Shared Mental Models: A Conceptual Analysis

Catholijn M. Jonker
EEMCS, TU Delft
Delft, The Netherlands
c.m.jonker@tudelft.nl

M. Birna van Riemsdijk
EEMCS, TU Delft
Delft, The Netherlands
m.b.vanriemsdijk@tudelft.nl

Bas Vermeulen
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 [3, 2, 13, 12]. 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 [16] 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, [15] identify "creating shared understanding between human and agent teammates" as the biggest challenge facing developers of human-agent teams. Moreover, [11] 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, 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 using several related notions. We illustrate our definitions by means of an example.

Cite as: The title of your paper should be written here, Author1, Author2 and Author3, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. XXX-XXX.

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., [3, 13]). Task work concerns the task or job that the team is to perform, while team work concerns what has to be done in order to complete a task as a team. In particular, task work concerns the equipment (equipment functioning and likely failures) and the task (task procedures and likely contingencies). Team work concerns 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 [7]. 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., [5] for a discussion in the context of system dynamics). In this paper, we use the following often cited, functional definition as proposed in [14]:

> 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 [10]. 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 [14] 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 [3, 13]. 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., [2, 13, 12]). 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 [3] 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 [3], 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) [3, 13].

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.

We use UML rather than formal ontology languages such as description logics [1], since it suffices for our purpose. We develop the ontology not for doing sophisticated reasoning, but rather to get a better understanding of the essential concepts that are involved. Also, the developed ontologies are relatively manageable and do not rely on involved concept definitions.

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.

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., [6]), 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.

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 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 defi-

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

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

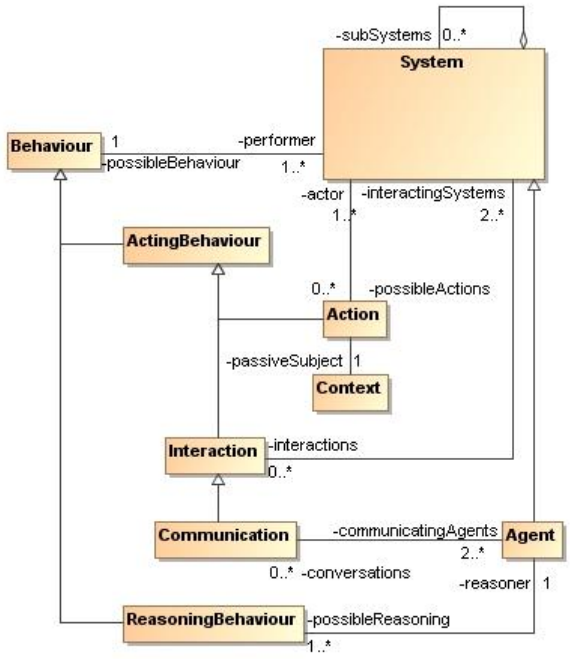

Figure 1: System

nition. 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. Actions are executed in a certain context.

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 [3, 13]. 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 [3, 13] since they discuss teams in the context of shared mental models. The most important realization for the sequel is that we define a team as a system. The framework can be instantiated with other definitions of team if needed.

In [3, 13], 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 specifically team interaction as a kind of interaction. We see team interaction as interaction induced by executing the team activity. Moreover, by defining that physical components are the subsystems of a team, we can deduce from Figure 1 that interaction occurs among physical components. Moreover, by defining a team member as an agent, we can deduce from Figure 1 that it is the team members that have the reasoning behavior and that can communicate.

These considerations are reflected in the UML model below.

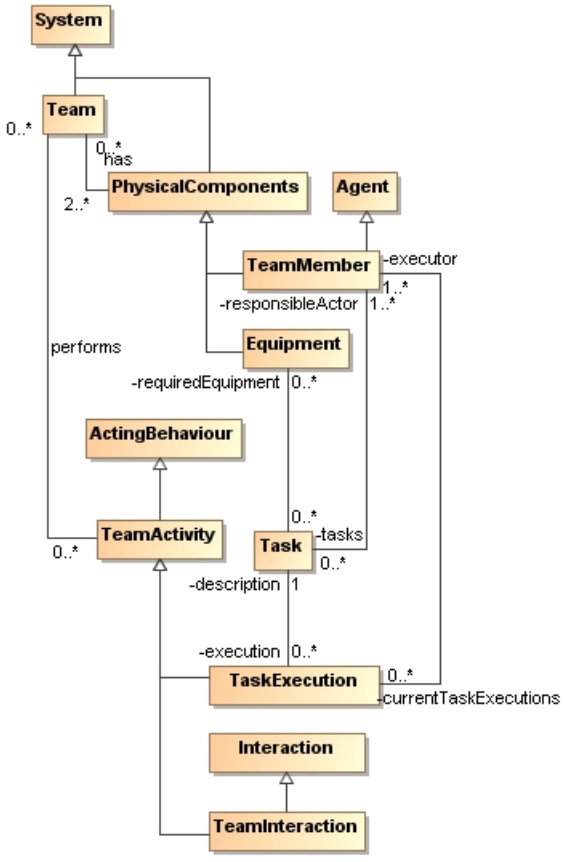

Figure 2: Team

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. A nice feature of this distinction is that it allows us to easily express how the notion of *extended mind* [4] 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 [2].

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 below.

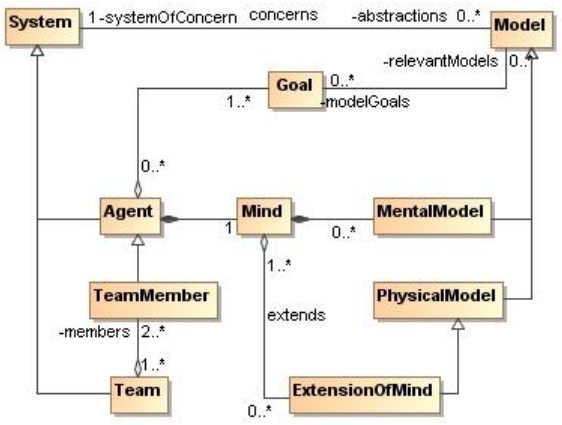

Figure 3: Mental Model

Given this conceptualization, we can express that an agent has 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 [3, 13], where the parts of a team mental model are called mental models themselves.

3.4 Accuracy of Models

In research on shared mental models, the relation of both *accuracy*[3] and *similarity* of mental models to team performance has been investigated [12]. As noted in [13], "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. The notion of similarity, on the other hand, *does* involve comparing mental models of team members. Although both accuracy and similarity affect team performance [12], we maintain that conceptually, only similarity is to be used for defining the notion of shared mental model. For reasons of space, 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.[4] 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. In research on shared mental models, this is the approach taken to determine accuracy of mental models of team members [12].

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 pro-

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

[4]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.

gramming languages (see, e.g., [8]), the knowledge representation language of the respective language can be used.

In order to define to what extent a model is similar to another model, we need to express the content of the model. Rather than considering the entire model, we focus on those aspects of the model that are relevant for the *goal* for which the model is to be used (Figure 3). In order to identify what the model has to say with respect to aspects relevant for the goal, we propose to use *questions* that can be posed to the model. A set of questions is typically denoted by Q. 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. We write $M \vdash \mathit{answer}(a, q)$ to express that M answers a to question q. As usual, we use $\mid s \mid$ 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.

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. A similar issue also arises 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 [12, 13]. 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.

We propose to use questions to identify the content of models because we believe it can be applied naturally to software agents and human agents alike (see the example in the sequel). Asking agents to describe relationships among concepts is more difficult to translate to software agents, unless they are endowed with capabilities for ontological reasoning. Moreover, with some mental flexibility one can use questions both for mental as well as for physical models, as illustrated by the examples provided above.

4.2 Definitions

In the following, let M_1 and M_2 be models of systems S, 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* $\mathit{OverAns}(M_1, M_2, Q) = \{q \in Q \mid \exists a_1, a_2 : M_1 \vdash \mathit{answer}(a_1, q)$ *and* $M_2 \vdash \mathit{answer}(a_2, q)\}$. *Then, we define the level of subject overlap between the model* M_1 *and* M_2 *with respect to set of questions* Q *as* $\mathit{SO}(M_1, M_2, Q) = \mid OverAns(M_1, M_2, Q) \mid / \mid Q \mid$.

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* $\mathit{IncompAns}(M_1, M_2, Q) = \{q \in Q \mid \exists a_1, a_2 : M_1 \vdash \mathit{answer}(a_1, q)$ *and* $M_2 \vdash \mathit{answer}(a_2, q)$ *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:*
$\mathit{C}(M_1, M_2, Q) = 1 - (\mid IncompAns(M_1, M_2, Q) \mid / \mid Q \mid)$.

Note that our definition of compatibility does not investigate more complex ways in which the so determined set 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* $\mathit{AgrAns}(M_1, M_2, Q) = \{q \in Q \mid \exists a_1, a_2 : M_1 \vdash \mathit{answer}(a_1, q)$ *and* $M_2 \vdash \mathit{answer}(a_2, q)$ *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:*
$\mathit{A}(M_1, M_2, Q) = \mid AgrAns(M_1, M_2, Q) \mid / \mid Q \mid$.

These measures of similarity are related in the following way.

PROPOSITION 1 (RELATIONS BETWEEN MEASURES). *We always have that* $\mathit{A}(M_1, M_2, Q) \leq \mathit{SO}(M_1, M_2, Q)$. *Moreover, if* $\mathit{SO}(M_1, M_2, Q) = 1$, *we have* $\mathit{A}(M_1, M_2, Q) \leq \mathit{C}(M_1, M_2, Q)$.

PROOF. The first part follows from the fact that $\mathit{AgrAns}(M_1, M_2, Q) \subseteq \mathit{OverAns}(M_1, M_2, Q)$. The second part follows from the fact that if $\mathit{SO}(M_1, M_2, Q) = 1$, all questions are answered by both models. Then we have $\mathit{AgrAns}(M_1, M_2, Q) \subseteq (Q \setminus \mathit{IncompAns}(M_1, M_2, Q))$, 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 [9]. 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 fill a number of bins with colored blocks that they have to pick up in separate rooms as quickly as a possible. Each bin is to be filled with blocks of a specific color in a specific order. The agents are allowed to communicate with each other but their visual range is limited to the room they are in.To perform this task effectively, the agents have to share a mental model on the order in which tasks are performed, when to communicate, the current task allocation, current location of blocks etc.

The system in our example consists of the whole environment, i.e. the rooms with the blocks, the corridors between the rooms, the bins and the agents. For this system we constructed a set Q of questions regarding, e.g., the current time, the number of blocks per color per room, the required color per position in the three bins, the knowledge about communication requirements, tasking of agents and previous communications. The questions are formulated in such a way that the answer is atomic in the sense that it is not composed of answers to sub-questions.

For example, we formulated 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. Given that there are 12 rooms and 3 colors (white, blue, and red), we formulated 36 questions of the atomic kind for rooms and the number of blocks per color. Similarly, for the three bins, each having three positions, we formulated questions such as "What is the required color at position 1 in bin 1?", leading to 9 questions of this kind. 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 per bin). Taking this set of 45 question Q, then we have that the mental model of agent A, M_A, answers 13 questions out of a total of 45, while M_B, the model of agent B only answers 9 questions. The subject overlap is $SO(M_A, M_B, Q) = 9/46$, and the compatibility is $C(M_A, M_B, Q) = 1$. Also the level of agreement between the models is $A(M_A, M_B, Q) = 9/46$, which in this case equal 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 13/46, 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.

Above, we have considered only questions related to the environment and to the team task, which in this case is also visible in the environment. Of another level are the questions that provide insight into the agents, their tasks, intentions and communication strategies. For this 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 per agent the following questions "What is the preferred task order of agent A?", "Which task does agent A have?" , "What is the intention of agent A?". Note that the intention of agents changes over time during the task execution. Finally, we can pose general questions such as "What information was communicated by agent A at time X?", where of course X varies over time, thus leading to an incremental number of questions as the team is at work.

For example, consider that regarding the question "Under which conditions should agents inform other agents?" agent A would answer "An agent communicates when it knows something it knows other agents need to know and everything it intends itself", while B's response would be "An agent communicates when it knows something it knows other agents need to know". This implies higher order aspects of the mental models 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: "Which task T does agent A have?", "What information is relevant for task T?", and either object level questions of the form "How many red blocks does agent A believe to be in room 1?" or higher level questions of the form "When can you be sure that an agent knows something?". 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. 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.

Also, awareness of sharedness may be relevant for effective teamwork and worth investigating. 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 will also depend on the kind of task and environment.

7. REFERENCES

[1] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider. *The description logic handbook: Theory, implementation, and applications*. Cambridge University Press, 2003.

[2] C. Bolstad and M. Endsley. 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.

[3] J. A. Cannon-Bowers, E. Salas, and S. Converse. Shared mental models in expert team decision making. In N. J. Castellan, editor, *Individual and group decision making*, pages 221–245. Lawrence Erlbaum Associates, 1993.

[4] A. Clark and D. J. Chalmers. The extended mind. *Analysis*, 58:10–23, 1998.

[5] J. Doyle and D. Ford. Mental models concepts for system dynamics research. *System Dynamics Review*, 14(1):3–29, 1998.

[6] C. Francois. Systemics and cybernetics in a historical perspective. *Systems Research and Behavioral Science*, 16:203–219, 1999.

[7] D. Gentner and A. Stevens. *Mental Models*. Lawrence Erlbaum Associates, New Jersey, 1983.

[8] K. V. Hindriks. Programming rational agents in GOAL. In R. H. Bordini, M. Dastani, J. Dix, and A. El Fallah Seghrouchni, editors, *Multi-Agent Programming: Languages, Tools and Applications*. Springer, Berlin, 2009.

[9] M. Johnson, C. Jonker, M. B. van Riemsdijk, P. J. Feltovich, and J. M. Bradshaw. Joint activity testbed: Blocks world for teams (BW4T). In *Proceedings of the Tenth International Workshop on Engineering Societies in the Agents' World (ESAW'09)*, volume 5881 of *LNAI*, pages 254–256. Springer, 2009.

[10] P. N. Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge University Press, Cambridge, 1983.

[11] G. Klein, D. D. Woods, J. M. Bradshaw, R. R. Hoffman, and P. J. Feltovich. Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intelligent Systems*, 19(6):91–95, 2004.

[12] B. Lim and K. Klein. 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.

[13] E. Mathieu, T. S. Heffner, G. Goodwin, E. Salas, and J. Cannon-Bowers. The influence of shared mental models on team process and performance. *The Journal of Applied Psychology*, 85(2):273–283, 2000.

[14] W. Rouse and N. Morris. On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin*, 100(3):349–363, 1986.

[15] K. Sycara and G. Sukthankar. Literature review of teamwork models. Technical Report CMU-RI-TR-06-50, Carnegie Mellon University, 2006.

[16] J. Yen, X. Fan, S. Sun, T. Hanratty, and J. Dumer. Agents with shared mental models for enhancing team decision makings. *Decision Support Systems*, 41(3):634–653, 2006.