Awareness as a vital ingredient of teamwork

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

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

ABSTRACT

For successful coordination and cooperation in a multiagent system, participants need to establish a sufficiently accurate awareness of the current situation. Awareness is understood here as a limited form of consciousness: in the minimal form, it refers to the state of an agent's beliefs about itself, about others and about the environment. When considered in the context of agents' mental states, this leads to distinguishing three levels of awareness: intra-personal (about the agent itself), inter-personal (about other agents as individuals), and group awareness.

Problems in modeling agents' awareness on all three levels are analyzed. It turns out that both the communication medium and agents' cognitive and computational limitations make the achievement of awareness difficult. Cognitive science is used to analyze and explain problems in human awareness, based on the concept of bounded rationality. The BDI framework, originally designed to formally define agents' informational and motivational attitudes, turns out to be also fruitful both for precisely formulating the problems concerning agents' awareness, and, even more importantly, for formulating and comparing possible solutions. Thus, the two fields of cognitive science and MAS mutually benefit from each other's viewpoints, especially in the light of the currently growing need for teamwork in which both computational agents and humans are involved.

In this paper, some possible avenues to solutions for defining and achieving appropriate levels of awareness are suggested. In some cases, these are concrete formal solutions, which have been adopted in our theory of collective motivational attitudes, presented in a number of conference and journal papers [5, 6, 7]. They give rise to more generic solutions that can be of use in any advanced BDI system, especially in those aiming to realize teamwork.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed AI—*multiagent systems*

General Terms

Theory, Human factors, Design

Keywords

teamwork, intra-personal awareness, inter-personal awareness, group awareness, BDI logic

1. INTRODUCTION

Multiagent systems (MAS) as a research area have been created to enable solving of complex problems, that usually are beyond capabilities of individuals (agents or humans), and which usually need expertise and/or capabilities of many different kinds. Very often these problems need *teamwork* to be solved. In the first phase of designing multiagent systems, the emphasis was put on cooperating teams of agents, while nowadays there is a growing need for teams consisting of computational agents as well as humans [19].

For teamwork to be successful, the first need for the participants is to establish a common view of the environment, which can be built by communication, reasoning and various forms of observation, for example of other agents' behavior and changes in the environment. Then, the next steps of Distributed Cooperative Problem Solving (DCPS) need to be implemented. These phases include building the entire variety of individual, social and collective motivational attitudes, as described in [4, 5, 6, 7]. In this paper we will show that in these complex processes which together constitute teamwork, agents' *awareness* about others and the environment is a vital ingredient.

Awareness is understood here as a limited form of consciousness: in the minimal form, it refers to the state of an agent's beliefs about *itself*, about *others* and about *the environment*. These aspects, when considered in the context of agents' mental states, lead to distinguishing different levels of agents' awareness: *intra-personal* (about the agent itself), *inter-personal* (about other agents as individuals), and *group awareness*. For example, an agent's awareness of itself refers to its beliefs about sentences of the form "I (do not) believe φ". In general, an agent's awareness of φ certainly implies that it does not believe that φ is false, but in

AAMAS'06 May 8–12 2006, Hakodate, Hokkaido, Japan.
Copyright 2006 ACM 1-59593-303-4/06/0005 ...$5.00.

many contexts, much stronger kinds of awareness are needed (see subsection 4.1).

In practice of MAS, agents' awareness is expressed in terms of *beliefs*. In the majority of applications, agents make do with belief instead of knowledge for at least the following reasons. First, agents' perception provides the main background for beliefs. In a dynamic unpredictable environment the natural (e.g. computational) limits of perception may give rise to false beliefs or to beliefs that, while true, still cannot be fully justified by the agent. Second, communication channels may be of uncertain quality, so that even if a trustworthy sender knows a certain fact, the receiver may only believe it. In the sequel, we will use the term of awareness in a generic way, sometimes pointing out by what means it is expressed. Usually, there will be different levels of beliefs (see subsection 4.1). However, we will not go into the phase of creating those beliefs, assuming generally that perception, communication and reasoning are basic steps in this process.

This paper is meant as a voice in the discussion on the general subject of formal specification of both static and dynamic aspects of teamwork. In our considerations we will address two issues. Firstly, we will argue that agents' awareness becomes a first class citizen in contemporary multiagent applications. Secondly, we will point out problems in modeling agents' awareness, suggesting some possible avenues to solutions. In some cases, we will present a concrete formal solution. These solutions have been adapted in our theory of collective motivational attitudes, that was presented in a series of journal papers [5, 6, 7] and [1]. However we will try to show that they give rise to more generic solutions, which can be of use in any advanced BDI system, especially in those aiming to realize teamwork.

Our postulates, even though they are formulated in logic, may be particularly interesting for system developers when tailoring a multi-agent system for a specific application, especially if both computational agents and humans are involved in teamwork. When looking for possible solutions, our aim has been to forge synergy between the viewpoints on awareness within teams from the fields of cognitive science and multi-agent systems. Cognitive science has been important in analyzing and explaining problems in human awareness. The BDI framework used in MAS turns out to be fruitful both for formulating the problems around agents' awareness in a very precise way, and, even more importantly, for formulating and comparing possible solutions. Thus, the two fields mutually benefit from each other's viewpoints.

The paper is structured as follows. Section 2 focuses on agents' awareness about their own mental states, and the effects of their bounded rationality. In section 3, attention is given to problematic aspects of agents' models of other individuals' mental states. These strands come together in section 4 where we show that awareness in groups is of vital importance. We show some pitfalls in achieving it, and some possibilities for system developers to flexibly adapt the type of group awareness in a multi-agent system to the environment and the envisioned kind of organization.

2. ON INTRA-PERSONAL AWARENESS

Intra-personal awareness, or consciousness of one's own mental states, also called meta-consciousness, plays an important role in an agent's thinking and reasoning. Such introspection has for long been held as totally unproblematic:

> Consciousness was often viewed as though it was *the* defining feature of human thought. The philosophical traditions that have had the strongest influence on psychology are those of Locke and Descartes, and while these two didn't agree on much, the one proposition they shared was that cognitive states are transparent to introspection [13].

In the second half of the twentieth century, however, cognitive scientists started to study phenomena like implicit cognition, where experimental subjects could correctly recognize well-formed strings of abstract languages by learning from examples, without being able to formulate the complex underlying rule [13]. Thus, humans are often not aware of their own knowledge and beliefs. In this section, we will see how the epistemic logics that are usually used in BDI systems to model agents' knowledge do not always form an accurate model of human cognitive abilities.

2.1 Problems of logical omniscience

The standard logic of beliefs is a modal logic with possible worlds semantics. The formula $\mathrm{BEL}(i, \varphi)$ stands for "agent i believes that φ", and it is defined to be true at a possible world w if φ is true at all worlds which are belief-accessible from w. In this paper, we do not go deeper into the semantics, but see [9, 14].

In order to represent beliefs, in our logical framework for teamwork in MAS [7], we adopt the standard $KD45_n$-system for n agents as explained in [9], containing axioms and rules for a n agents named $i = 1, \ldots, n$. Similar axioms hold in the standard epistemic logic for knowledge (KNOW), except that there, A6 is replaced by the stronger A3, namely $\mathrm{KNOW}(i, \varphi) \to \varphi$ (Veracity of knowledge). Here follows the axiom system $KD45_n$, in which K refers to the basic modal logic (axioms A1, A2 and rules R1, R2), D refers to axiom A6, 45 refers to axioms A4 and A5, and the subscript n refers to the number of agents:

A1 All instantiations of propositional tautologies

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

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

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

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

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

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

A first problem for modeling humans is that, as mentioned above, they often lack positive or negative introspection into their beliefs. As a counter-example to negative introspection, one may be completely unaware of a sentence φ that one doesn't believe, and thus not believe that one does not believe it. A counter-example to positive introspection is formed by the implicit cognition experiments mentioned above. (On introspection for humans, see also [22].) In multi-agent systems, in order for agents to model themselves properly, the developer needs to take care that a modicum of (especially positive) introspection is present.

Another problem of systems based on epistemic logic is that we have the following theorems (similar ones hold for knowledge instead of belief):

$\models \varphi \Rightarrow \models \mathrm{BEL}(i, \varphi)$ (Belief of valid formulas)

$\models \varphi \rightarrow \psi \Rightarrow \models \mathrm{BEL}(i,\varphi) \rightarrow \mathrm{BEL}(i,\psi)$ (closure under valid implication)

These are examples of *logical omniscience*: agents 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 humans nor to computational agents, who have only limited time available. It is unrealistic to assume that they believe every logical theorem, however complicated.

Two belief-related problems of logical omniscience are:

$\mathrm{BEL}(i,\varphi) \rightarrow \neg\mathrm{BEL}(i,\neg\varphi)$ (consistency of beliefs)

$\mathrm{BEL}(i,(\mathrm{BEL}(i,\varphi) \rightarrow \varphi))$ (belief of having no false beliefs)

The first one is problematic because the agent may believe two sentences which are in fact (equivalent to) each other's negation without the agent being aware of it. The second one (which follows from A6 and A5) makes an agent too arrogant about its beliefs: it is not aware of its limitations.

There are several possible solutions to the problems of logical omniscience, involving non-standard semantics or syntactic operators for awareness and explicit belief. Good logical references to the logical omniscience problem and its possible solutions are [14, Chapter 2] and [9, Chapter 9].

2.2 Bounded rationality

According to Herbert Simon, who coined the term *bounded rationality*, "boundedly rational agents experience limits in formulating and solving complex problems and in processing (receiving, storing, retrieving, transmitting) information" [21]. We agree with Simon that models which present humans as logically omniscient or as perfectly rational in the sense of optimizing their own utility are problematic. Moreover, we would propose to extend this discussion to software agents, because they need to reason under bounded rationality as well as humans.

Another example where one needs to be aware of limitations on agents' information occurs in a quantified setting. The distinction de re / de dicto stems from the philosophy of language [18]. A sentence of the form $\exists x \mathrm{BEL}(j, A(x))$ is a *de re* belief attribution which relates j to a *res*, an individual that the belief is about. On the other hand, $\mathrm{BEL}(j, \exists x A(x))$ is a *de dictum* belief attribution, relating j to a *dictum*, namely the proposition $\exists x A(x)$. Luckily, here the distinction is easily modeled in quantified modal logic, where $\exists x \mathrm{BEL}(j, A(x))$ logically implies $\mathrm{BEL}(j, \exists x A(x))$, but not vice versa. For example, it usually takes the detective a whole novel to move from "I believe (or know) that somebody murdered the victim" to "I believe (or know) of a specific suspect that he/she murdered the victim", thus from general awareness that something is wrong, to a much more specific awareness. In subsection 5.3.2, we show that developers also need to take this distinction into account when deciding what kinds of awareness are appropriate for teamwork in the given circumstances.

3. ON INTER-PERSONAL AWARENESS

Bounded rationality plays a role not only in limiting intrapersonal awareness, it also constraints agents' inter-personal awareness. Formal models of human reasoning, such as those in epistemic logic and game theory, assume that humans can faultlessly reason about other people's individual knowledge and beliefs, for example in card games such as happy families [20]. However, recent research in cognitive psychology reveals that adults do not always correctly use their theory of what others know in concrete situations [12, 11].

In Keysar's experiments, some adult subjects could not correctly reason in a practical situation about another person's lack of knowledge (first-order theory of mind reasoning of the form "a does not know p") [12]. Hedden and Zhang, when describing their experiments involving a sequence of dyadic games, suggested that players generally began with first-order reasoning. When playing against first-order co-players, some began to use second-order reasoning (for example, of the form "a does not know that I know that p"), but most of them remained on the first level [11].

In recent experiments by Lisette Mol, it turns out that humans *can* learn to play a version of symmetric Mastermind involving natural language utterances such as "some colors are right". After mastering the first task, namely to play the game according to its rules, many of them learn to perform a second task, namely to develop a winning strategy for the game by using a higher-order theory of mind: "Which sentences reveal the least information while still being true?" "What does the opponent think I am trying to make him think?" [15]. Thus, some awareness of others' mental states can be learned.

3.1 Inter-personal awareness in BDI

Here follow some examples of theorems of the logic of knowledge and belief related to inter-personal reasoning:

$\mathrm{KNOW}(i, \mathrm{KNOW}(j,\varphi)) \rightarrow \mathrm{KNOW}(i,\varphi)$ (Transparency)

This is not realistic, for example a child may know that her father has proved Fermat's last theorem, without knowing the theorem herself (where knowing includes being able to justify it). The transparency problem has been treated by using an alternative semantics of "local worlds" [10]. Also the following theorem may be unrealistic:

$\mathrm{BEL}(j, \mathrm{BEL}(i,\varphi)) \rightarrow \mathrm{BEL}(j, \neg\mathrm{BEL}(k,\neg\varphi))$

(This theorem follows from A6 and R2). The above and similar theorems with even higher stacks of belief operators unrealistically presuppose that agents are constantly aware that other agents monitor the consistency of their beliefs (and follow other epistemic rules) as well as they themselves do, and that others are in their turn aware of still other agents following the logical rules.

4. ON GROUP AWARENESS

When analyzing different aspects of teamwork from the viewpoint of BDI systems, one of the first purposes is to define the scope and strength of motivational and informational attitudes needed for successful team action. These determine the strength and scope of the necessary awareness within a cooperating team.

4.1 Epistemic logic for group awareness

Knowledge, which always corresponds to the facts and can be justified by a formal proof or less rigorous argumentation, is the strongest and therefore preferred informational attitude. The strongest notion of knowledge in a group is *common knowledge*, which is the basis of all conventions and the preferred basis of coordination. Halpern and Moses proved that common knowledge of certain facts is on the one hand necessary for coordination in well-known examples, while on the other hand, it cannot be established by communication if there is uncertainty about the communication channel [9].

4.1.1 General belief and common belief

Common belief is the notion of group belief which is constructed in a similar way as common knowledge, except that a collective belief among a group that φ need not imply that

φ is true. Here follows a short reminder of the axioms governing group beliefs. Let $G \subseteq \{1, \ldots, n\}$ be a group. The formula E-BEL$_G(\varphi)$ (group G has φ as a *general belief*) is meant to stand for "every agent in group G believes φ":

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

C-BEL$_G(\varphi)$ (group G has φ as a *common belief*) is meant to be true if everyone in G believes φ, everyone in G believes that everyone in G believes φ, etc. Let E-BEL$^1_G(\varphi)$ be an abbreviation for E-BEL$_G(\varphi)$, and let E-BEL$^{k+1}_G(\varphi)$ for $k \geq 1$ be an abbreviation of E-BEL$_G$(E-BEL$^k_G(\varphi)$). Thus, we have intuitively C-BEL$_G(\varphi)$ iff E-BEL$^k_G(\varphi)$ for all $k \geq 1$; it turns out that this seemingly infinitary property can be captured by the following (finitary) axiom and rule.

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 system containing individual axioms of subsection 2.1 together with the above axioms and rule is called $KD45^C_n$ [9]. Although using common belief as the intended type of awareness puts less constraints on the communication media than common knowledge, it is still logically highly complex.

4.1.2 Degrees of beliefs important in teamwork

It is well-known that for teamwork, as well as for coordination, it often does not suffice that a group of agents has a general believe about something (E-BEL$_G(\psi)$), but they should collectively believe it (C-BEL$_G(\psi)$). An example is formed by collective actions where the success of each individual agent is vital to the result, for example, lifting a heavy object together or coordinated attack. It has been proved that for such an attack to be guaranteed to succeed, the starting time of the attack must be a collective belief (even common knowledge) for the generals involved [9].

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

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

Parikh has introduced a hierarchy of levels of knowledge between individual knowledge and common knowledge and, together with Krasucki, proved a number of interesting mathematical properties. It turns out that, due to the lack of the truth axiom, the similarly defined hierarchy between individual belief and collective belief is structurally different from the knowledge hierarchy [17].

If even limited orders of theory of mind, as in inter-personal awareness, present such difficulties for humans, it seems that creating group awareness is impossible: reasoning about common belief and common knowledge apparently involves an infinitude of levels. From the time when these notions were first studied, there has been a puzzle about their establishment and assessment, the so-called *Mutual Knowledge Paradox*, most poignantly described in [3]. How can it be that to check whether one makes a felicitous reference when saying "Have you seen the movie showing at the Roxy tonight", one has to check an infinitude of facts about reciprocal knowledge, but people seem to do this in a finite, indeed short, time? Clark's solution for human communication was that such *common ground* (common knowledge) about a sentence can be created if a number of conditions is met, namely co-presence, mutual visibility, mutual audibility, co-temporality, simultaneity, sequentiality, reviewability and revisability. Most of these conditions do not hold in multiagent systems, where agents communicate over non-instantaneous and possibly faulty communication media.

Even though common knowledge cannot in general be established by communication, we have shown that common belief can. In [8] we presented a somewhat tricky procedure that, under some assumptions about the communication channels, trust among group members and temporary persistence of some relevant beliefs (e.g. the group should be aware of the procedure), establishes a common belief C-BEL$_G(\varphi)$. The idea is essentially that one initiator first broadcasts the message φ to all agents in the group, based on a standard low-level communication protocol such as TCP, ensuring that it knows at a certain point that E-BEL$_G(\varphi)$; then the initiator broadcasts the message that C-BEL$_G(\varphi)$ to all of them. We have discussed this procedure and the needed assumptions in [8]. Intuitively, the reason that this procedure can establish common belief, whereas common knowledge can never be established, is that common beliefs need not be true.

5. TUNING AS A METHOD TO EXPRESS DEGREES OF TEAM AWARENESS

The next part of this paper refers to a larger research program, aiming at investigating and, finally, understanding the role of collective motivational attitudes in Distributed Cooperative Problem Solving (DCPS). The analysis resulted, firstly, in a *static* characterization of DCPS, with collective intention and collective commitment as central notions. (A complete theory of collective motivational attitudes in teamwork was presented in [5, 7]). Secondly, an application of this theory in a dynamic and unpredictable environment led to the reconfiguration problem, formally treated in [6].

When building the static and dynamic parts of this theory we've experienced that our theory evolved in the course of the evolution of our thinking about awareness. Over a couple of years this resulted in a generalization and relaxation of some, previously rather strong, conditions. As the role of awareness in agents' setting became more and more clear, this led us to understanding the vital need of *explicitly* expressing awareness in these systems. This conclusion resulted in a tuning schema for collective commitments ([7]), further developed for social (bilateral) commitment and collective intention in the sequel.

While investigating the role of awareness within cooperating teams of agents and humans, let us turn to the commonsense meaning of teamwork. It is clear that there are different *gradations* of being a team. For example, 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, all aspects of teamwork are openly discussed, and team members keep each other informed about relevant changes in the plan. This non-hierarchical teamwork may be contrasted with a group of spies who all work for the same (top secret) goal. In their case a plan is designed by one mastermind, who divides the roles and divulges to each participant *solely* the information that is absolutely necessary for him to do his own part. Thus, members may not know the main goal, 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 former one, we would still like to speak about DCPS, albeit a non-typical case.

These two examples, showing a different gradation of being a team, differ significantly as individual and collective awareness about the ingredients of DCPS (like the main goal and the plan to achieve it) ranges from very high in the first example to very low in the second. One way to express this subtle gradation of awareness is to *characterize* its different degrees, taking into account specific circumstances (e.g. the state of the environment), as well as the application in question. In the majority of applications, different degrees of awareness could be formally expressed in terms of different kinds of beliefs, defined in the previous section.

For efficiency reasons it is often important to minimize the degree of awareness between agents, which usually allows to minimize the level of communication among agents. This degree should be tuned to the circumstances under consideration. Thus in some situations individual belief may suffice, while in other situations, general belief (E-BEL$_G$ of a relevant proposition within a group G) ensures a proper level of awareness, and again in others the strongest notion of common belief is needed. This tuning mechanism is adopted below in our definitions of collective intention, social commitment, and collective commitment. As discussed in [5, 7, 1], collective intentions and collective commitments are the two essential team attitudes that allow teamwork to happen and be successful. While collective intention may be viewed as a sort of glue that holds a group together, collective commitment reflects the concrete manner to achieve the goal in question. Social commitments, finally, represent the concrete agent-to-agent "promises" to fulfill agents' individual actions that make up the team's social plan on which the collective commitment is based.

5.1 Tuning scheme for collective intentions

In strictly cooperative teams the notion of collective intention seems to be rather strong. A necessary condition for a collective intention is that all members i of the team G have the associated individual intention $\mathrm{INT}(i, \varphi)$ towards φ. (Note that individual intentions can be modeled using possible worlds semantics based on intention-accessibility relations similar to those for beliefs.) However, the combination of individual intentions is not sufficient. Imagine that two agents want to achieve the same goal 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 $\text{M-INT}_G(\varphi)$. In order to formalize this condition, $\text{E-INT}_G(\varphi)$ (standing for "everyone intends") is defined first:

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

The mutual intention $\text{M-INT}_G(\varphi)$ is meant to be true if everyone in G intends φ, everyone in G intends that everyone in G intends φ, etc. As we do not have infinite formulas to express this, let $\text{E-INT}^1_G(\varphi)$ be an abbreviation for $\text{E-INT}_G(\varphi)$, and let $\text{E-INT}^{k+1}_G(\varphi)$ for $k \geq 1$ be an abbreviation of $\text{E-INT}_G(\text{E-INT}^k_G(\varphi))$. Thus again we have intuitively that $\text{M-INT}_G(\varphi)$ iff $\text{E-INT}^k_G(\varphi)$ for all $k \geq 1$.

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

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

The resulting system is called $KD_n^{\text{M-INT}_G}$, and it is sound and complete with respect to Kripke models where all n accessibility relations are serial [5].

Furthermore, all members of the team need to be aware of this mutual intention. Adding this condition completes the definition of collective intention.

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

Instantiating the above schema corresponds to tuning the $\textit{awareness}_G$-dials from $\emptyset$, through individual beliefs and different degrees of E-BEL^k_G, to collective belief, and/or analogously for degrees of knowledge.

Even though $\text{C-INT}_G(\varphi)$ seems to be an infinite concept, collective intentions may be established in practice in a finite number of steps. As defined by **M3**, collective intentions are appropriate to model those situations in which communication, in particular announcements, work, especially if one initiator establishes the team. We have showed in detail in [4] how team formation in such an ideal case may actually work in terms of the first two stages of collective problem solving, namely potential recognition and team formation, and how at these stages the proper attitudes are established through dialogues consisting of the appropriate speech acts.

The degree of awareness in the definition of collective intention clearly depends on the circumstances, and varies from just recognizing the situation by perception when communication is difficult or impossible, through simply confirming what situation we deal with, to more complex cases. An example of a collective intention where $awareness_G$ is instantiated as E-BEL_G occurs in a situation where two persons are needed to save a third one from disaster and have severely limited means to communicate (e.g. because of a storm), but can see each other act. In this case the fact that they see each other running toward the victim may be sufficient for them to conclude that the mutual intention is present, thus $\text{M-INT}_G(\varphi) \wedge \text{E-BEL}_G(\text{M-INT}_G(\varphi))$ is achieved. Other examples of variants of collective intentions are discussed in [5].

5.2 Tuning scheme for social commitments

In teamwork, bilateral "promises" to execute one's allocated actions in order to achieve the main goal are very important. Such a *social 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 [2]), need to be enhanced by the condition expressing the agents' awareness about the situation, i.e. about their individual attitudes. In our earlier papers, such awareness was expressed in terms of collective belief [7]. In order to relax the framework to be better adaptable to the environment, here follow the schemes for defining social commitments with respect to actions α and propositions φ:

$$\begin{aligned} \text{COMM}(i,j,\alpha) \leftrightarrow\ & \text{INT}(i,\alpha) \wedge \text{GOAL}(j, \mathit{done}(i,\alpha)) \wedge \\ & \mathit{awareness}_{\{i,j\}}(\text{INT}(i,\alpha) \wedge \text{GOAL}(j, \mathit{done}(i,\alpha))) \end{aligned}$$

$$\begin{aligned} \text{COMM}(i,j,\varphi) \leftrightarrow\ & \text{INT}(i,\varphi) \wedge \text{GOAL}(j, \mathtt{stit}(i,\varphi)) \wedge \\ & \mathit{awareness}_{\{i,j\}}(\text{INT}(i,\varphi) \wedge \text{GOAL}(j, \mathtt{stit}(i,\varphi))) \end{aligned}$$

Here $\mathit{done}(i,\alpha)$ means that agent i has just executed action α and $\mathtt{stit}(i,\varphi)$ means that agent i sees to it (takes care) that φ is achieved. For a formal treatment of *done* and `stit`, see [6]. Instantiating the above schema again corresponds to tuning the $awareness_G$-dials from $\emptyset$, through individual beliefs and different degrees of E-BEL_G^k, to collective belief, and/or analogously for degrees of knowledge.

5.3 Tuning scheme for collective commitments

After a group is constituted on the basis of collective intention, as a next step a *collective commitment* between the team members may be established. While a collective intention may be viewed as an inspiration for team activity, the plan-based 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. The allocation is concluded when agents accept social commitments to realize their individual actions. We claim that it is important for system developers to make appropriate decisions about the type or gradation of teamwork, and consequently the type of agents' awareness, needed to achieve the goal in given circumstances. Thus, it is useful for them to have a mechanism that helps to choose the corresponding type of group commitment. See [1] for a thorough discussion on this matter.

5.3.1 General schema of collective commitment

Clearly, the two examples in the introduction to section 5 cannot be covered by *one* generic type of collective commitment. In [7], we give a generic method for the system developer to tune the type of collective commitment to the application in question, the organizational structure of the group or institution, and to the environment, especially to its communicative possibilities. This generic solution allows to provide a full range of types of collective commitments and weaker group attitudes covering the range from proper teams to more loosely connected groups involved in DCPS. In our generic description we define solely basic ingredients constituting collective commitments, leaving room for case-specific extensions. The obligatory ingredients are related to different aspects of teamwork:

1. Mutual intention $\text{M-INT}_G(\varphi)$ between a group of agents.

   The team is *based* on this attitude, and exists as long as the mutual intention 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 goal φ. The predicate $\mathit{cons}(\varphi, P)$ informally stands for "P is a correct social plan to achieve φ".

   For definitions of social plans and $\mathit{cons}(\varphi, P)$, see [6]. Here it is enough to say that social plan P consists of actions to be executed sequentially or in parallel.

3. Pairwise social commitments $\text{COMM}(i,j,\alpha)$ for actions from the plan.

   The group splits the tasks according to social plan P, and each agent takes on responsibility to do its part by accepting relevant social commitments (see 5.2).

Next to the above ingredients, different degrees of awareness about them may be present in a team. 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':

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

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

Instantiating the above schema again corresponds to tuning the $awareness_G$-dials from $\emptyset$, through individual beliefs and different degrees of general belief, to collective belief, and/or analogously for degrees of knowledge.

5.3.2 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 DCPS. The degree of this awareness varies, as explained before. In the sequel, the strongest version is considered, namely collective belief about considered aspects of DCPS. For this reason it is justified to speak about *collective awareness* in this context, while in other circumstances, the degree of awareness can be weakened by using general belief E-BEL_G (or another E-BEL_G^k) instead of collective belief C-BEL_G. Let us discuss the relevant aspect of building pair-wise commitments, where the notion of awareness plays a crucial role, more in detail. For extensive discussion of other aspects of commitments see [7].

When a plan P as a recipe is in place, then the particular actions α from it need to be allocated to particular team members (i, j) in order to create pairwise social commitments. This way a social team structure is built, and the plan acquires the property of being social. The type of awareness connected with this phase may be twofold.

- A *detailed* collective awareness of each social commitment:

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

  This corresponds to the interpretation *de re.*

- A *global* collective awareness of the bare existence of social commitments:

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

  This corresponds to the interpretation *de dicto.*

Clearly, the distinction de dicto versus de re (of section 2.2) is also fruitful for complex epistemic operators such as collective belief. Note that C-BEL_G in (a) and (b) distributes over conjunction ($\bigwedge_{\alpha \in P}$), so that only the position of C-BEL_G with respect to $\bigvee_{i,j \in G}$ matters. The cognitive multi-agent distinction between both interpretations is the following. In the detailed collective awareness everybody knows every single detail, while in the global collective awareness everybody knows that all details are somehow arranged, and that things are under control.

The above aspects of awareness serve to distinguish different strengths of collective commitments. Exemplar definitions are produced by keeping the $awareness_G$-dial fixed to a choice between $\emptyset$ and collective belief. In [7], we started from the strongest form of collective commitment, fully reflecting the collective aspects of DCPS. Subsequently, some underlying assumptions were relaxed, leading ultimately to weaker notions of team and distributed commitment. In ongoing research, the interrelations between different forms of commitments and management theory are investigated.

6. DISCUSSION AND CONCLUSIONS

In this paper we have shown that agents working as a team depend, for good coordination and cooperation, on awareness of their own and others' mental states, as well as of the group attitudes. We have shown that both communicative possibilities and cognitive and computational limitations make the achievement of awareness difficult. Some examples showed that in BDI logics, used to make high-level specifications of MAS behavior, it is possible to show rather precisely where the limitations lie. It is important for system developers to keep these limitations in mind and to adapt to them.

Let us sum up some of the possible avenues of adaptation the system developer might choose. With respect to intrapersonal awareness, to enable agents' positive and negative introspection, it helps to restrict the language to a limited number of propositional variables, for which agents can easily check whether they believe in them (respectively know them) or not. In order to create a specification framework in which the problems of logical omniscience does not occur, the system developer may adapt a framework of non-normal modal logic as presented in [9] to the application at hand.

With respect to inter-personal awareness, a system developer designing a combined agent-human multiagent system should adapt to the fact that humans can only correctly reason about at most level two theory of mind ("you do not know that I know p"). Thus, it may be useful to model their reasoning about other individuals by a framework of local reasoning adapted from the proposal in [10].

Finally, in order to adapt to the problems in creating group awareness, a system developer may use the tuning mechanisms of section 5 to adapt the kind of awareness needed in collective intentions, social commitments, and collective commitments to one that is appropriate for the environment and kind of organization. This seems to be a rather sensitive mechanism allowing to calibrate the expressive power of the notion in question. Moreover, this is a very flexible method of defining various common-sense notions, as any case-specific extensions can be easily incorporated into the collective intention or collective commitment schema. Again, the resulting definition and its properties can be viewed as a high level specification of a system.

Emma Norling [16] also combines human modeling with BDI systems. She focuses on extending the BDI framework with characteristics of human reasoning such as timing, learning and memory, all understood in a *folk psychological* sense, in order to allow developers of training simulators to model human-like synthetic agents. We, on the other hand, remain within the scope of the BDI framework in that we focus on agents' informational and motivational attitudes, but pose the question where BDI models should be adapted to better model human and computational information processing constraints and cognitive limitations. Thus, we build on *cognitive science* studies of human awareness about themselves, others and the groups they are part of.

The outcome of our considerations on possible solutions can still be combined with some other, for example strictly technical, elements. Also, when creating teams in which humans and agents cooperate, system developers may divide roles in such a way that they make use of strong points of humans, such as common sense knowledge, as well as those of software agents, such as their memory, which enhances monitoring the team's activity.

7. REFERENCES

[1] J. Brzezinski, P. Dunin-Kęplicz, and B. Dunin-Kęplicz.

Collectively cognitive agents in cooperative teams. In M. P. Gleizes, A. Omicini, and F. Zambonelli, editors, *Engineering Societies in the Agents World V, (ESAW 2004): Revised Selected and Invited Papers*, volume 3451 of *LNCS*, pages 191–208, Berlin, 2005. Springer.

[2] C. Castelfranchi. Commitments: From individual intentions to groups and organizations. In V. Lesser, editor, *Proceedings First International Conference on Multi-Agent Systems*, pages 41–48, San Francisco, 1995. AAAI-Press and MIT Press.

[3] H. H. Clark and C. Marshall. Definite reference and mutual knowledge. In A. Joshi, B. Webber, and I. Sag, editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge University Press, 1981.

[4] F. Dignum, B. Dunin-Kęplicz, and R. Verbrugge. Creating collective intention through dialogue. *Logic Journal of the IGPL*, 9:145–158, 2001.

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

[6] B. Dunin-Kęplicz and R. Verbrugge. Evolution of collective commitments during teamwork. *Fundamenta Informaticae*, 56:329–371, 2003.

[7] B. Dunin-Kęplicz and R. Verbrugge. A tuning machine for cooperative problem solving. *Fundamenta Informaticae*, 63:283–307, 2004.

[8] B. Dunin-Kęplicz and R. Verbrugge. Creating common beliefs in rescue situations. In B. Dunin-Keplicz, A. Jankowski, A. Skowron, and M. Szczuka, editors, *Proceedings of Monitoring, Security and Rescue Techniques in Multiagent Systems (MSRAS)*, Advances in Soft Computing, pages 69–84, Berlin, 2005. Springer.

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

[10] P. Gochet and E. Gillet. On Professor Weingartner's contribution to epistemic logic. In *Advances in Scientific Philosophy*, pages 97–115. Rodopi, Amsterdam, 1991.

[11] T. Hedden and J. Zhang. What do you think I think you think? Strategic reasoning in matrix games. *Cognition*, 85:1–36, 2002.

[12] B. Keysar, S. Lin, and D. Barr. Limits on theory of mind use in adults. *Cognition*, 89:25–41, 2003.

[13] L. Litman and A. Reber. Implicit cognition and thought. In K. Holyoak and R. Morrison, editors, *The Cambridge Handbook of Thinking and Reasoning*, pages 431–453. Cambridge University Press, Cambridge, 2005.

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

[15] L. Mol, N. Taatgen, L. Verbrugge, and P. Hendriks. Reflective cognition as secondary task. In B. Bara, L. Barsalou, and M. Bucciarelli., editors, *Proceedings of Twenty-seventh Annual Meeting of the Cognitive Science Society*, pages 1925–1930, Mahwah (NJ), 2005. Erlbaum.

[16] E. Norling. Folk psychology for human modelling: Extending the BDI paradigm. In *International Conference on Autonomous Agents and Multi Agent Systems (AAMAS)*, pages 202–209, New York, 2004. ACM Press.

[17] R. Parikh and P. Krasucki. Levels of knowledge in distributed computing. *Sadhana: Proceedings of the Indian Academy of Sciences*, 17:167–191, 1992.

[18] W. Quine. Quantifiers and propositional attitudes. *Journal of Philosophy*, 53:177–187, 1956.

[19] K. Sycara and M. Lewis. Integrating intelligent agents into human teams. In E. Salas and S. Fiore, editors, *Team Cognition: Understanding the Factors that Drive Process and Performance*, pages 203–232, Washington (DC), 2004. American Psychological Association.

[20] W. van der Hoek and R. Verbrugge. Epistemic logic: A survey. In L. Petrosjan and V. Mazalov, editors, *Game Theory and Applications*, pages 53–94. Nova Science Publishers, vol. 8, New York, 2002.

[21] O. Williamson. The economies of organization: The transaction cost approach. *American Journal of Sociology*, 87:548–577, 1981.

[22] T. Williamson. *Knowledge and its Limits*. Oxford University Press, Oxford, 2000.