Shared Mental Models
A Conceptual Analysis

Catholijn M. Jonker[1], M. Birna van Riemsdijk[1], and Bas Vermeulen[2]

[1] EEMCS, Delft University of Technology, Delft, The Netherlands
{m.b.vanriemsdijk,c.m.jonker}@tudelft.nl
[2] ForceVision, Den Helder, The Netherlands
bas.vermeulen@forcevision.nl

Abstract. The notion of a shared mental model is well known in the literature regarding team work among humans. It has been used to explain team functioning. The idea is that team performance improves if team members have a shared understanding of the task that is to be performed and of the involved team work. We maintain that the notion of shared mental model is not only highly relevant in the context of human teams, but also for teams of agents and for human-agent teams. However, before we can start investigating how to engineer agents on the basis of the notion of shared mental model, we first have to get a better understanding of the notion, which is the aim of this paper. We do this by investigating which concepts are relevant for shared mental models, and modeling how they are related by means of UML. Through this, we obtain a mental model ontology. Then, we formally define the notion of shared mental model and related notions. We illustrate our definitions by means of an example.

1 Introduction

The notion of a shared mental model is well known in the literature regarding team work among humans [6,3,22,21]. It has been used to explain team functioning. The idea is that team performance improves if team members have a shared understanding of the task that is to be performed and of the involved team work.

We maintain that shared mental model theory as developed in social psychology, can be used as an inspiration for the development of techniques for improving team work in (human-)agent teams. In recent years, several authors have made similar observations. In particular, in [27] agents are implemented that use a shared mental model of the task to be performed and the current role assignment to proactively communicate the information other agents need. Also, [25] identify "creating shared understanding between human and agent teammates" as the biggest challenge facing developers of human-agent teams. Moreover, [20,19] identify common ground and mutual predictability as important for effective coordination in human-agent teamwork.

In this paper, we aim to lay the foundations for research on using shared mental model theory as inspiration for the engineering of agents capable of effective teamwork. We believe that when embarking on such an undertaking, it is important to get a better understanding of the notion of shared mental model. In this paper, we do this by investigating which concepts are relevant for shared mental models (Section 2), and

M. De Vos et al. (Eds.): COIN 2010 International Workshops, LNAI 6541, pp. 132–151, 2011.

modeling how they are related by means of UML (Section 3). Through this, we obtain a mental model ontology. Then, we formally define the notion of shared mental model using several related notions (Section 4). We illustrate our definitions by means of an example in Section 5 and discuss related work in Section 7.

2 Exploration of Concepts

This section discusses important concepts related to the notion of shared mental models.

2.1 Working in a Team

An abundance of literature has appeared on working in teams, both in social psychology as well as in the area of multi-agent systems. It is beyond the scope of this paper to provide an overview. Rather, we discuss briefly how work on shared mental models distinguishes aspects of teamwork. Since we are interested in shared mental models, we take their perspective on teamwork for the analyses in this paper. We do not suggest that it is the only (right) way to view teamwork, but it suffices for the purpose of this paper.

An important distinction that has been made in the literature on shared mental models, is the distinction between *task work* and *team work* (see, e.g., [6,22]). Task work concerns the task or job that the team is to perform, while team work concerns what has to be done only because the task is performed by a team instead of an individual agent. In particular, task work mental models concern the equipment (equipment functioning and likely failures) and the task (task procedures and likely contingencies). Team work mental models concern team interaction (roles and responsibilities of team members, interaction patterns, and information flow), and team members (knowledge, skills, and preferences of teammates).

2.2 Mental Models

In order to be able to interact with the world, humans must have some internal representation of the world. The notion of *mental model* has been introduced to refer to these representations. A mental model can consist of knowledge about a physical system that should be understood or controlled, such as a heat exchanger or an interactive device [11]. The knowledge can concern, e.g., the structure and overall behavior of the system, and the disturbances that act on the system and how these affect the system. Such mental models allow humans to interact successfully with the system.

Different definitions of mental models have been proposed in the literature (see, e.g., [9] for a discussion in the context of system dynamics). In this paper, we use the following often cited, functional definition as proposed in [24]:

> Mental models are the mechanisms whereby humans are able to generate descriptions of system purpose and form, explanations of system functioning and observed system states, and predictions of future system states.

Central to this definition is that mental models concern a *system* and that they serve the purpose of *describing, explaining, and predicting the behavior of the system*.

Another important view of mental models was proposed in [17]. The idea proposed there focuses on the way people reason. It is argued that when people reason, they do not use formal rules of inference but rather think about the possibilities compatible with the premises and with their general knowledge. In this paper, we use the definition of [24] because as we will show, it is closely related to the definition of shared mental model that we discuss in the next section.

2.3 Shared Mental Models

Mental models have not only been used to explain how humans interact with physical systems that they have to understand and control, but they have also been used in the context of team work [6,22]. There the *system that mental models concern is the team*. The idea is that mental models help team members predict what their teammates are going to do and are going to need, and hence they facilitate coordinating actions between teammates. In this way, mental models help explain team functioning.

Mental models have received a lot of attention in literature regarding team performance. Several studies have shown a positive relation between team performance and similarity between mental models of team members (see, e.g., [3,22,21]). That is, it is important for team performance that team members have a shared understanding of the team and the task that is to be performed, i.e., that team members have a *shared mental model*. The concept of shared mental model is defined in [6] as:

> knowledge structures held by members of a team that enable them to form accurate explanations and expectations for the task, and, in turn, coordinate their actions and adapt their behavior to demands of the task and other team members.

Shared mental models thus help *describe, explain and predict the behavior of the team*, which allows team members to coordinate and adapt to changes. In [6], it is argued that shared mental model theory does not imply identical mental models, but "rather, the crucial implication of shared mental model theory is that team members hold compatible mental models that lead to common expectations for the task and team."

In correspondence with the various aspects of teamwork as discussed above, it has been argued that multiple different types of shared mental models are relevant for team performance: shared mental models for task work (equipment model and task model) and for team work (team interaction model and team member model) [6,22].

In this paper, we are interested in the notion of shared mental model both in humans and in software agents, but at this general level of analysis we do not distinguish between the two. Therefore, from now on we use the term "agent" to refer to either a human or a software agent.

3 Mental Model Ontology

We start our analysis of the notion of shared mental model by analyzing the notion of mental model. We do this by investigating the relations between notions that are

essential for defining this concept, and provide UML[1] models describing these relations. The UML models thus form a mental model ontology. This means that the models are not meant as a design for an implementation. As such, attributes of and navigability between concepts is not specified. For example, we model that a model concerns a system by placing a relation between the concepts. But that does not mean that if one would build an agent with a mental model of another agent, that the first would be able to navigate to the contents of the mind of the other agent. We have devided the ontology in three figures for reasons of space and clarity of presentation. We have not duplicated all relations in all diagrams to reduce the complexity of the diagrams.

We use UML rather than (formal) ontology languages such as frames [23] or description logics [2], since it suffices for our purpose. We develop the ontology not for doing sophisticated reasoning or as a design for a multi-agent system, but rather to get a better understanding of the essential concepts that are involved and their relations. Also, the developed ontologies are relatively manageable and do not rely on involved concept definitions. We can work out more formal represention in the future when developing techniques that allow agents to reason with mental models.

We present the UML models in three steps. First, since the concept of a mental model refers to systems, we discuss the notion of *system*. Then, since shared mental models are important in the context of teams, we show how a *team* can be defined *as a system*. Following that, we introduce the notion of agent into the picture and show how the notions of agent, system, and mental model are related.

In UML, classes (concepts) are denoted as rectangles. A number of relations can be defined between concepts. The generalization relation is a relation between two concepts that is denoted like an arrow. This relation represents a relationship between a general class and a more specific class. Every instance of the specific class is also an instance of the general class and inherits all features of the general class. A relationship from a class A to class B with an open diamond at side one of the ends is called a shared aggregate, defined here as a part-whole relation. The end of the association with the diamond is the whole, the other side is the part. Because of the nature of this relationship it cannot be used to form a cycle. A composite aggregation is drawn as an association with a black diamond. The difference with a shared aggregation is that in a composite aggregation, the whole is also responsible for the existence, persistence and destruction of the parts. This means that a part in a composite aggregation can be related to only one whole. Finally, a relationship between two concepts that is represented with a normal line, an association, can be defined. The nature of this relationship is written along the relationship. This can either be done by placing the name of the association in the middle of the line or by placing a role name of a related concept near the concept. The role name specifies the kind of role that the concept plays in the relation. Further, numbers can be placed at the ends of the shared aggregation, composite aggregation and associations. They indicate how many instances of the related concepts can be related in one instance of the relationship. Note that we have not duplicated all relations and concepts in all figures. This is done to keep the figures of the separate parts of our conceptualization clean.

[1] `http://www.omg.org/spec/UML/2.2/`

3.1 System

The previous section shows that the concept of a mental model refers to systems. In this section, we further analyze the notion of system in order to use it to define a team as a system. For this purpose, the basic definition provided by Wikipedia[2] suffices as a point of departure: *A system is a set of interacting or independent entities, real or abstract, forming an integrated whole*. This definition captures the basic ingredients of the notion of system found in the literature (see, e.g., [10]), namely static structures within the system as well as the dynamic interrelations between parts of the system.

Our conceptualization of systems is supported by the UML diagram in Figure 1.

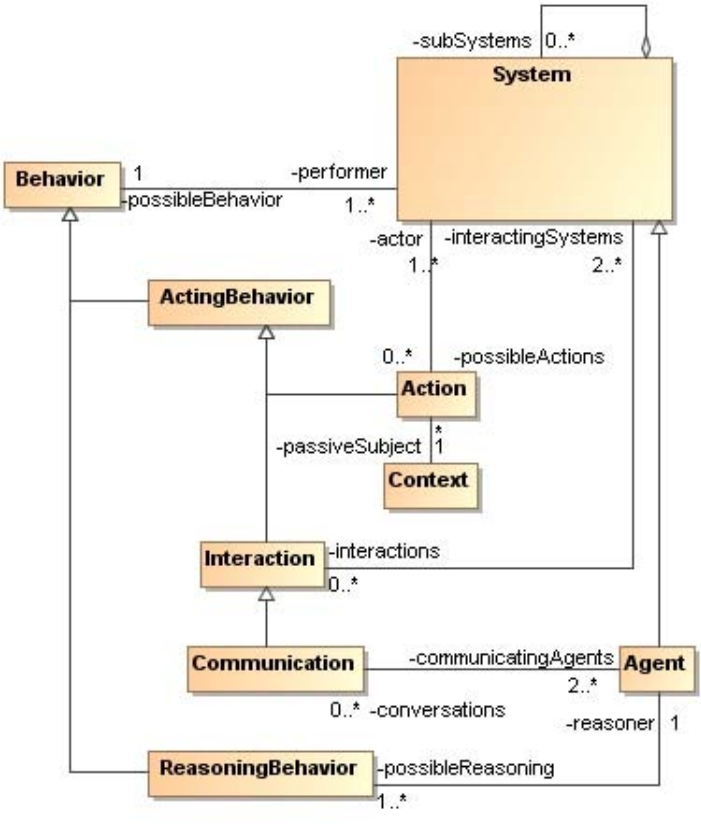

Fig. 1. System

The upper-right corner of the diagram depicts that a system may be a composite, i.e., it may be composed of other systems. This modeling choice makes it easier to define in the following section the notion of team as a system. In particular, the compositionality of the concept system in terms of other systems makes the compositionality of mental

[2] `http://en.wikipedia.org/wiki/System`

models straightforward in the next sections. Regarding the definition, this part addresses the sub-phrase that a system is a set of entities.

The system forms an integrated whole, according to the definition. Therefore, the whole shows behavior. As we do not distinguish between natural or designed systems, living or otherwise, we chose behavior to represent the dynamics of the system as a whole. Note that we further distinguish between reasoning behavior and acting behavior. Not all systems will show both forms of behavior. Acting behavior refers to either actions or interactions. An action is a process that affects the environment of the system and/or the composition of the system itself. Interaction is a process with which a sub-system of the system (or the system as a whole) affects another sub-system of the system. Communication is a special form of interaction, in which the effect of the interaction concerns the information state of the other element. Communication is a term we restricted for the information-based interaction between two agents. The term reasoning behavior is also reserved for agents. The concept "context" refers to both the environment of the system as well as the dynamics of the situation the system is in. The system executes its actions in its context. Thus one context is related to multiple actions.

3.2 Team as a System

The notion of system is central to the definition of mental model. In the context of shared mental models we are especially interested in a certain kind of system, namely a team. According to the definition of system, a team can be viewed as a system: it consists of a set of interacting team members, forming an integrated whole.

As noted above, several aspects are relevant for working in a team. We take as a basis for our model the distinction made in [6,22]. As noted in Section 2.1, we by no means claim that this is the only suitable definition of a team or that it captures all aspects. We start from this research since it discusses teams in the context of shared mental models. The most important realization for the sequel is that we define a team as a system and that it has as a set of team members that are agents. Other aspects of the team definition can be varied if nessecary.

The following aspects are distinguished: *equipment* and *task* (related to task work), and *team interaction* and *team members* (related to team work). In our model, we include these four aspects of working in a team. However, we divide them not into team work and task work, but rather into *physical components* and *team activity*, where team members and equipment are physical components and task and team interaction are team activities The reason for making this distinction is that we argue that physical components can in turn be viewed as systems themselves, while team activities cannot, as reflected by the link from physical components to system in Figure 2 below. Moreover, we make another refinement and make a distinction between a task and *task execution*. We argue that task execution is a team activity, even though a task might be performed by only one team member. The task itself describes what should be executed. The concept task is also linked to equipment, to express the equipment that should be used for executing the task, and to team member, to describe which team members are responsible for a certain task.

We link this conceptualization of the notion of team to the general notion of system of Figure 1 by defining a team activity as a kind of acting behavior, and more specifi-

cally team interaction as a kind of interaction[3]. We see team interaction as interaction induced by executing the team activity. Moreover, by defining that physical components are systems, we can deduce from Figure 1 that they can have interactions with each other. Moreover, by defining a team member as an agent, we can deduce from Figure 1 that team members can have reasoning behavior and that they can communicate. The reasoning of a team is built up from the interaction between team members and the individual reasoning of these team members during the interaction. A fully specified example of two agents Arnie and Bernie that have to cooperate to solve an identification task is provided in [18]. It contains examples of team reasoning.

These considerations are reflected in the UML model of Figure 2.

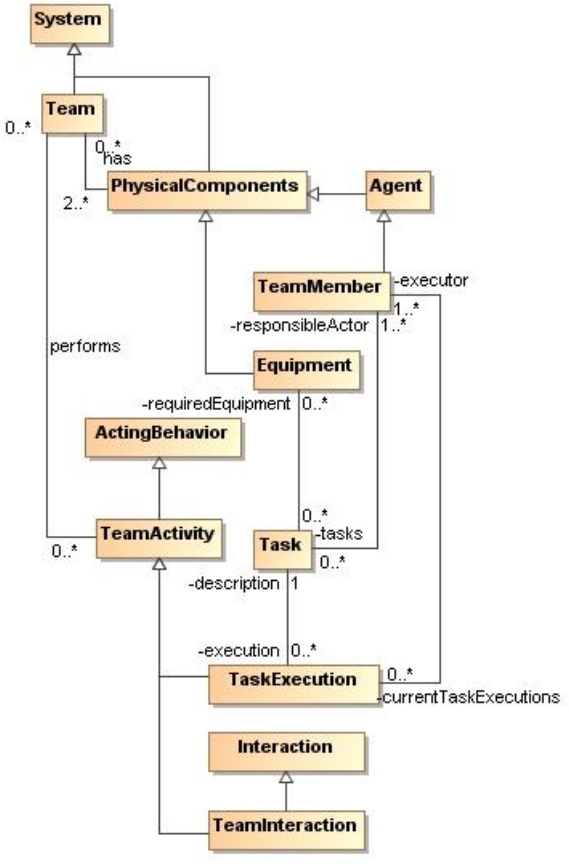

Fig. 2. Team

[3] We could have distinguished "interaction" as a description of an activity from the "performance of the interaction", similarly to the distinction between task and task execution. This is done in the case of task (execution) to be able to express that a team member is responsible for a task, which when executed becomes a team activity. We omit this distinction for interaction for reasons of simplicity.

3.3 Mental Model

Now that we have conceptualized in some detail the notion of system and of a team as a system, we are ready to zoom in on the notion of mental model.

As noted above, mental models are used by humans, i.e., humans have mental models. However, since in this paper we use the notion of agent as a generalization of human and software agent, here we consider that agents have mental models. Moreover, a mental model concerns a system. The basic structure of how mental models are related to systems and agents is thus that an agent has mental models and a mental model concerns a system.

However, we make several refinements to this basic view. First, we would like to express where a mental model resides, namely in the *mind* of an agent. As such, mental models can be contrasted with *physical models*. In order to do this, we introduce the notion of a *model*, and define that physical models and mental model are kinds of models. Both kinds of models can concern any type of system. A nice feature of this distinction is that it allows us to easily express how the notion of *extended mind* [7] is related. The notion of extended mind is being developed in research on philosophy of mind, and the idea is that some objects in the external environment of an agent, such as a diary to record a schedule of meetings or a shared display, are utilized by the mind in such a way that the objects can be seen as extensions of the mind itself. The notion is relevant to research on shared mental models because agents in a team may share an extended mind, and through this obtain a shared mental model [3].

Another aspect that we add to the conceptualization, is the notion of *goal* to express that a mental model is used by an agent for a certain purpose, expressed by the goal of the model.

This is captured in the UML model of Figure 3.

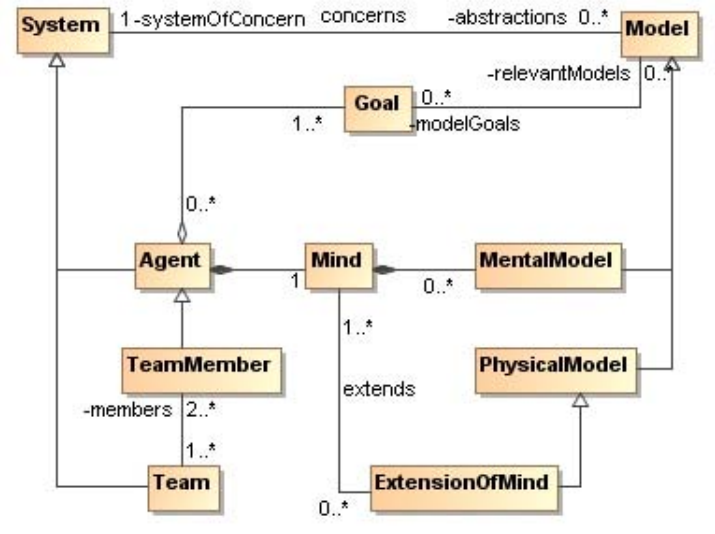

Fig. 3. Mental Model

Given this conceptualization, we can express that an agent can have a mental model of a team. An agent can have a mental model, since it has a mind and a mind can have mental models. A mental model can concern a team, since a mental model is a model and a model concerns a system, and a team is a kind of system. However, since team interaction is not by itself a system (see previous subsection), our model does not allow to express, for example, that the agent has a team interaction mental model. What our conceptualization does allow to express, is that the team mental model has a part that describes team interaction, since the team mental model concerns a team, and a team has team interaction. According to our model, we thus cannot call this part a mental model. However, we will for the sake of convenience refer to that part as a team interaction model (and similarly for the other parts of a team mental model). This is in line with [6,22], where the parts of a team mental model are called mental models themselves. We have modelled the relation between team and team member as a normal association instead of by an aggregation because modelling this relation as an aggregation would mean that an agent's mind is part of a team, which does not conform to intuition.

3.4 Accuracy of Models

In research on shared mental models, the relation of both *accuracy*[4] and *similarity* of mental models to team performance has been investigated [21]. As noted in [22], "similarity does not equal quality - and teammates may share a common vision of their situation yet be wrong about the circumstances that they are confronting".

We suggest that the notions of accuracy and similarity not only have different meanings, but play a different role in the conceptualization of shared mental models. That is, the notion of accuracy of a mental model can be defined by comparing the mental model against some standard or "correct" mental model, i.e., it does not (necessarily) involve comparing mental models of team members. Depending on what one views as a correct model one gets a different notion of accuracy. We have defined two such notions below. The notion of similarity, on the other hand, *does* involve comparing mental models of team members. Although both accuracy and similarity affect team performance [21], we maintain that conceptually, only similarity is to be used for defining the notion of shared mental model. We therefore discuss accuracy informally, and omit the formalizations. We discuss accuracy and similarity with respect to models in general, rather than to only mental models.

We identify two kinds of accuracy, depending on what one takes to compare the model with. The first is what we call *system accuracy*, which assumes that one has a "bird's eye view" of the system and can see all relevant aspects, including the mental models of agents in the system. In general, this is only of theoretical relevance, since one typically has limited access to the various parts of a system[5]. Another notion of accuracy that is easier to operationalize, is *expert accuracy*. In expert accuracy, the idea is to compare a model to an expert model (see e.g. [21] for an example of how to obtain an expert model). Expert accuracy may be defined as the extent to which the model agrees

[4] Here, accuracy is meant in the sense of "freedom from errors", not in the sense of exactness.

[5] In a multi-agent system where one has access to the environment and internal mental states of all agents, one *would* be able to obtain all necessary information.

(see Section 4.2) with the expert model. This then assumes that the expert has a correct model. In research on shared mental models, this is the approach taken to determine accuracy of mental models of team members [21]. That work also describes how this can be operationalized. If the questions we pose to the model should result in a set of answers, then the measures of precision, defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents and recall from the field of information retrieval are good ways to measure the accuracy of the answers [5]. However, in this paper we have only considered questions with single answers.

4 Similarity of Models

As we suggested in the previous section, the essence of the concept of shared mental model is the extent to which agents have *similar* mental models. The word "shared" suggests full similarity, but this is typically not the case. Rather, we propose that *measures* of similarity should be used, which allow the investigation of when models are similar enough for a good team performance, or, in general, good enough for achieving certain goals. We introduce a formal framework in order to be able to express several definitions of notions of similarity. We define sharedness in terms of those notions.

4.1 Formal Framework

The definitions of similarity are based on the concepts and their relations as discussed above. The basic concept that we use in all definitions is *model* (Figure 3). We denote a model typically as M. In this paper, we abstract from the knowledge representation language used for representing the model. Depending on the context, different languages may be chosen. For example, when investigating shared mental models in the context of cognitive agent programming languages (see, e.g., [14]), the knowledge representation language of the respective language can be used. In that context, following Figure 3, the agent is programmed in an agent programming language, it has a mind which is represented by the agent program, this mind can contain mental models which would typically be represented in the so-called mental state of the agent, these mental models concern systems, which can in particular be the team of which the agent is a part.

In order to define to what extent a model is similar to another model, we need to express the content of the model. Depending on which system the model concerns, the content may differ. In particular, in case of mental models concerning a team, the content would represent information about the physical components and activity of the team, which in turn consist of information about equipment and team members, and about task execution and team interaction (Figure 2).

In order to compare models, one could (in principle) inspect the content of these models and compare this content directly. However, this is not always practicable, in particular when considering people: one cannot open up the mind of people to inspect the content of their mental models. Moreover, not all content of a model is always relevant. Depending on what one wants to use the model for, i.e., depending on the

goal for which the model is to be used (Figure 3), different parts of the model may be relevant, or different levels of detail may be needed. For these reasons we propose to use a set of *questions* Q that can be posed to the model in order to determine its contents, thereby treating the model as a black box. For example, a mental model that is to be used for weather predictions should be able to answer a question such as what the weather will be tomorrow in a certain city. A physical model of our solar system should be able to answer a question such as whether the Earth or Mars is closer to the sun.

Choosing an appropriate set of questions is critical for obtaining useful measures of similarity. For example, posing questions about the solar system to a model for weather predictions will not be useful for measuring the similarity of the weather prediction model to another such model. Moreover, posing only questions about whether it will rain to a weather prediction model, will not provide a useful measure of the weather model's similarity to another model in predicting the weather in general. If the model concerns a team, the questions will have to concern the team's physical components and the team activity (Figure 2). With some mental flexibility one can use questions both for mental as well as for physical models, as illustrated by the examples provided above (cf. Figure 3).

Designing a set of questions is also done in research on shared mental models in social psychology. In that work, researchers commonly assess mental models by presenting respondents with a list of concepts and asking them to describe the strength of relationships among the concepts [21,22]. These concepts are carefully chosen based on, for example, interviews with domain experts. The operationalization of our definitions thus requires methods and techniques to determine the appropriate sets of questions Q for the team tasks, respecting the characteristics of the domain/environment in which the team has to function. The methods and techniques we consider important are those for knowledge engineering and elicitation and should take into account social theories about team building and team performance (as partly conceptualized in Figure 2). In the definitions that follow, we abstract from the content of models and assume a set of relevant questions is given. A more thorough investigation of how to define the set of questions is left for future work.

We write $M \vdash \mathit{answer}(a, q)$ to express that M answers a to question q. As usual, we use $|\, s \,|$ to denote the number of elements of a set s. If the model is represented using a logical knowledge representation language, $\vdash$ can be taken to be the entailment relation of the logic. If this is not the case, $\vdash$ should be interpreted more loosely.

4.2 Definitions

In the following, let M_1 and M_2 be models concerning the same system, and let Q be the set of questions identified as relevant for the goal for which M_1 and M_2 are to be used. Let T be a background theory used for interpreting answers. In particular, equivalence is defined with respect to T. For example, the answers "1,00 meter" and "100 centimeter" are equivalent with respect to the usual definitions of units of length.

The first definition of similarity that we provide, is what we call *subject overlap*. Subject overlap provides a measure for the extent to which models provide answers to the set of relevant questions Q. These answers may be different, but at least an answer should be given. We assume that if the answer is not known, no answer is provided.

For example, posing a question about the weather in a certain city to a model of the solar system would typically not yield an answer. Also, we assume that answers are individually consistent.

Definition 1 (subject overlap). *Let the set of questions for which the models provide answers (not necessarily similar answers) be* $OverAns(M_1, M_2, Q) = \{q \in Q \mid \exists a_1, a_2 : M_1 \vdash answer(a_1, q)$ *and* $M_2 \vdash answer(a_2, q)\}$. *Then, we define the level of subject overlap between the model* M_1 *and* M_2 *with respect to set of questions* Q *as* $SO(M_1, M_2, Q) = |OverAns(M_1, M_2, Q)| \,/\, |Q|$.

Since the literature (see Section 2.3) says that shared mental model theory implies that team members hold compatible mental models, we define a notion of compatibility of models. It is defined as the extent to which models do not provide contradictory answers.

Definition 2 (compatibility). *Let the set of questions for which the models provide incompatible answers be* $IncompAns(M_1, M_2, Q) = \{q \in Q \mid \exists a_1, a_2 : M_1 \vdash answer(a_1, q)$ *and* $M_2 \vdash answer(a_2, q)$ *and* $T, a_1, a_2 \vdash \bot\}$. *Then, we define the level of compatibility between the model* M_1 *and* M_2 *with respect to set of questions* Q *as:* $C(M_1, M_2, Q) = 1 - (|IncompAns(M_1, M_2, Q)| \,/\, |Q|)$.

Note that our definition of compatibility does not investigate more complex ways in which the set so determined might lead to inconsistencies. Also note that non-overlapping models are maximally compatible. This is due to the fact that we define incompatibility based on inconsistent answers. If the models do not provide answers to the same questions, they cannot contradict, and therefore they are compatible.

Next, we define *agreement* between models, which defines the extent to which models provide *equivalent* answers to questions.

Definition 3 (agreement). *Let the set of questions for which the models agree be* $AgrAns(M_1, M_2, Q) = \{q \in Q \mid \exists a_1, a_2 : M_1 \vdash answer(a_1, q)$ *and* $M_2 \vdash answer(a_2, q)$ *and* $a_1 \equiv_T a_2\}$. *Then, we define the level of agreement between the model* M_1 *and* M_2 *with respect to set of questions* Q *as:* $A(M_1, M_2, Q) = |AgrAns(M_1, M_2, Q)| \,/\, |Q|$.

These measures of similarity are related in the following way.

Proposition 1 (relations between measures). *We always have that* $A(M_1, M_2, Q) \leq SO(M_1, M_2, Q)$. *Moreover, if* $SO(M_1, M_2, Q) = 1$, *we have* $A(M_1, M_2, Q) \leq C(M_1, M_2, Q)$.

Proof. The first part follows from the fact that $AgrAns(M_1, M_2, Q) \subseteq OverAns(M_1, M_2, Q)$. The second part follows from the fact that if $SO(M_1, M_2, Q) = 1$, all questions are answered by both models. Then we have $AgrAns(M_1, M_2, Q) \subseteq (Q \setminus IncompAns(M_1, M_2, Q))$, using the assumption that answers are consistent.

Next we define what a shared mental model is in terms of the most important characteristics. The model is a mental model, thus it must be in the mind of an agent. Sharedness is defined with respect to a relevant set of questions Q. Furthermore, we have to indicate by which agents the model is shared. The measure of sharedness is defined in terms of the aspects of similarity as specified above.

Definition 4 (shared mental model). *A model M is a mental model that is shared to the extent θ by agents A_1 and A_2 with respect to a set of questions Q iff there is a mental model M_1 of A_1 and M_2 of A_2, both with respect to Q, such that*

1. $SO(M, M_1, Q) = 1$, *and* $SO(M, M_2, Q) = 1$
2. $A(M, M_1, Q) \geq \theta$, *and* $A(M, M_2, Q) \geq \theta$

The definition is easily extendable for handling an arbitrary number n of agents. The definition allows for two important ways to tune it to various situations: varying θ gives a measure of sharedness, varying Q allows to adapt to a specific usage of the model. For example, for some teamwork it is not necessary for every team member to know exactly who does what, as long as each team member knows his own task. This is possible if the amount of interdependencies between sub-tasks is relatively low. For other teamwork in which the tasks are highly interdependent and the dynamics is high, e.g., soccer, it might be fundamental to understand exactly what the others are doing and what you can expect of them. This can also be expressed more precisely by defining expectations and defining sharedness as full agreement of expectations. Making this precise is left for future research.

5 Example: BW4T

In this section, we illustrate the concepts defined in the previous sections using an example from the Blocks World for Teams (BW4T) domain [16]. BW4T is an extension of the classic blocks world that is used to research joint activity of heterogeneous teams in a controlled manner. A team of agents have to deliver colored blocks from a number of rooms to the so-called drop zone in a certain color sequence. The agents are allowed to communicate with each other but their visual range is limited to the room they are in.

We distinguish questions on three levels: *object level*, which concerns the environment (e.g., which blocks are in which rooms, which other agents are there, etc.), *informational and motivational level*, which concerns, e.g., beliefs of agents about the environment, and task allocation and intentions, and *strategic level*, which concerns the reasoning that agents are using to solve problems. These levels correspond to physical components and team activity in Figure 2, and reasoning behavior of agents in Figure 1, respectively.

For the object level, we constructed a set Q of questions regarding, e.g., the number of blocks per color per room, the required color per position in the required color sequence. For example, one can formulate questions such as "How many red blocks are there in room 1?". The answer to such a question is a number that can easily be compared to the answer given by another model. Assuming that there are 12 rooms and 3 colors (white, blue, and red), one can formulate 36 questions of the atomic kind for rooms and the number of blocks per color. Similarly, assuming that the required color sequence (the team task) has 9 positions, one can formulate questions such as "What is the required color at position 1?", leading to 9 questions of this kind (in BW4T the team task is displayed in the environment). In this way, we constructed 36 + 9 questions that refer to the current state of the environment. Note that over time, the situation changes, because the agents move the blocks around.

Suppose room 1 contains 2 red blocks, 2 white blocks and no blue blocks. Furthermore assume, that agent A, having just arrived in room 1 has been able to observe the blocks in this room, whereas agent B is still en route to room 2 and has no idea about the colors of the blocks in the various rooms as yet. Assume that both agents have an accurate picture of the team task (which color has to go to which position). Taking this set of 45 question Q, then we have that the mental model of agent A, M_A, answers 12 questions out of a total of 45, while M_B, the model of agent B only answers 9 questions. The subject overlap is then $SO(M_A, M_B, Q) = 9/45$, and the compatibility is $C(M_A, M_B, Q) = 1$. Also the level of agreement between the models is $A(M_A, M_B, Q) = 9/45$, which in this case equals the subject overlap since the answers do not differ. In order to identify a shared mental model between these agents, we have to restrict the questions to only the part concerning the team task. This model is shared to extent 1. Now, if agent A communicates his findings to agent B, then somewhat later in time the overlap and agreement could grow to 12/45, and the shared mental model would grow when modifying the set of questions accordingly. As the agents walk through the environment, they could achieve the maximum number on measures for these models, as long as they keep informing each other. If this is not done effectively, it may be the case that an agent believes a block to be in a room, while another agent believes it is not there anymore. This would lead to a decreased agreement.

For the informational and motivational level, one may, e.g., formulate the following questions: "Under which conditions should agents inform other agents?" which regards what each agent thinks is the common strategy for the team, and For the task level, one may formulate for each agent A the questions like "What is the preferred task order of agent A?", "Which task does agent A have?", "What is the intention of agent A?", and "What information was communicated by agent A at time X?". Note that the intention of agents changes over time during the task execution, and also X varies over time, thus leading to an incremental number of questions as the team is at work.

For the strategic level, one may consider questions like "Under which conditions should agents inform other agents?". Agent A might answer "An agent communicates when it knows something it knows other agents need to know and everything it intends itself", while B's response may be "An agent communicates when it knows something it knows other agents need to know". The formalizations of these statements could be:

belief(hasTask(Agent,Task)) $\wedge$ *belief(requires(Task,Info))* $\wedge$
hasInfo(self,Info) $\wedge$ *Agent* $\neq$ *self* $\wedge$ *belief*($\neg$ *hasInfo(Agent,Info))*
$\rightarrow$ *toBeCommunicatedTo(Info,Agent))*

intends(self, X) $\wedge$ *belief*($\neg$ *hasInfo(Agent,hasTask(self,X)))*
$\rightarrow$ *toBeCommunicatedTo(hasTask(self,X),Agent)*

This implies higher order aspects of the mental models that these agents need to have, i.e., a good image of what other agents know about the current situation, knowledge about the tasks and their dependence on information, and information about who has what task. For this example domain, this means that the questions need to be extended to include, e.g., "What information is relevant for task T?", and either informational and motivational level questions of the form "How many red blocks does agent A believe to be in room 1?" or strategic questions of the form "When can you be sure that an agent

knows something?", to which an answer could be *observed(Info, self)* $\vee$ *communicatedBy(Info, Agent)*. Note that the complexity of computing the measures of similarity depends heavily on the complexity of the logic underlying the questions and thus the answers to the questions. The operationalization of testing these measures might require advanced logical theorem proving tools or model checkers.

6 Agent Reasoning with Shared Mental Models

The concepts introduced in Section 4 which were illustrated in Section 5, consider similarity between mental models from a *bird's eye* perspective. One could say that questions are posed to the mental models by an outside observer. However, this does not demonstrate how the notion of shared mental model can be *operationalized* and used in agents' reasoning. In this section we sketch the latter, using the Two Generals' Problem [1] (see also `http://en.wikipedia.org/wiki/Two_Generals%27_Problem`). The operationalization is done on the strategic level, with shared mental models in the lower two levels as a result. The aim is not to argue that the way this problem is solved using shared mental models is better than other solutions. The example is used only for illustration purposes.

> Two armies, each led by a general, are preparing to attack a fortified city. The armies are encamped near the city, each on its own hill. A valley separates the two hills, and the only way for the two generals to communicate is by sending messengers through the valley. Unfortunately, the valley is occupied by the city's defenders and there's a chance that any given messenger sent through the valley will be captured. Note that while the two generals have agreed that they will attack, they haven't agreed upon a time for attack before taking up their positions on their respective hills.
>
> The two generals must have their armies attack the city at the same time in order to succeed. They must thus communicate with each other to decide on a time to attack and to agree to attack at that time, and each general must know that the other general knows that they have agreed to the attack plan. Because acknowledgement of message receipt can be lost as easily as the original message, a potentially infinite series of messages is required to come to consensus.

The problem the generals face is that they are aware that they do not have a mental model of the attack time that is shared between them. Thus, the communication stream that they initiate is an attempt to come to a shared mental model and to know that they have a shared mental model.

By introducing the concept of a shared mental model, the problem can be formulated internally within the code of the agents (`gen_a` and `gen_b`) as follows. The notation we use resembles that of the agent programming language GOAL [14], giving an indication of how the reasoning can be programmed in an agent. GOAL uses Prolog for expressing the agent's knowledge, which represents general (static) knowledge of the domain and environment. Goals represent what agents want to achieve. The program section has rules of the form `if <condition> then <action>`, where the condition refers to the beliefs and/or goals of the agent. Percept rules are used to process percepts and/or execute multiple send actions. In each cycle of the agent's reasoning,

all instantations of percept rules are applied (meaning that the actions in the consequent are executed if the conditions in the antecedent hold), after which one action rule of which the condition holds is applied.

```
knowledge{
  conquer(city) :-
    simultaneous_attack.
  simultaneous_attack :-
    attacks_at(gen_a, T), attacks_at(gen_b, T).
  requires(shared_mental_model(attack_planned_at),
           hasInfo(A, attack_planned_at(B, T))).

}

goals{ conquer(city). }

program{
  if a-goal(conquer(city)) then
    adopt(simultaneous_attack) +
    adopt(shared_mental_model(attack_planned_at)).

  if a_goal(G) then insert(hasGoal(self,G)).

  <code to determine attack time T>

  if bel(hasInfo(gen_a, attack_planned_at(gen_a, T))),
     bel(hasInfo(gen_a, attack_planned_at(gen_b, T))),
     bel(hasInfo(gen_b, attack_planned_at(gen_a, T))),
     bel(hasInfo(gen_b, attack_planned_at(gen_b, T)))
  then do(attack_at(T)).
}

perceptrules{
  % the agents perceive the predicate "attacks_at(A,T)"
  % for any agent at the T the attack is performed.

  % Generic reflection rule for informing teammates
  if bel(hasGoal(Agent,Goal)),
     bel(requires(Goal,Info)),
     bel(Info),
     not(Agent = self),
     not(bel(hasInfo(Agent,Info)))
  then sendonce(Agent, Info) + insert(hasInfo(Agent,Info)).

}
```

The knowledge line about conquer city expresses that the city will be successfully conquered if the generals simultaneously attack at some time T and share a mental model with respect to the predicate `attacks_at`. The knowledge line about the requirement of a shared mental model about `attacks_at` explains that all agents A (thus both `gen_a` and `gen_b`) should have information about when all agents B (thus both `gen_a` and `gen_b`) will attack.

The initial goal of conquer city will lead to subsequent goals for the agents to attack simultaneously and to have a shared mental model with respect to the attack time, by applying the first rule in the program section.

The generic reflection rule in the perceptrules section cannot be executed by GOAL directly, but has to be interpreted as a scheme of rules that should be instantiated with concrete predicates for the kind of information to be sent in a specific domain. Using (instantiations of) this rule, the generals will start to inform each other of choices they made regarding the time to attack. This is done based on the goal of having a

shared mental model concerning the attack plan (adopted through applying the first action rule), and the fact that for this certain information is required (as specified in the knowledge base).

The rest of the code of the agents, which is omitted here for brevity, should consist of code to get to the same time T at which they will attack. A simple solution is that e.g., `gen_a` is the boss, and `gen_b` will accept his proposal for the attack time. Once a common time has been established, the generals attack as specified in the last action rule.

Note that the formulation chosen does not require the infinite epistemic chain of `hasInfo` that is part of the thought experiment that the Two Generals' Problem is. Simply put, each of the agents will attack if it believes that it has the same idea about the attack time as the other agent. The agents as formulated above do not reflect again, that both should also share a mental model with respect to the predicate `hasInfo`. This would of course be interesting to model, but will lead to the long, infinitely long, process of informing each other of their plans as is explained in the literature on the Two Generals' Problem. We choose to stop here to explain a possible explicit use of the concept of a shared mental model.

7 Related Work

In this section, we discuss how our work is related to existing approaches to (human-)agent teamwork. An important difference between our work and other approaches is that to the best of our knowledge, few other approaches are based directly on shared mental model theory (see below for an exception). Moreover, our focus is on a conceptualization of the involved notions rather than on reasoning techniques that can be applied directly when developing agent teams, since this is one of the first papers that aims at bringing shared mental model theory to agent research. We believe it is important to get a better understanding of the concepts, thereby developing a solid foundation upon which reasoning techniques inspired by shared mental model theory can be built.

Although most existing approaches to (human-)agent teamwork are not based *directly* on shared-mental model theory, similar ideas have been used for developing these approaches. Many of these approaches advocate an explicit representation of teamwork knowledge (see, e.g., [15,12,26,4]). Such teamwork knowledge may concern, e.g., rules for communication to team members, for example if the execution of a task is not going according to plan, and for establishing a joint plan or recipe on how to achieve the team goal. By making the teamwork representations explicit and implementing agents that behave according to them, agents inherently have a shared understanding of teamwork. Moreover, these representations often incorporate strategies for obtaining a shared view on the concrete team activity that the agents engage in. Jennings [15] and Tambe [26] propose work that is based on joint intentions theory [8]. A joint intention is defined as "a joint commitment to perform a collective action while in a certain shared mental state". The latter refers to an important aspect of a joint intention, which is that team members mutually believe they are committed to a joint activity.

These approaches thus already provide concrete techniques for establishing shared mental models to some extent. However, the notion of shared mental model is implicit

in these approaches. We believe that considering (human-)agent teamwork from the perspective of shared mental models could on the one hand yield a unifying perspective on various forms of shared understanding that are part of existing teamwork frameworks, and on the other hand could inspire new research by identifying aspects related to shared mental models that are not addressed by existing frameworks. An example of the latter is the development of techniques for dealing with an observed lack of sharedness. Existing approaches provide ways of trying to prevent this from occurring, but in real-word settings this may not always be possible. Therefore, one needs techniques for detecting and dealing with mental models that are not shared to the needed extent. This is important, for example in human-agent teamwork where humans cannot be programmed to always provide the right information at the right time.

An approach for agent teamwork that incorporates an explicit notion of shared mental model is [27]. The paper presents an agent architecture that focuses on proactive information sharing, based on shared mental models. An agent in this architecture is composed of several models, including an individual mental model and a shared mental model. The individual mental model stores beliefs (possibly including beliefs about others) and general world knowledge. The shared mental model stores information and knowledge shared by all team members. This concerns information about the team structure and process, and dynamic information needs such as the progress of teammates.

This notion of shared mental model differs from ours. In particular, while we do consider mental models to be part of agents' minds (Figure 3), we do not consider a shared mental model to be a component of an agent. Rather, we suggest that the essence of the notion of shared mental model is the extent to which agents have similar mental models, i.e., a shared mental model is a mental model that is shared to some extent between agents. We thus consider shared mental model a derived concept which expresses a property of the relation between mental models, rather than an explicit component inside an agent. This makes our notion fundamentally different from the one proposed by [27].

An approach for representing mental models of other agents in agent programming is proposed in [13]. In that work, mental states of agent are represented by means of beliefs and goals, as is common in cognitive agent programming languages. Besides the agent's own mental state, an agent has mental models for the other agents in the system, which consist of the beliefs and goals the agent thinks other agents have. These are updated through communication. For example, if an agent A informs another agent B of some fact p, agent B will update its model of A to include that agent A believes p (assuming agents do not send this information if they do not believe it). A similar mechanism applies to goals. This approach can be extended by applying similarity measures on the mental state of the agent and of the mental models it has of other agents, to determine what should be communicated.

8 Conclusion

In this paper, we have studied the notion of shared mental model, motivated by the idea of taking shared mental model theory as inspiration for the engineering of agents capable of effective teamwork. We have analyzed the notion starting from an analysis

of the notion of mental model, and continuing with definitions of similarity of models, leading to a definition of shared mental model. We have illustrated how these definitions can be operationalized using an example in the BW4T domain.

As for future work, there are conceptual as well as engineering challenges. We aim to investigate how theory of mind (agents that have mental models about other agents) fits into this framework. We will study in more detail models of agent teamwork in which a notion of sharedness plays a role (e.g., [15,12,26,4]), and analyze how these approaches compare to our notion of shared mental model. As in joint intentions theory, awareness of sharedness may be relevant for effective teamwork and worth investigating from the perspective of shared mental models. From an engineering perspective, a main challenge for future research is the investigation of mechanisms that lead to a shared mental model that is shared to the extent needed for effective teamwork, which may also depend on the kind of task and environment. A thorough comparison of existing approaches for agent teamwork with our notion of shared mental model will form the basis for this.

References

1. Akkoyunlu, E., Ekanadham, K., Huber, R.: Some constraints and tradeoffs in the design of network communications. In: Proceedings of the Fifth ACM Symposium on Operating Systems Principles (SOSP 1975), pp. 67–74. ACM, New York (1975)
2. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F.: The description logic handbook: Theory, implementation, and applications. Cambridge University Press, Cambridge (2003)
3. Bolstad, C., Endsley, M.: Shared mental models and shared displays: An empirical evaluation of team performance. Human Factors and Ergonomics Society Annual Meeting Proceedings 43(3), 213–217 (1999)
4. Bradshaw, J., Feltovich, P., Jung, H., Kulkami, S., Allen, J., Bunch, L., Chambers, N., Galescu, L., Jeffers, R., Johnson, M., Sierhuis, M., Taysom, W., Uszok, A., Hoof, R.V.: Policy-based coordination in joint human-agent activity. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, pp. 2029–2036 (2004)
5. Buckland, M., Gey, F.: The relationship between recall and precision. Journal of the American Society for Information Science 45(1), 12–19 (1994)
6. Cannon-Bowers, J.A., Salas, E., Converse, S.: Shared mental models in expert team decision making. In: Castellan, N.J. (ed.) Individual and Group Decision Making, pp. 221–245. Lawrence Erlbaum Associates, Mahwah (1993)
7. Clark, A., Chalmers, D.J.: The extended mind. Analysis 58, 10–23 (1998)
8. Cohen, P., Levesque, H.: Teamwork. Nous, 487–512 (1991)
9. Doyle, J., Ford, D.: Mental models concepts for system dynamics research. System Dynamics Review 14(1), 3–29 (1998)
10. Francois, C.: Systemics and cybernetics in a historical perspective. Systems Research and Behavioral Science 16, 203–219 (1999)
11. Gentner, D., Stevens, A.: Mental Models. Lawrence Erlbaum Associates, New Jersey (1983)
12. Grosz, B., Kraus, S.: Collaborative plans for complex group action. Journal of Artifical Intelligence 86(2), 269–357 (1996)
13. Hindriks, K., van Riemsdijk, M.B.: A computational semantics for communicating rational agents based on mental models. In: Braubach, L., Briot, J.-P., Thangarajah, J. (eds.) ProMAS 2009. LNCS (LNAI), vol. 5919, pp. 31–48. Springer, Heidelberg (2010)

14. Hindriks, K.V.: Programming rational agents in GOAL. In: Bordini, R.H., Dastani, M., Dix, J., El Fallah Seghrouchni, A. (eds.) Multi-Agent Programming: Languages, Tools and Applications. Springer, Berlin (2009)
15. Jennings, N.: Controlling cooperative problem solving in industrial multi-agent systems using joint intentions. Artificial Intelligence Journal 74(2) (1995)
16. Johnson, M., Jonker, C., van Riemsdijk, M.B., Feltovich, P.J., Bradshaw, J.M.: Joint activity testbed: Blocks world for teams (BW4T). In: Aldewereld, H., Dignum, V., Picard, G. (eds.) ESAW 2009. LNCS, vol. 5881, pp. 254–256. Springer, Heidelberg (2009)
17. Johnson-Laird, P.N.: Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness. Cambridge University Press, Cambridge (1983)
18. Jonker, C., Treur, J.: Compositional verification of multi-agent systems: a formal analysis of pro-activeness and reactiveness. International Journal of Cooperative Information Systems 11, 51–92 (2002)
19. Klein, G., Feltovich, P., Bradshaw, J., Woods, D.: Common ground and coordination in joint activity. In: Organizational Simulation, pp. 139–184 (2004)
20. Klein, G., Woods, D.D., Bradshaw, J.M., Hoffman, R.R., Feltovich, P.J.: Ten challenges for making automation a "team player" in joint human-agent activity. IEEE Intelligent Systems 19(6), 91–95 (2004)
21. Lim, B., Klein, K.: Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy. Journal of Organizational Behavior 27(4), 403 (2006)
22. Mathieu, E., Heffner, T.S., Goodwin, G., Salas, E., Cannon-Bowers, J.: The influence of shared mental models on team process and performance. The Journal of Applied Psychology 85(2), 273–283 (2000)
23. Minsky, M.: A framework for representing knowledge. The Psychology of Computer Vision (1975)
24. Rouse, W., Morris, N.: On looking into the black box: Prospects and limits in the search for mental models. Psychological Bulletin 100(3), 349–363 (1986)
25. Sycara, K., Sukthankar, G.: Literature review of teamwork models. Technical Report CMU-RI-TR-06-50, Carnegie Mellon University (2006)
26. Tambe, M.: Towards flexible teamwork. Journal of Artificial Intelligence Research 7, 83–124 (1997)
27. Yen, J., Fan, X., Sun, S., Hanratty, T., Dumer, J.: Agents with shared mental models for enhancing team decision makings. Decision Support Systems 41(3), 634–653 (2006)