively):

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

Using this property, it can be shown that the following fixed-point axiom and rule can be soundly added to the union of $KD_n$ (the axiom system for individual intentions given in Subsection 3.3) and **M1**:

**M2** M-INT$_G(\varphi) \leftrightarrow$ E-INT$_G(\varphi \wedge$ M-INT$_G(\varphi))$
**RM1** From $\varphi \rightarrow$ E-INT$_G(\psi \wedge \varphi)$ infer $\varphi \rightarrow$ M-INT$_G(\psi)$ (Induction Rule)

The resulting system is called $KD_n^{\text{M-INT}_G}$, and it is sound and complete with respect to Kripke models where all $n$ accessibility relations are serial (see [17] or [16, Ch. 3] for a proof).

**Remark** There is an interesting difference between the properties of common belief and those of mutual intention, even though their definitions are very similar. The distinction is due to a difference between the underlying frameworks for individual beliefs and intentions. As a reminder, for common beliefs it holds that

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

(see Lemma 1 and the subsequent remark). Similarly, it is easy to show from the axioms that

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

However, surprisingly, the other direction does not hold, as can be shown by the counter-model in Figure 1, in which $\mathcal{M}, s_1 \models$ M-INT$_G$(M-INT$_G(p))$, but $\mathcal{M}, s_1 \not\models$ M-INT$_G(p)$. The possibility of the counter-example crucially hinges on the fact that $KD_n$ corresponds to frames in which the intention accessibility relations are serial, but not necessarily euclidean.

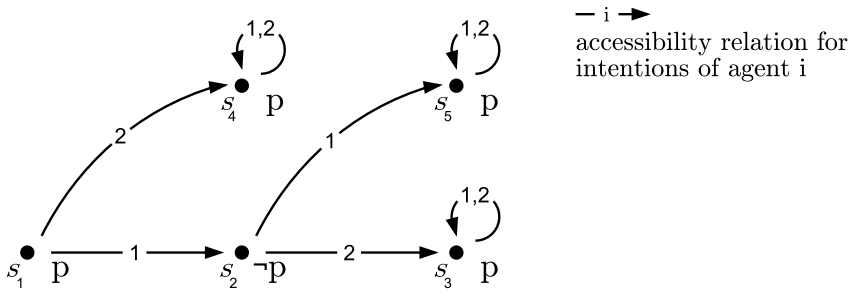


**Fig. 1.** A $KD_n^{\text{M-INT}_G}$ model with relations $I_i$. The following hold for $G = \{1, 2\}$: $\mathcal{M}, s_1 \not\models \text{M-INT}_G(p)$, in particular $\mathcal{M}, s_1 \not\models \text{E-INT}_G(p)$, because $(s_1, s_2) \in I_1$ and $\mathcal{M}, s_2 \not\models p$. However, $\mathcal{M}, s_1 \models \text{M-INT}_G(\text{M-INT}_G(p))$, because we have $\mathcal{M}, s_i \models \text{M-INT}_G(p)$ for all $i \neq 1$, i.e., $\text{M-INT}_G(p)$ holds in all $s_i$ that are $G_I$-reachable from $s_1$.

**From mutual to collective intention** 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 [48, 49]. The definition would be even stronger if common knowledge instead of common belief appeared in **M3**. However, because common knowledge is almost impossible to establish in multi-agent systems due to (among other circumstances) the unreliability of communication media, we do not pursue this strengthening further.

The resulting system, which we call $KD_n^{\text{C-INT}_G}$, is the union of $KD_n^{\text{M-INT}_G}$ (for mutual intentions), $KD45_n^C$ (for common beliefs) and axiom **M3**. In [17] and [16, Ch. 3] we have also defined alternative versions of collective intention that are applicable in different circumstances.

Let us give an informal **example** of the establishment of a collective intention. Two violinists, $a$ and $b$, have studied together and have toyed with the idea of giving a concert together someday. Later this becomes more concrete: they both intend to perform together the solo parts of the Bach Double Concerto ($\text{INT}(a, \varphi)$ and $\text{INT}(b, \varphi)$, where $\varphi$ stands for "$a$ and $b$ perform the solo parts of the Bach Double Concerto"). After communicating with each other about this, they start practicing together. Clearly, a mutual intention as defined in **M2** is now in place (involving nested intentions like $\text{INT}(a, \text{INT}(b, \text{INT}(a, \varphi)))$ and so on). The communication established a common belief $\text{C-BEL}_G(\varphi)$ (with $G = \{a, b\}$) about their mutual intention, according to **M3**. As sometimes happens in life, when people are ready, an opportunity appears: Carnegie Hall plans a concert for Christmas Eve, including the Bach Double Concerto. Now they refine their collective intention to a more concrete $\text{C-INT}_G(\psi)$ (where $\psi$ stands for "$a$ and $b$ perform the solo parts of the Bach Double Concerto at

the Christmas Eve concert in Carnegie Hall"). It happens that our two violinists are chosen from among a list of candidates to be the soloists, and both sign the appropriate contract. Because they do this together, we can speak about common knowledge, not merely common belief, of their mutual intention: M-INT$_G(\psi) \wedge$ C-KNOW$_G$(M-INT$_G(\psi)$).

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

Even though C-INT$_G(\varphi)$ seems to be an infinite concept, collective intentions may be established in practice in a finite number of steps. As defined by **M3**, collective intentions are appropriate to model those situations in which communication, in particular announcements, work, especially if one initiator establishes the team. We have showed in detail in [30] and [16, Ch. 8] how team formation in such an ideal case may actually work in terms of the first two stages of cooperative problem solving, namely potential recognition and team formation, and how at these stages the proper attitudes are established through dialogues consisting of the appropriate speech acts.

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

The amount of necessary communication, expressed by C-BEL$_G$(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 $\varphi$ be a formula and $G \subseteq \{1, \ldots, n\}$. Then the following holds:*

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

**Remark** Interestingly, the converse of the above property of awareness of collective intentions does not hold, so we can *not* derive C-BEL$_G$(C-INT$_G(\varphi)$) $\rightarrow$ C-INT$_G(\varphi)$. As in the case of the interesting property of mutual intentions, this can again be shown by a counter-model plus some extra reasoning. First, we show that

$$KD_n^{\text{C-INT}_G} \nvdash \text{C-BEL}_G(\text{M-INT}_G(\varphi)) \rightarrow \text{M-INT}_G(\varphi).$$

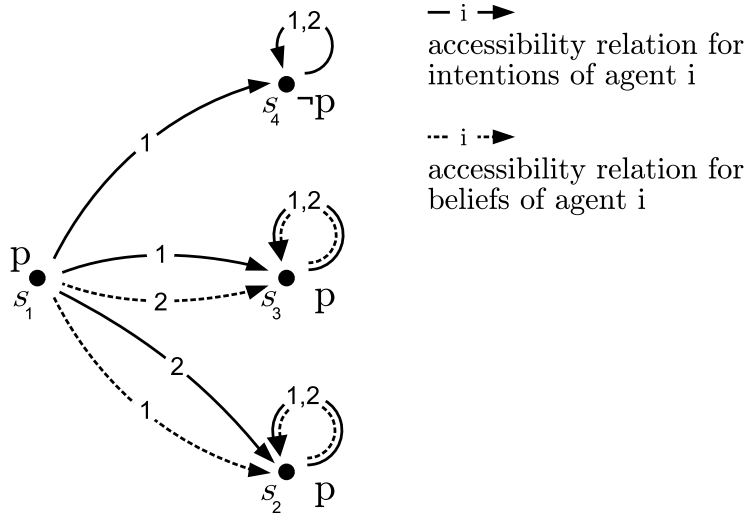


**Fig. 2.** A $KD_n^{\text{C-INT}_G}$ model with relations $I_i$ and $B_i$. The following hold for $G = \{1, 2\}$: $\mathcal{M}, s_1 \not\models \text{M-INT}_G(p)$, in particular $\mathcal{M}, s_1 \not\models \text{E-INT}_G(p)$, because $(s_1, s_4) \in I_1$ and $\mathcal{M}, s_4 \not\models p$. However, $\mathcal{M}, s_1 \models \text{C-BEL}_G(\text{M-INT}_G(p))$, because we have $\mathcal{M}, s_2 \models \text{M-INT}_G(p)$ and $\mathcal{M}, s_3 \models \text{M-INT}_G(p)$, and $s_2$ and $s_3$ are the only worlds that are $G_B$ reachable from $s_1$.

For this, the counter-model in Figure 2 suffices: $\mathcal{M}, s_1 \models \text{C-BEL}_G(\text{M-INT}_G(\varphi))$, but $\mathcal{M}, s_1 \not\models \text{M-INT}_G(p)$.

It turns out that the same model enables us to show that

$$KD_n^{\text{C-INT}_G} \not\vdash \text{C-BEL}_G(\text{C-INT}_G(\varphi)) \rightarrow \text{C-INT}_G(\varphi).$$

Firstly, the fact that $\mathcal{M}, s_1 \not\models \text{M-INT}_G(p)$ immediately implies that

$$\mathcal{M}, s_1 \not\models \text{C-INT}_G(p),$$

because by **M3** collective intention implies mutual intention, and common belief distributes over this implication. On the other hand, similarly to Lemma 2, we can use Lemma 1 to show that

$$KD_n^{\text{C-INT}_G} \vdash \text{C-BEL}_G(\text{M-INT}_G(\varphi)) \rightarrow \text{C-BEL}_G(\text{C-INT}_G),$$

so the fact that $\mathcal{M}, s_1 \models \text{C-BEL}_G(\text{M-INT}_G(\varphi))$ implies that

$$\mathcal{M}, s_1 \models \text{C-BEL}_G(\text{C-INT}_G(\varphi)).$$

**Shared cooperative activity** One could wonder whether our definition of collective intention does not cover inappropriate cases where real teamwork is out of the question. Bratman [44, Ch. 5] characterizes shared cooperative activity. He

presents various examples of situations where the agents share some attitudes, but truly shared cooperative activity is out of the question. It turns out that our definition excludes these cases, as well. For example,

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

Thus, taking $\varphi =$"$a$ and $b$ go to New York" with $a$ for "you" $b$ for "me", $G$ for $\{a, b\}$, in the situation above E-INT$_G(\varphi)$ and possibly also C-BEL$_G$(E-INT$_G(\varphi)$) holds, but M-INT$_G(\varphi)$ and C-INT$_G(\varphi)$ do not. Specifically, it seems unlikely that INT$(b,$INT$(a, \varphi))$ holds for the Mafioso. Note that Rao, Georgeff and Sonenberg's definition of a *joint intention* among $G$ to achieve $\varphi$ is defined as E-INT$_G(\varphi) \wedge$ C-BEL$_G$(E-INT$_G(\varphi)$) (translated to our notation), thus it erroneously ascribes a joint intention to go to New York among the agents in the example. Incidentally, a similar one-level definition of mutual goals was also given in [47].

**Comparison with the two-level definition** In previous work, we gave a somewhat weaker definition of collective intention than the one above (see [48, 49]). It consisted of two levels of reciprocal intentions in a team, and a common belief about this; so it did not erroneously assign a collective intention to sets whose members are in individual competition, as a one-level definition does (such as it appears in e.g. [46]). Here follows the two-level definition:

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

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

If we adapt the definition above to make it consist of three levels of intention instead of two, the troublesome example would be solved. However, one may

invent similar (admittedly artificial) examples for any $k$, using coalitions of $k$ people from among a base set of at least $k+1$ agents. Thus, the infinitary mutual intention of the previous section was derived to avoid all such counterexamples. One can see, however, that in practical situations, for any fixed finite group, a finite number of levels of the mutual intention is sufficient to construct the collective intention among them.

## 5 A tuning machine for collective commitment

After a group is constituted on the basis of collective intention, another stage of cooperative problem solving, namely *plan formation* is started, 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 by agents accepting pairwise (i.e. social) commitments to realize their individual actions. This way, our approach to collective commitments is plan-based.

Let us turn for a moment to the commonsense meaning of teamwork. 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 cooperative problem solving, albeit a non-typical case.

In the two examples above, individual and collective awareness about the ingredients of cooperative problem solving (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. In the past 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 [40, 48, 46, 50]. These definitions of

collective attitudes were independent of organisational structures and communication possibilities. In contrast, here we 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 cooperative problem solving. 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.

While investigating the calibration of group commitment to the particular purpose and the specific circumstances, we isolated and separately characterized invariant ingredients of collective commitments. These are:

- 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', 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.

### 5.1 General schema of collective commitment

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

1. Mutual intention M-INT$_G(\varphi)$ between a group of agents.
   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.
   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. The predicate $cons(\varphi, P)$ informally stands for "$P$ is a correct social plan to achieve $\varphi$". For a definition and examples of social plans, see Section 3; for a formal definition of $cons(\varphi, P)$, see [16, Chapter 6].
3. Pairwise social commitments COMM$(i, j, \alpha)$ for actions from the plan.
   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. It may vary from the lack of any awareness to common

belief about the given aspect, as it was discussed before. 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' :

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

In words, group $G$ has a collective commitment to achieve overall goal $\varphi$ based on social plan $P$ (C-COMM$_{G,P}(\varphi)$) iff all of the following hold. The group mutually intends $\varphi$ (with or without being aware); moreover, successful execution of social plan $P$ leads to $\varphi$ ($\mathit{cons}(\varphi, P)$) (with or without the group being aware of this); and finally, for every one of the actions $\alpha$ from a 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 (COMM$(i, j, \alpha)$) (with or without the group being aware of this).

Instantiating the above schema corresponds to tuning the $\mathit{awareness}_G$-dials from $\emptyset$, through individual beliefs and different degrees of E-BEL$^k_G$, to common belief, and/or analogously for degrees of knowledge. These degrees have been discussed in Subsection 3.2.

### 5.2 Different aspects of agents' awareness

The notion of collective commitment, whatever 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 concerns different aspects of awareness that appear in a group of agents in the course of cooperative problem solving. The degree of this awareness, characterized in terms of different types of beliefs, may be different. In the sequel, the strongest version is considered, including common belief about considered aspect of cooperative problem solving. 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$^k_G$) 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:
   C-BEL$_G$(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:
C-BEL$_G(cons(\varphi, P))$

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

The distinction de re / de dicto stems from the philosophy of language [51]. A sentence of the form $\exists x$BEL$(j, A(x))$ is a *de re* belief attribution which relates $j$ to a *res*, an individual that the belief is about. On the other hand, BEL$(j, \exists x A(x))$ is a *de dictum* belief attribution, relating $j$ to a *dictum*, namely the proposition $\exists x A(x)$. This distinction is also fruitful for complex epistemic operators such as common 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.

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

### 5.3 Different notions of collective commitment

The following exemplar definitions are produced by keeping the *awareness*$_G$-dial fixed to a choice between $\emptyset$ and common belief, and the dial for 'kind of mutual intention' fixed as the standard definition of Subsection 4.1. We will start from the strongest form of collective commitment, fully reflecting the collective aspects of teamwork. Subsequently, some underlying assumptions will be relaxed, leading ultimately to weaker notions of team and distributed commitment.

**Robust collective commitment** Our discussion on different types of collective commitments will start from the two cases based on collective planning. Additionally, for every one of the actions $\alpha$ 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, in a *robust collective commitment* (R-COMM$_{G,P}$), there is a detailed awareness of social

commitments in the team:

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

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.

**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.

**Strong collective commitment** In contrast to robust collective commitment, in the case of *strong collective commitment* ($\text{S-COMM}_{G,P}$), there is a global 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:

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

As the reponsibility 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 there is no common belief about pairwise social commitments, 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 subteams. Typically, planning is done collectively. However, establishing bilateral commitments is not done publicly in the whole team, but in subgroups. Sometimes this suffices, and it is sometimes preferable in order not to waste energy.

**Weak collective commitment** In the weaker cases of collective commitment, the degree of team awareness is even more limited. When the plan as a whole is not known to the team, for example, if no collective decision making is assumed, there is no awareness that the plan leads to proper realization of the goal ($\text{C-BEL}_G cons(\varphi, P)$ is not in place). In this case we deal with a *weak collective commitment* ($\text{W-COMM}_{G,P}$):

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

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

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 does believe in the plan's correctness $cons(\varphi, P)$, and this may suffice.

**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. In this situation, there is no common belief that all actions have been adopted by other committed members, but a team as a structure still exists :

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

Because of the presence of collective intention, the overall goal and composition of the team are commonly 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 its employees. Information about the precise involvement of colleagues and other aspects of the plan may be confidential.

**Distributed commitment** The last case distinguished here is *distributed commitment* (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:

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

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, e.g. in order to avoid competition.

## 6 Discussion and conclusions

This chapter falls within a larger research program, the first part of which presents a *static* characterization of teamwork with collective commitment as

a central notion [17]. This theory is built incrementally starting from individual intentions, through social commitments, leading ultimately to the collective level of intentions and commitments. All these notions play a crucial role in practical reasoning. They are defined in multi-modal logics with clear semantics (cf. [17]), comprising a descriptive view on collective motivational attitudes.

The reader will note that collective intentions are not introduced here as primitive modalities, with some restrictions on the semantic accessibility relations (as in e.g. [52]). We do give necessary and sufficient conditions for such collective motivational attitudes to be present. In this way, we hope to make the behavior of a team easier to predict. We have tried to find *minimal* conditions for collective intentions to be present, and not to weigh down the definitions with all aspects that play a part in the establishment of collective intentions. Such elements as conventions, abilities, opportunities, power relations and social structure (see [53, 45, 40] for a thorough discussion) certainly are important, and we leave open the possibility of defining and using them in specific cases where they play a crucial role. For example, abilities and opportunities are important in the dialogues leading up to the establishment of a collective intention and a team based on it [30],[16, Ch. 8]. Power relations and social structure, on the other hand, are reflected in the definitions of collective commitments.

The definitions of collective commitments are not overloaded, and therefore easy to understand and to use. In contrast to [48], 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. Such an approach is especially strong when re-planning is needed. The multi-modal logic framework allows to express subtle aspects of cooperative problem solving, modeling different situations occurring in practical domains.

The presented system defining collective intentions and collective commitments is decidable. It is known to be EXPTIME-complete [32], so in general it is not feasible to give automated proofs of desired properties; at least there is no single algorithm that performs well on all inputs. As with other modal logics, the better option would be to develop a variety of different algorithms and heuristics, each performing well on a limited class of inputs. For example, it is known that restricting the number of propositional atoms to be used or the depth of modal nesting may reduce the complexity (cf. [54–57, 32]). Also, when considering specific applications it is possible to reduce some of the infinitary character of common beliefs and intentions to more manageable proportions (cf. [21, Ch. 11]).

The static definitions of relevant motivational attitudes express solely vital aspects of teamwork, leaving room for case-specific extensions. This set of *teamwork axioms* constitutes a definition of motivational attitudes in BDI systems. This way they may serve a system designer as a specification to create a correct and complete system, as well as to verify a system.

The presented analysis of social and collective attitudes in teams of agents assumes a rather high level of idealization: solely strictly cooperative teams are considered. This leads to a strong definition of collective intention, based on agents' mutual intentions, and then to a plan-based collective commitment. Even though in [33] we introduced a general tuning mechanism, presented here in Section 5, to calibrate the strength of collective commitments fitting to a variety of circumstances, an essential ingredient of these definitions — agents' awareness — is formalized by means of a strong notion of common belief. After investigating and formalizing this basic case, it is time to relax some of the strong assumptions underlying this research in order to take a closer look on weaker and more distributed forms of cooperation.

Also, this normal modal framework, like any logic based on standard Kripke semantics, suffers from well-known problems related to logical omniscience. Agents are supposed to know and intend all tautologies; moreover, they are supposed to know all logical consequences of their knowledge, and to intend all logical consequences of their intentions. This is clearly unrealistic. For epistemic logic, several solutions to the omniscience problem have been proposed, mostly based on non-normal modal logics ([21, Ch. 9]). Similar solutions were proposed for individual intentions. For future research, we plan to design a non-normal multi-modal logic suitable to solve logical omniscience problems for our framework characterizing collective motivational attitudes.

For a much more extensive discussion of TeamLog$^{dyn}$, our full logical theory of teamwork in dynamic multi-agent environments, as well as precise comparisons to related work, see our recent book [16].

## 7 Acknowledgements

We are grateful to the Netherlands Institute for Advanced Study (NIAS) for providing an opportunity to write the first version of this chapter as Fellows in Residence. Furthermore, we would like to thank the anonymous referee for fruitful suggestions. Harmen Wassenaar carefully made the two illustrations, for which we are grateful. Finally, we thank the Polish MNiSW grant N N206 399334 for supporting Barbara Dunin-Kęplicz' research, and the Netherlands Organization for Scientific Research for a Replacement grant NWO 400-05-710 and Vici grant NWO 277-80-001, enabling Rineke Verbrugge's research.

## References

1. Tomasello, M.: Why We Cooperate. MIT Press, Cambridge, MA (2009)
2. Gärdenfors, P.: The cognitive and communicative demands of cooperation. In van Eijck, J., Verbrugge, R., eds.: Games, Actions and Social Software. Texts in Logic and Games (FOLLI subseries of LNCS). Springer Verlag, Berlin (2011) XX–YY
3. Borrill, C., West, M.: The psychology of effective teamworking. In Gold, N., ed.: Teamwork. Palgrave McMillan, Basingstoke and New York (2005) 136–160
4. Huczynski, A., Buchanan, D.: Organizational Behaviour: An Introductory Text. Sixth edn. Pearson Education Ltd., Essex (2007)

5. Kozlowski, S., Bell, B.: Work groups and teams in organizations. In Borman, W., Ilgen, D., Klimoski, R., eds.: Handbook of Psychology: Industrial and Organizational Psychology. Wiley, Chichester (2003) 333–375
6. Kozlowski, S.W.J., Ilgen, D.R.: Enhancing the effectiveness of work groups and teams. Psychological Science in the Public Interest (2006) 77–124
7. Akkerman, S., van den Bossche, P., Admiraal, W., Gijselaers, W., Segers, M., Simons, R., Kirschner, P.: Reconsidering group cognition: From conceptual confusion to a boundary area between cognitive and socio-cultural perspectives? Educational Research Review **2** (2007) 39–63
8. Dennett, D.: The Intentional Stance. MIT Press, Cambridge, MA (1987)
9. Bratman, M.: Intention, Plans, and Practical Reason. Harvard University Press, Cambridge, MA (1987)
10. Rao, A., Georgeff, M.: Modeling rational agents within a BDI-architecture. In Fikes, R., Sandewall, E., eds.: Proceedings of the Second Conference on Knowledge Representation and Reasoning, Morgan Kaufman (1991) 473–484
11. Weiss, G., ed.: Multiagent Systems. MIT Press, Cambridge, MA (1999)
12. Levesque, H., Cohen, P., Nunes, J.: On acting together. In: Proceedings Eighth National Conference on AI (AAAI90), Menlo Park (CA), Cambridge (MA), AAAI-Press and MIT Press (1990) 94–99
13. Cohen, P., Levesque, H.: Intention is choice with commitment. Artificial Intelligence **42** (1990) 213–261
14. Dunin-Kęplicz, B., Verbrugge, R.: A reconfiguration algorithm for distributed problem solving. Engineering Simulation **18** (2001) 227 – 246
15. Dunin-Kęplicz, B., Verbrugge, R.: Evolution of collective commitments during teamwork. Fundamenta Informaticae **56** (2003) 329–371
16. Dunin-Kęplicz, B., Verbrugge, R.: Teamwork in Multi-Agent Systems: A Formal Approach. Wiley, Chichester (2010)
17. Dunin-Kęplicz, B., Verbrugge, R.: Collective intentions. Fundamenta Informaticae **51(3)** (2002) 271–295
18. Ferber, J.: Multi-agent Systems: An Introduction to Distributed Artificial Intelligence. Addison Wesley, Reading, MA (1999)
19. Osborne, M.: An Introduction to Game Theory. Oxford University Press, New York, NY (2004)
20. Rasmusen, E.: Games and Information. Fourth edn. Blackwell, Malden, MA (2008)
21. Fagin, R., Halpern, J., Moses, Y., Vardi, M.: Reasoning about Knowledge. MIT Press, Cambridge, MA (1995)
22. van Linder, B., van der Hoek, W., Meyer, J.: Formalising abilities and opportunities of agents. Fundamenta Informaticae **34** (1998) 53–101
23. Wooldridge, M.: Reasoning About Rational Agents. MIT Press, Cambridge, MA (2000)
24. Harel, D., Kozen, D., Tiuryn, J.: Dynamic Logic. MIT Press, Cambridge, MA (2000)
25. Meyer, J., van der Hoek, W.: Epistemic Logic for AI and Theoretical Computer Science. Cambridge University Press, Cambridge (1995)
26. Halpern, J., Zuck, L.: A little knowledge goes a long way: Simple knowledge-based derivations and correctness proofs for a family of protocols. In: 6th ACM Symposium on Principles of Distributed Computing. (1987) 268–280
27. Stulp, F., Verbrugge, R.: A knowledge-based algorithm for the internet protocol TCP. Bulletin of Economic Research **54** (2002) 69–94
28. Broersen, J., Dastani, M., van der Torre, L.: Realistic desires. Journal of Applied Non-Classical Logics **12** (2002) 287–308

29. Su, K., Sattar, A., Lin, H., Reynolds, M.: A modal logic for beliefs and pro attitudes. In: Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI)- Volume 1, AAAI Press (2007) 496–501
30. Dignum, F., Dunin-Kęplicz, B., Verbrugge, R.: Creating collective intention through dialogue. Logic Journal of the IGPL **9** (2001) 145–158
31. Dignum, F., Conte, R.: Intentional agents and goal formation: Extended abstract. In Singh, M., Rao, A., Wooldridge, M., eds.: Preproceedings Fourth International Workshop on Agent Theories, Architectures and Languages, Providence, Rhode Island (1997) 219–231
32. Dziubiński, M., Verbrugge, R., Dunin-Kęplicz, B.: Complexity issues in multiagent logics. Fundamenta Informaticae **75(1-4)** (2007) 239–262
33. Dunin-Kęplicz, B., Verbrugge, R.: A tuning machine for cooperative problem solving. Fundamenta Informaticae **63** (2004) 283–307
34. Castelfranchi, C.: Commitments: From individual intentions to groups and organizations. In Lesser, V., ed.: Proceedings First International Conference on Multi-Agent Systems, San Francisco, AAAI-Press and MIT Press (1995) 41–48
35. Segerberg, K.: Bringing it about. Journal of Philosophical Logic **18** (1989) 327–347
36. Horty, J.F., Belnap, N.: The deliberative Stit: A study of action, omission, ability, and obligation. Journal of Philosophical Logic **24** (1995) 583–644
37. Broersen, J., Herzig, A., Troquard, N.: Embedding Alternating-time Temporal Logic in strategic STIT logic of agency. Journal of Logic and Computation **16** (2006) 559–578
38. Castelfranchi, C., Miceli, M., Cesta, A.: Dependence relations among autonomous agents. [58]
39. Jennings, N.: Commitments and conventions: The foundation of coordination in multi-agent systems. Knowledge Engineering Review **3** (1993) 223–250
40. Wooldridge, M., Jennings, N.: Cooperative problem solving. Journal of Logic and Computation **9** (1999) 563–592
41. Haddadi, A.: Communication and Cooperation in Agent Systems: A Pragmatic Theory. Volume 1056 of LNAI. Springer Verlag, Berlin (1995)
42. Tuomela, R., Miller, K.: We-intentions. Philosophical Studies **53** (1988) 367–390
43. Gilbert, M.: A theoretical framework for the understanding of teams. In Gold, N., ed.: Teamwork. Palgrave McMillan, Basingstoke and New York (2005) 22–32
44. Bratman, M.: Faces of Intention. Cambridge University Press, Cambridge (1999)
45. Tuomela, R.: The Importance of Us: A Philosophical Study of Basic Social Notions. Stanford Series in Philosophy. Stanford University Press, Stanford (CA) (1995)
46. Rao, A., Georgeff, M., Sonenberg, E.: Social plans: A preliminary report. [58] 57–76
47. Wooldridge, M., Jennings, N.: Towards a theory of collective problem solving. In Perram, J., Muller, J., eds.: Distributed Software Agents and Applications. Volume 1069 of LNAI. Springer Verlag, Berlin (1996) 40–53
48. Dunin-Kęplicz, B., Verbrugge, R.: Collective commitments. In Tokoro, M., ed.: Proceedings Second International Conference on Multi-Agent Systems, Menlo Park (CA), AAAI-Press (1996) 56–63
49. Dunin-Kęplicz, B., Verbrugge, R.: Collective motivational attitudes in cooperative problem solving. In Gorodetsky, V., ed.: Proceedings of the First International Workshop of Eastern and Central Europe on Multi-agent Systems (CEEMAS'99), St. Petersburg (1999) 22–41
50. Grosz, B., Kraus, S.: The evolution of SharedPlans. In Rao, A., Wooldridge, M., eds.: Foundations of Rational Agency. Kluwer, Dordrecht (1999) 227–262

51. Quine, W.: Quantifiers and propositional attitudes. Journal of Philosophy **53** (1956) 177–187
52. Cavedon, L., Rao, A., Tidhar, G.: Social and individual commitment (preliminary report). In Cavedon, L., Rao, A., Wobcke, W., eds.: Intelligent Agent Systems: Theoretical and Practical Issues. Volume 1209 of LNAI. Springer Verlag, Berlin (1997) 152–163
53. Singh, M.: Commitments among autonomous agents in information-rich environments. In Boman, M., de Velde, W.V., eds.: Multi-Agent Rationality (Proceedings of MAAMAW'97). Volume 1237 of LNAI. Springer Verlag, Berlin (1997) 141–155
54. 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
55. Hustadt, U., Schmidt, R.: On evaluating decision procedures for modal logics. In Pollack, M., ed.: Proceedings IJCAI'97, Los Angeles (CA), Morgan Kauffman (1997)
56. Graedel, E.: Why is modal logic so robustly decidable? Bulletin of the EATCS **68** (1999) 90–103
57. Vardi, M.: Why is modal logic so robustly decidable? DIMACS Series on Discrete Mathematics and Theoretical Computer Science **31** (1997) 149–184
58. Werner, E., Demazeau, Y., eds.: Decentralized A.I.-3. In Werner, E., Demazeau, Y., eds.: Decentralized A.I.-3, Amsterdam, Elsevier (1992)