Towards an Approach for Modelling Uncertain Theory of Mind in Multi-Agent Systems

Ștefan Sarkadi[1], Alison R. Panisson[2], Rafael H. Bordini[2], Peter McBurney[1], and Simon Parsons[1]

[1] King's College London, Department of Informatics, London, UK
[2] PUCRS, School of Technology, Porto Alegre, Brazil
{stefan.sarkadi,peter.mcburney,simon.parsons}@kcl.ac.uk,
alison.panisson@acad.pucrs.br, rafael.bordini@pucrs.br

Abstract. Applying Theory of Mind to multi-agent systems enables agents to model and reason about other agents' minds. Recent work shows that this ability could increase the performance of agents, making them more efficient than agents that lack this ability. However, modelling others agents' minds is a difficult task, given that it involves many factors of uncertainty, e.g., the uncertainty of the communication channel, the uncertainty of reading other agents correctly, and the uncertainty of trust in other agents. In this paper, we explore how agents acquire and update Theory of Mind under conditions of uncertainty. To represent uncertain Theory of Mind, we add probability estimation on a formal semantics model for agent communication based on the BDI architecture and agent communication languages.

Keywords: Multi-Agent Systems, Theory of Mind, Uncertainty, Socially-Aware AI.

1 Introduction

It is reasonable to expect that agents could be more effective at achieving their goals during social interactions with other agents if they understand the other agents involved. However, understanding other agents requires the capability of modelling and reasoning about other agents' mental attitudes. These characteristics are intrinsic to Theory of Mind (ToM) [12].

Normally, agents operate under conditions of uncertainty due to the dynamism of the environments in which they are situated [45]. Modelling other agents' minds also involves uncertainty. ToM involves uncertainty not only due to the dynamism of other agents' mental attitudes, e.g., agents might change their beliefs constantly, but also because it involves the uncertainty of a message reaching its audience (i.e., the uncertainty of the communication channels working properly), the uncertainty of other (autonomous) agents telling/acting truly, and the uncertainty of an agent reading other agents' mental attitudes correctly during interactions.

Various studies have investigated the application of ToM in Multi-Agent Systems (MAS). Among them, [9, 8] investigated the advantages of using different levels of ToM in games played by agents, [33] investigated the role of ToM in modelling dishonest attitudes in MAS, and [3, 15, 16, 27, 38] show the advantages of modelling the opponent when considering strategies in argumentation-based dialogues, even though ToM is not represented explicitly. It seems that modelling other agents' minds is an important topic of research, and existing results show important contributions to MAS.

However, as described in [42], most of the work considering the modelling of other agents' minds assume such model as a given, which is an understandable assumption due to the complexity of the problem, but unrealistic. Unfortunately, the question of how to represent the uncertainty of beliefs about others' beliefs when agents acquire and update the model of other agents' minds, i.e., uncertain ToM, has not been fully investigated in the literature. We believe that agents should also to be able to reason and make decisions using ToM. Therefore, taking inspiration from others who have investigated the use of others agents' model during reasoning and decision-making, e.g., [9, 8, 3, 15, 16, 27, 38], we propose an approach to model ToM in software agents that reflects the uncertainty present in agent communication. We also took some inspiration from the STAPLE language. STAPLE (Social and Team Agents Programming Language) has its semantics based on joint intention theory [19]. STAPLE has the goal of reaching a fault-tolerant approach to program teamwork, in which the authors argue that a team is more than a collection of individuals working together to achieve a common goal. The agents in a team must have a shared goal as well as a shared mental state [21]. Thus, STAPLE enables agents to specify the models of other agents, as well temporal properties of actions and events, allowing them to reason about group beliefs, team intentions, and team commitments [20]. [3]

Our first contribution is the proposal of an approach to model ToM that reflects the uncertainty of information that agents infer about other agents' minds through communication. To the best of our knowledge, our work is the first to propose a formal model of how agents acquire and update ToM during communication in MAS given the uncertainty of other agents' model, particularly in the practical context of a BDI based Agent-Oriented Programming Language (AOPL). This approach allows us to implement multi-agent communication that reflects some desired properties from communication and common knowledge theories [6]. For example, how agents increase the certainty of their ToM by communicating more and, consequently, how communicating more reinforces the already existing model of the mental attitudes of those agents. Our second contribution is showing how agents may use our approach to reach (or not) shared beliefs under conditions of uncertainty, and how agents make decisions using ToM and probabilistic reasoning.

[3] Note that our approach is more general than that, in which ToM could be used to implement similar approaches for teamwork, which is a likely research direction for our work.

2 Background

2.1 Theory of Mind and The Problem of Other Minds

ToM is the ability of humans to ascribe elements such as beliefs, desires, and intentions, and relations between these elements to other human agents. In other words, it is the ability to form mental models of other agents [1]. There are two major theories about ToM. One theory of ToM is the Theory-Theory of Mind (henceforth TT). TT can be described as a theory based approach for assigning states to other agents. While some argue TT is nothing else but folk psychology, others say that it is a more scientific way of mindreading [13]. The other major theory is Simulation Theory of Mind (henceforth ST), which is described to be 'process-driven rather than theory-driven' [2]. In other words, ST emphasises the process of putting oneself into another's shoes. TT argues for a hypothesis testing method of model extraction, whereas ST argues for a simulation based method for model selection.

An important factor that influences the acquisition, formation, or modification of ToM is uncertainty. Inferring the beliefs of others is a notorious epistemological issue named by philosophers *The Problem of Other Minds* [17]. The problem still stands since the times of Descartes [23]. It would be unreasonable for one to assume that ToM is absolute or that ToM is a universal set of beliefs shared by all agents in a system. Therefore, we believe that a reasonable approach to model how ToM is acquired and updated by artificial agents has to be able to represent the uncertainty with which agents infer beliefs about other agents' beliefs.

2.2 Agent Communication Languages

Agent communication languages have been developed based on the speech act theory [41]. Speech act theory is concerned with the role of language as actions. Among the agent communication languages which emerged from the speech act theory, FIPA-ACL [11] and KQML [10] are the best known.

In this work, for practical reasons, we choose KQML, which is the standard communication language in the Jason platform [5], the multi-agent platform we choose to implement this work. Knowledge Query and Manipulation Language (KQML) was designed to support interaction among intelligent software agents, describing the message format and message-handling protocol to support run-time agent communication [10, 25].

In order to make KQML broadly applicable, a semantic framework for KQML was proposed in [22]. The semantics for KQML-based messages in the AgentSpeak programming language, as given in [43] and implemented in Jason [5], formalises how the locutions successively operate on the states of agents, making the references to the mental attitudes of BDI agents explicit, thus addressing some of the problems of ACLs pointed out in [44]. Based on that semantics, we put forward the idea that agents are able to infer the likely model of other agents' minds during the process of communication, i.e., agents are able to acquire and update ToM, as we describe later in this paper.

2.3 Agent Oriented Programming Languages

Among the many AOPL and platforms, such as Jason, Jadex, Jack, AgentFactory, 2APL, GOAL, Golog, and MetateM, as discussed in [4], we chose the Jason platform [5] for our work. Jason extends the AgentSpeak language, an abstract logic-based AOPL introduced by Rao [37], which is one of the best-known languages inspired by the BDI architecture.

Besides specifying agents with well-defined mental attitudes based on the BDI architecture, the Jason platform [5] has some other features that are particularly interesting for our work, for example strong negation, belief annotations, and (customisable) speech-act based communication. Strong negation helps the modelling of uncertainty, allowing the representation of propositions that the agent: (i) believes to be true, e.g., $\texttt{about}(\texttt{paper1}, \texttt{uncertain\_tom})$; (ii) believes to be false, e.g., $\neg\texttt{about}(\texttt{paper2}, \texttt{uncertain\_tom})$; (iii) is ignorant about, i.e., the agent has no information about whether a paper is about `uncertain_tom` or not. Also, Jason automatically generates annotations for all the beliefs in the agents' belief base about the source from where the belief was obtained (which can be from sensing the environment, communication with other agents, or a mental note created by the agent itself). The annotation has the following format: $\texttt{about}(\texttt{paper1}, \texttt{tom})[\texttt{source}(\texttt{reviewer1})]$, stating that the source of the belief that `paper1` is about the topic `tom` is `reviewer1`. The annotations in Jason can be easily extended to include other meta-information, for example, trust and time as used in [26, 30]. Another interesting feature of Jason is the communication between agents, which is done through a predefined (internal) action. There are a number of performatives allowing rich communication between agents in Jason, as explained in detail in [5]. Further, new performatives can be easily defined (or redefined) in order to give special meaning to them[4], which is an essential characteristic for this work.

3 Running Example

As a running example, we will take the following university scenario with five agents. The first agent, *John*, plays the role of a professor in the university, and the other agents, named *Bob*, *Alice*, *Nick*, and *Ted*, play the role of students. *John* has a relation of *adviser* with the *students*. Also, *John* is responsible for distributing tasks to students, which the students can accept or refuse. *John* keeps information about the students, in order to assign tasks that the students are more likely to accept.

Our model can be formally defined as $\langle Ag, \mathcal{T}, \mathcal{A}, \mathcal{S}\rangle$, in which Ag represents the set of agents, $\mathcal{T}$ the set of tasks of the kind $\mathcal{T} \subseteq \mathcal{A} \times \mathcal{S}$, describing an action from $\mathcal{A}$, requiring knowledge about a subset of subjects from $\mathcal{S}$, that might be executed to achieve the task $\mathcal{T}$. In our example, we consider the following actions, subjects, and tasks:

[4] For example, [31, 32] propose new performatives for argumentation-based communication between Jason agents.

- $\mathcal{A} = \{\texttt{write\_paper}, \texttt{review\_paper}, \texttt{paper\_seminar}\}$
- $\mathcal{S} = \{\texttt{mas}, \texttt{kr}, \texttt{tom}\}$
- $\mathcal{T} = \left\{ \begin{array}{l} \texttt{task}(\texttt{write\_paper}, [\texttt{mas}, \texttt{tom}]) \\ \texttt{task}(\texttt{review\_paper}, [\texttt{kr}]) \\ \texttt{task}(\texttt{paper\_seminar}, [\texttt{tom}, \texttt{mas}]) \end{array} \right\}$

For example, the task to *write a paper with the subjects MAS and ToM*, $\texttt{task}(\texttt{write\_paper}, [\texttt{mas}, \texttt{tom}])$, requires competence on both subjects: `mas` and `tom`. Thus, this task has a greater likelihood to be accepted by a student who desires to execute that particular task, or who likes to execute the action `write_paper` and believes that itself knows the necessary subjects (e.g., $\texttt{knows}(\texttt{mas})$ and $\texttt{knows}(\texttt{tom})$ are necessary to execute this example task). Thus the probability of an agent ag to accept a task t_i is given by the following equation:

$$P(accepts(ag, task_i)) = \begin{cases} P(Des_{ag}(task_i)) & \text{if } Des_{ag}(task_i) \in \Delta_{John} \\ P(Bel_{ag}(likes(a_i))) \times P(Bel_{ag}(knows(S'))) & \text{otherwise} \end{cases}$$

with

$$P(Bel_{ag}(knows(S'))) = \prod_{s_i \in S'} P(Bel_{ag}(knows(s_i)))$$

where $task_i = \texttt{task}(\texttt{a}_\texttt{i}, \texttt{S}')$, for $task_i \in \mathcal{T}$, $a_i \in \mathcal{A}$, and $S' \subseteq \mathcal{S}$. Δ_{John} represents *John*'s knowledge.

Thus, considering our scenario, when *John* knows that some student ag likely desires to execute a particular task $task_i$, i.e., $Des_{ag}(task_i)$, it can use this information to assign the task. Otherwise, *John* can calculate the likely acceptance for each student ag, based on the probability of each student to like executing that action, $P(Bel_{ag}(likes(a_i)))$, and the knowledge the student has about each of the required subjects $P(Bel_{ag}(knows(S')))$. Note that, while modelling the students' desires is more difficult to obtain in our scenario, the students' beliefs are easily obtained by *John*, given that *John* frequently talks to students about these subjects and tasks.

In reality, agents operate with uncertain information, especially in the cases of thinking about other agents' minds. The minds of others are considered to be some sort of black boxes that are more or less accessible depending on the given scenario. Reasoning under uncertainty is a classic case where bounded rationality acts as a major constraint on what agents can infer from their beliefs. However, even if agents are constrained by their access to information, it does not mean that the agents cannot reach reasonable conclusions about the minds of other agents [14, 23].

In our scenario, *John* will reason and make decisions based on information it has about the students' minds, i.e., information from its ToM. Thus *John* will reach conclusions based on uncertain information, given that its ToM contains information about students' minds that has been estimated through the communication *John* has had with the students. Considering that an approach to reason about uncertain information, uncertain ToM in our case, is using probabilistic

reasoning, as described in [14], we have modelled *John*'s decision-making process based on the probability of each information in *John*'s ToM to be correct, considering some factors of uncertainty we will describe further in this paper.

4 Modelling ToM from Other Agents' Actions

In this paper, we are going to consider the modelling of ToM based on communication only, which can be considered a general approach for any application, based on the semantics of each speech act used. On the other hand, the semantics for other actions, e.g., actions agents execute in the environment, might have different meaning according to different application domains.

In order to describe our approach, we use the following notation: $Bel_{ag}(\psi)$ means that an agent ag believes ψ; $Des_{ag}(\psi)$ means that an agent ag desires ψ; Δ_{ag} represents the ag's knowledge base. Two distinct agents are represented using ag_i and ag_j, with $ag_i,\ ag_j \in Ag$, and $ag_i \neq ag_j$. We label the updates agents execute in their ToM with γ, which can be used to represent the uncertainty of that information. In Section 5, we propose an approach for uncertain ToM, which is a particular instance for such γ label.

The speech acts considered in this particular work and their semantics are based on our work in [29]. Messages are represented as $\langle \texttt{sender}, \texttt{receiver}, \texttt{performative}, \texttt{content} \rangle$, and the meaning of each message is associated with the performative used:

- $\langle ag_i, ag_j, \texttt{tell}, \psi \rangle$ means a message sent by agent ag_i to agent ag_j, with the `tell`performative, and content ψ. When ag_i sends this message, it carries out the following update[5] on its ToM:
$$\Delta_{ag_i} = \Delta_{ag_i} \cup Bel_{ag_j}(\psi)_{[\gamma]} \tag{1}$$
When ag_j receives this message, it carries out the following update on its ToM:
$$\Delta_{ag_j} = \Delta_{ag_j} \cup Bel_{ag_i}(\psi)_{[\gamma]} \tag{2}$$
- $\langle ag_i, ag_j, \texttt{ask}, \psi \rangle$ means a message sent by agent ag_i to agent ag_j, with the `ask` performative, and content ψ. When ag_i sends this message, it carries out the following update on its ToM:
$$\Delta_{ag_i} = \Delta_{ag_i} \cup Bel_{ag_j}(Des_{ag_i}(\psi))_{[\gamma]} \tag{3}$$
When ag_j receives this message, it carries out the following update on its ToM:
$$\Delta_{ag_j} = \Delta_{ag_j} \cup Des_{ag_i}(\psi)_{[\gamma]} \tag{4}$$

Before introducing our approach for uncertain ToM, imagine that ToM could be modelled without uncertainty, i.e., that we could ignore γ in our semantic rules. Then, based on these simple semantic rules for agent communication, we are able to show that agents can reach shared beliefs in a relatively straightforward way [29].

[5] Note that we are ignoring any other updates agents execute in their mental attitudes, given we are interested only in the updates agents make on their ToM.

Definition 1 (Shared Beliefs using ToM) *An agent ag_i will reach a state of shared beliefs with another agent ag_j when, for a belief φ, it is able to match its own belief φ with a ToM about ag_j believing φ, i.e., $\varphi \wedge Bel_{ag_j}(\varphi)$.*

Example (shared beliefs without uncertainty): Following the scenario introduced, imagine that two students, *Alice* and *Bob*, need to work together to accomplish a particular task `paper_seminar`, which requires the subjects `mas` (Multi-Agent Systems) and `tom` (Theory of Mind). Also, while *Alice* only knows the subject of `mas`, *Bob* only knows the subject of `tom`. Considering that both *Alice* and *Bob* need to know both topics in order to help each other during the paper seminar, they decide to exchange knowledge about these topics. Thus, they might reach some shared beliefs (knowledge) about both topics. Note that, in this scenario, *Alice* and *Bob* assume that both are cooperating and both are rational. Thus, *Bob* starts the dialogue telling *Alice* that "*Theory of Mind is an approach to model others' minds*", i.e., $\langle alice, \texttt{tell}, \texttt{def}(\texttt{tom}, \text{“an approach to model others' mind”})\rangle$. At that moment, following the semantics for the `tell` performative (equation (1)), *Bob* updates its ToM with the following information $Bel_{alice}(\texttt{def}(\texttt{tom}, \text{“an approach to model others' minds”}))$. After that, when *Alice* receives this message, following the semantics for the `tell` performative (equation (2)), *Alice* updates its belief base with the following information $\texttt{def}(\texttt{tom}, \text{“an approach to model others' mind”})$, as well as *Alice* updates its ToM about *Bob* with $Bel_{bob}(\texttt{def}(\texttt{tom}, \text{“an approach to model other minds”}))$. At this moment, both *Alice* and *Bob* reach a state of shared belief about the definition of `tom`, according to Definition 1.

However, agents operate under conditions of uncertainty in a MAS, and the previous assumptions are hard to obtain; thus, agents will face uncertainty about their ToM, and consequently about their shared beliefs. For example, when an agent sends a message, it faces the uncertainty of the communication channel, i.e., the uncertainty of the message reaching the receiver. Also, when receiving a message, an agent faces the uncertainty of the truth of that statement, e.g., an agent is not able to verify if the other agents are acting maliciously [40, 33], thus it needs to consider the uncertainty of information it receives for those agents based on how much it trusts them [34, 35, 26].

One manner to overcome the uncertainty and reach a more accurate ToM, following the literature on *common knowledge* [6], is increasing the communication between agents. Thus, an agent is able to increase the certainty on a given agent ag_j believing φ, confirming whether its ToM about agent ag_j believing φ is correct. That is, the agent is able to infer that ag_j believes φ by reinforcing this belief through communication. Henceforth we describe our model for uncertain ToM, which is compatible with that behaviour.

5 A Model of Uncertain ToM

In this section we propose an approach to model ToM that reflects the uncertainty present in MAS. In order to show our approach, we are going to consider

some parameter values. The first, α, reflects the uncertainty of the communication channel when sending a message. The second, β, reflects the uncertainty of the other agents telling the truth, i.e., when an agent ag_i tells φ to agent ag_j, agent ag_j is able to model that ag_i believes on φ with a degree of certainty equal to β. For simplicity, we will assume that an agent will model its ToM about the other agents with a degree of certainty equal to the trust it has on the source[6], following the ideas introduced in [35, 34].

Definition 2 *The label γ will be instantiated with $\gamma = (\alpha, t)$ for an agent sending a message, and $\gamma = (\beta, t)$ for an agent receiving a message, where α represents the uncertainty of the message reaching the target, β the uncertainty of the sender telling the truth, and t a discrete representation of the time of the MAS in which the message was exchanged.*

Thus, following Definition 2, a trace of different updates on the ToM is constructed over time. Note that α and β reflect the uncertainty of an update at a given time. In order to execute reasoning over the ToM, agents are able to use the trace of these updates to calculate the degree of certainty on their model. Using this trace, we are able to model some desired behaviour from communication theory in agent communication, as we will describe later in this paper.

For example, considering our scenario, when *Bob* tells *Alice* that "Theory of Mind is an approach to model others' minds", considering also that *Bob* knows that the efficiency of the communication channel is 0.9, i.e., $\alpha = 0.9$, *Bob* will update its ToM, following the semantics for the `tell` performative (equation (1)) and Definition 2, with the information $Bel_{alice}(\texttt{def(tom}, \text{"an approach to model others' minds"}\texttt{)})_{[(0.9,t_i)]}$, with t_i the discrete time when the communication occurred. When *Alice* receives this message, considering that the trust *Alice* has on *Bob* telling the truth is 0.8, i.e., $\beta = 0.8$, and following the semantics for the `tell` performative (equation (2)) and Definition 2, *Alice* updates its ToM with $Bel_{bob}(\texttt{def(tom}, \text{"an approach to model other minds"}\texttt{)})_{[(0.8,t_j)]}$, with t_j the discrete time at which the message was received, with $t_i < t_j$. Both *Alice* and *Bob* model uncertainty of their ToM about each other believing on the definition of ToM.

Considering uncertain ToM, we need to redefine shared beliefs, in order to reflect the uncertainty of agents' models.

Definition 3 (Shared Beliefs using Uncertain ToM) *An agent ag_i will reach a state of shared beliefs with another agent ag_j when, for a belief φ, it is able to match its own belief φ with a ToM about ag_j believing φ with a predetermined degree of certainty χ, i.e., $\varphi \wedge P(Bel_{ag_j}(\varphi)) \geq \chi$, with χ a value describing the certainty necessary to consider φ a shared belief.*

Following the literature on common knowledge [6], if two individuals ag_i and ag_j can repeatedly communicate, then they can repeatedly reinforce their mental

[6] In [28], the authors show that trust aggregates not only the sincerity of the source but also the expertise the source has about the information communicated.

state regarding an information φ. For example, telling each other that φ is true, they should increase the certainty of each others' belief in φ. In order to model this desired behaviour in our model, we maintain the trace of all updates an agent executes in its ToM, and using this trace we are able to aggregate different pieces of evidence in order to increase the certainty on ToM. There are many different ways to model this desired behaviour on agent communication, and it could consider the particularities of each application domain. In our scenario, the information communicated by agents, e.g. a concept definition, does not change over time. Thus, for simplicity, we do not weight each information according to the time it was received and the current time of the MAS, we only consider the number of evidences about that information. Thus, we model this desired behaviour using the following equation:

$$P(Bel_{Ag}(\varphi)) = \begin{cases} f(Bel_{ag}(\varphi)) & \text{if } f(Bel_{ag}(\varphi)) <= 1 \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

$$f(Bel_{ag}(\varphi)) = \frac{\sum\limits_{t_i \in \Delta T} v \mid Bel_{ag}(\varphi)_{[(v,t_i)]}}{|\Delta T|} + (\lambda \times |\Delta T|) \tag{6}$$

with ΔT the number of occurrences of $Bel_{ag}(\varphi)_{[(v,t_i)]}$ in the agent ToM, and λ the *evidence factor*, i.e., a parameter that reinforce the certainty on that information according to how often it occurs in the trace. Equation 6 calculates the average of the trace for $Bel_{ag}(\varphi)$ plus the evidence factor. Thus, following Definition 3, ag_i is able to reach a state of shared belief with another agent ag_j about a belief φ when it is able to infer $P(Bel_{ag_j}(\varphi)) \geq \chi$ from Equation 5 with $\chi = 1$, for example.

Proposition 1 (Reaching Shared Beliefs — ToM with Uncertainty) *When λ is a positive value, agents are able to eventually reach a state of shared beliefs, even considering $\chi = 1$, provided they communicate the same information repeatedly. Also, the greater the value of λ, the faster agents will reach the state of shared beliefs.*

Example (shared beliefs under conditions of uncertainty): Following our example, imagine that *Bob* wants to reach a state of shared beliefs with *Alice* about the definition of ToM under the conditions of uncertainty described above. Thus, after sending the previous message and updating its ToM with $Bel_{alice}(\texttt{def(tom}, \text{“an approach to model others' minds”}\texttt{)})_{[(0.9,t_i)]}$, *Bob* has two options to increase the certainty on its ToM about *Alice* believing on that definition: (i) telling *Alice* that definition more times, or (ii) asking *Alice* the definition of ToM and waiting for an answer from *Alice*, in which *Alice* tells *Bob* the definition of ToM. Considering $\lambda = 0.1$, in the first case, when *Bob* tells *Alice* about the definition of ToM one more time, following the semantics for the `tell` performative (equation (1)) and Definition 2, *Bob* adds $Bel_{alice}(\texttt{def(tom}, \text{“an approach to model others' minds”}\texttt{)})_{[(0.9,t_j)]}$ to its ToM, with $t_i < t_j$. Thus, Equation 5 returns 1, considering the average 0.9 + 0.2 from the evidence factor, which is 0.1 multiplied by the

number of evidences (equation (6)). Also, following the semantics for the `tell` performative (equation (2)) and Definition 2, *Alice* updates its ToM with $Bel_{bob}(\texttt{def(tom, "an approach to model other minds")})_{[(0.8,t_j)]}$, and Equation (5) returns 1, considering the average $0.8 + 0.2$ from the evidence factor (equation (6)). Thus, they reach a state of shared belief about the definition of ToM[7], considering $\chi = 1$ in Definition 3. In the other case, *Bob* asks to *Alice* to tell him the definition of ToM, and it waits for the answer. When *Alice* tells *Bob* the definition of ToM, *Alice* and *Bob* update their ToM with $Bel_{bob}(\texttt{def(tom, "an approach to model other minds")})_{[(0.9,t_j)]}$, $Bel_{alice}(\texttt{def(tom, "an approach to model others' minds")})_{[(0.8,t_i)]}$, respectively. For both, Equation (5) returns 1, considering the average $0.85 + 0.2$ from the evidence factor, reaching a state of shared beliefs about the definition of ToM according to Definition 3 with $\chi = 1$.

6 Decision Making Using Uncertain ToM

Apart from enabling agents to model other agents' minds and allowing them to improve their models during communicative interactions, it is also essential that agents are able to make decisions using these models. Normally, a decision-making process is associated with the application domain, i.e., it is domain dependent. Therefore, we will present the decision-making process for the task assignment problem introduced in Section 3.

In our scenario, during advising sessions, *John* asks students about different tasks they like to execute, as well as the different subjects the students are reading about (the subjects the students know about). Thus, *John* acquires ToM about the students, and its ToM becomes more accurate as they have more advising sessions, and consequently they communicate more with each other.

$$John_{_{ToM}} = \left\{ \begin{array}{l} Bel_{alice}(\texttt{likes(paper\_seminar)})_{[0.8]} \\ Bel_{alice}(\texttt{likes(write\_paper)})_{[0.7]} \\ Bel_{bob}(\texttt{likes(review\_paper)})_{[0.9]} \\ Bel_{bob}(\texttt{likes(write\_paper)})_{[0.8]} \\ Bel_{nick}(\texttt{likes(review\_paper)})_{[0.6]} \\ Bel_{nick}(\texttt{likes(write\_paper)})_{[0.5]} \\ Bel_{ted}(\texttt{likes(write\_paper)})_{[0.8]} \\ Bel_{ted}(\texttt{likes(paper\_seminar)})_{[0.4]} \\ Bel_{ted}(\texttt{likes(review\_paper)})_{[0.6]} \end{array} \right\}$$

For example, *John* has asked (in different meetings and times) *Bob*, *Alice*, *Nick*, and *Ted* which academic tasks they like to execute, e.g., $\langle bob, AskIf, \texttt{likes(T)} \rangle$. After receiving this message, according to the semantic rule for the `ask` performative (equation (4)), each student knows that *John* desires to know which task they like to execute. Based on this knowledge, each student has answered

[7] When considering $\gamma = 0.1$ and α and $\beta >= 0.8$, agents are able to reach shared beliefs communicating only 2 messages.

to *John* the tasks they like to execute, *John* has received these messages and updated its ToM as shown in $John_{_{ToM}}$[8].

Continuing with the example, during a meeting *Alice* asks *John* if there is any scheduled paper seminar about ToM and MAS, i.e., $\langle \texttt{john}, AskIf, \texttt{task}(\texttt{paper\_seminar}, [\texttt{tom}, \texttt{mas}]\rangle$. Thus, based on the semantic rule for the `ask` performative (equation (4)), *John* models that *Alice* is likely to desire that task, i.e., $Des_{alice}(\texttt{task}(\texttt{paper\_seminar}, [\texttt{tom}, \texttt{mas}]))_{[0.7]}$, answering positively. Also, imagine that *John* has asked the students which subject they have knowledge about, resulting in the following additional information to *John*'s ToM:

$$John_{_{ToM}} = \left\{ \begin{array}{ll} Bel_{alice}(\texttt{knows}(\texttt{tom}))_{[0.8]} & Bel_{bob}(\texttt{knows}(\texttt{mas}))_{[0.8]} \\ Bel_{alice}(\texttt{knows}(\texttt{mas}))_{[0.9]} & Bel_{bob}(\texttt{knows}(\texttt{kr}))_{[0.9]} \\ Bel_{nick}(\texttt{knows}(\texttt{kr}))_{[0.8]} & Bel_{ted}(\texttt{knows}(\texttt{tom}))_{[0.8]} \\ Bel_{nick}(\texttt{knows}(\texttt{mas}))_{[0.7]} & Bel_{ted}(\texttt{knows}(\texttt{kr}))_{[0.5]} \\ Bel_{nick}(\texttt{knows}(\texttt{tom}))_{[0.8]} & Bel_{ted}(\texttt{knows}(\texttt{mas}))_{[0.8]} \end{array} \right\}$$

Using its ToM, *John* executes the probabilistic reasoning described in Section 3, which computes the likelihood for each student to accept each task as shown in Table 1. Note that the likelihood of *Alice* accepting the task `paper_seminar` is based on the information $Des_{alice}(\texttt{task}(\texttt{paper\_seminar}, [\texttt{tom}, \texttt{mas}]))_{[0.7]}$ in *John*'s ToM, while the other results are based on the likelihood of the students liking a particular task and knowing the subjects related to that task. Thus, *John* concludes that it is possible to increase the probability of each task to be accepted by the students by offering the task $\texttt{task}(\texttt{paper\_seminar}, [\texttt{tom}, \texttt{mas}])$ to *Alice*, offering $\texttt{task}(\texttt{review\_paper}, [\texttt{kr}])$ to *Bob*, and offering $\texttt{task}(\texttt{write\_paper}, [\texttt{mas}, \texttt{tom}])$ to *Ted*.

Student	Task	Likelihood
Alice	$\texttt{task}(\texttt{write\_paper}, [\texttt{mas}, \texttt{tom}])$	0.5
Alice	$\texttt{task}(\texttt{review\_paper}, [\texttt{kr}])$	0.0
Alice	$\texttt{task}(\texttt{paper\_seminar}, [\texttt{tom}, \texttt{mas}])$	**0.7**
Bob	$\texttt{task}(\texttt{write\_paper}, [\texttt{mas}, \texttt{tom}])$	0.0
Bob	$\texttt{task}(\texttt{review\_paper}, [\texttt{kr}])$	**0.8**
Bob	$\texttt{task}(\texttt{paper\_seminar}, [\texttt{tom}, \texttt{mas}])$	0.0
Nick	$\texttt{task}(\texttt{write\_paper}, [\texttt{mas}, \texttt{tom}])$	0.3
Nick	$\texttt{task}(\texttt{review\_paper}, [\texttt{kr}])$	**0.5**
Nick	$\texttt{task}(\texttt{paper\_seminar}, [\texttt{tom}, \texttt{mas}])$	0.0
Ted	$\texttt{task}(\texttt{write\_paper}, [\texttt{mas}, \texttt{tom}])$	**0.5**
Ted	$\texttt{task}(\texttt{review\_paper}, [\texttt{kr}])$	0.2
Ted	$\texttt{task}(\texttt{paper\_seminar}, [\texttt{tom}, \texttt{mas}])$	0.1

Table 1. Likelihood calculation for task assignment

[8] We do not represent the time at which the messages were communicated, but since they were communicated at different times we introduced different values for γ.

7 Future Work

Uncertainty does not only arise from noisy communication channels or levels of trust between agents. As future work, we plan to add an environment to our model in order to represent how agents infer ToM from the observation of actions performed by other agents in that environment. The modelling of ToM based on these aspects faces complex issues such as the ones mentioned in [7]: "*the slamming of a door communicates the slammer's anger only when the intended observer of that act realises that the slammer wanted both to slam the door in his face and for the observer to believe that to be his intention*". This means that there is both uncertainty about the slammer's intentions and uncertainty about the act of slamming the door, which could be caused by an accidental shove or by natural means, hence not represent a communicative act. Therefore, observing such an event occur should not cause the observer to make any inference about the slammer's mental state. That being said, modelling ToM based on environment observations requires more than only representing both intended and non-intended acts of communication. The slammer might very well not intend to communicate when slamming the door, but that does not stop the observer from reading a message when observing the slamming of the door. These complex issues arise with the inclusion of ToM because agents are able to project beliefs in the minds of other agents they share an environment, or even just a communication channel, with. Therefore, the agents that project beliefs can be subject to what is known as the *Mind Projection Fallacy* [18]. An agent commits this fallacy using ToM when the agent incorrectly assigns beliefs to another agent's mind[9]. In our future work, we hope to improve our model in order to be able to represent complex phenomena such as the mind projection fallacy.

8 Conclusions

We have proposed an approach for agents to acquire and update their ToM during communication whilst reflecting on the uncertainty of this process. To the best of our knowledge, our work is the first to explicitly address acquisition and update of uncertain ToM in MAS. In order to show how our approach allows us to model desired properties from communication and common knowledge theories [6], we have proposed a model for uncertain ToM. Using our approach, agents are able to reach accurate ToM and, consequently, accurate shared beliefs by reinforcing their mental attitudes through communication. Furthermore, in this work, we have shown not only how agents acquire and update ToM based on agent communication, but also how agents reason and make decisions using ToM.

The modelling of ToM in MAS under the condition of uncertainty is an important step towards obtaining more realistic and more socially aware artificial agents. We argue that the approach we used to model ToM in MAS is in tune

[9] It is similar to committing a type I error in a statistical analysis.

with what we believe to be an upcoming discipline in the field of AI, namely the study of machine behaviour as proposed by [36]. Thus, modelling ToM is relevant to both the AI community and to multi-disciplinary research groups because it offers the possibility to study how agents reach agreements with [24], cooperate with [39], or even behave dishonestly towards [40, 33] other agents using more realistic models of social interactions.

Acknowledgements

We gratefully acknowledge the partial support from CAPES and CNPq. Special thanks to Francesca Mosca for the support and for the feedback on this paper.

References

1. Apperly, I.A.: What is theory of mind? concepts, cognitive processes and individual differences. The Quarterly Journal of Experimental Psychology **65**(5), 825–839 (2012)
2. Barlassina, L., Gordon, R.M.: Folk psychology as mental simulation. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, summer 2017 edn. (2017)
3. Black, E., Atkinson, K.: Choosing persuasive arguments for action. In: The 10th International Conference on Autonomous Agents and Multiagent Systems. pp. 905–912. (2011)
4. Bordini, R.H., Dastani, M., Dix, J., Seghrouchni, A.E.F.: Multi-Agent Programming: Languages, Tools and Applications. Springer Publishing Company, Incorporated, 1st edn. (2009)
5. Bordini, R.H., Hübner, J.F., Wooldridge, M.: Programming Multi-Agent Systems in AgentSpeak using Jason (Wiley Series in Agent Technology). John Wiley & Sons (2007)
6. Chwe, M.S.Y.: Rational ritual. Culture, coordination, and common knowledge (2001)
7. Cohen, P.R., Perrault, C.R.: Elements of a plan-based theory of speech acts. In: Readings in Distributed Artificial Intelligence, pp. 169–186. Elsevier (1988)
8. de Weerd, H., Verheij, B.: The advantage of higher-order theory of mind in the game of limited bidding. In: Proc. Workshop Reason. About Other Minds, ceur workshop proceedings. vol. 751, pp. 149–164 (2011)
9. de Weerd, H., Verbrugge, R., Verheij, B.: Higher-order social cognition in rock-paper-scissors: A simulation study. In: Proc. of the 11th International Conference on Autonomous Agents and Multiagent Systems. pp. 1195-1196 (2012)
10. Finin, T., Fritzson, R., McKay, D., McEntire, R.: KQML as an agent communication language. In: Proc. of the 3rd international conference on Information and knowledge management. pp. 456–463. ACM (1994)
11. FIPA, T.: FIPA communicative act library specification. Foundation for Intelligent Physical Agents, http://www.fipa.org/specs/fipa00037/SC00037J.html (15.02.2018) (2008)
12. Goldman, A.I.: Theory of mind. In: The Oxford Handbook of Philosophy of Cognitive Science, vol. 1. Oxford Handbooks Online, 2012 edn. (2012)

13. Gopnik, A., Glymour, C., Sobel, D.M., Schulz, L.E., Kushnir, T., Danks, D.: A theory of causal learning in children: causal maps and bayes nets. Psychological review **111**(1), 3 (2004)
14. Gopnik, A., Wellman, H.M.: Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. Psychological bulletin **138**(6), 1085 (2012)
15. Hadidi, N., Dimopoulos, Y., Moraitis, P., et al.: Tactics and concessions for argumentation-based negotiation. In: COMMA. pp. 285–296 (2012)
16. Hadjinikolis, C., Siantos, Y., Modgil, S., Black, E., McBurney, P.: Opponent modelling in persuasion dialogues. In: International Joint Conference on Artificial Intelligence. pp. 164–170 (2013)
17. Hyslop, A.: Other minds. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, spring 2016 edn. (2016)
18. Jaynes, E.T.: Probability theory as logic. In: Maximum entropy and Bayesian methods, pp. 1–16. Springer (1990)
19. Kumar, S., Cohen, P.R.: STAPLE: An agent programming language based on the joint intention theory. In: Proc. of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems. pp. 1390–1391 (2004)
20. Kumar, S., Cohen, P.R., Huber, M.J.: Direct execution of team specifications in STAPLE. In: Proc. of the 1st International Joint Conference on Autonomous Agents and Multiagent Systems. pp. 567–568 (2002)
21. Kumar, S., Cohen, P.R., Levesque, H.J.: The adaptive agent architecture: Achieving fault-tolerance using persistent broker teams. In: Proc Fourth International Conference on MultiAgent Systems, 2000. pp. 159–166 (2000)
22. Labrou, Y., Finin, T.: A semantics approach for KQML - a general purpose communication language for software agents. In: Proceedings of the third international conference on Information and knowledge management. pp. 447–455. ACM (1994)
23. Leudar, I., Costall, A.: On the persistence of the problem of other minds in psychology: Chomsky, grice and theory of mind. Theory & Psychology **14**(5), 601–621 (2004)
24. Luck, M., McBurney, P.: Computing as interaction: agent and agreement technologies. In: IEEE international conference on distributed human-machine systems. IEEE Press. Citeseer (2008)
25. Mayfield, J., Labrou, Y., Finin, T.W.: Evaluation of KQML as an agent communication language. In: Wooldridge, M., Mller, J.P., Tambe, M. (eds.) ATAL. Lecture Notes in Computer Science, vol. 1037, pp. 347–360. Springer (1995)
26. Melo, V.S., Panisson, A.R., Bordini, R.H.: Argumentation-based reasoning using preferences over sources of information. In: 15th International Conference on Autonomous Agents and Multiagent Systems. (2016)
27. Oren, N., Norman, T.J.: Arguing using opponent models. In: International Workshop on Argumentation in Multi-Agent Systems. pp. 160–174 (2009)
28. Paglieri, F., Castelfranchi, C., da Costa Pereira, C., Falcone, R., Tettamanzi, A., Villata, S.: Trusting the messenger because of the message: feedback dynamics from information quality to source evaluation. Computational and Mathematical Organization Theory **20**(2), 176–194 (2014)
29. Panisson, A.R., Sarkadi, S., McBurney, P., Parsons, S., Bordini, R.H.: On the Formal Semantics of Theory of Mind in Agent Communication. In: 6th International Conference on Agreement Technologies. (2018)
30. Panisson, A.R., Melo, V.S., Bordini, R.H.: Using preferences over sources of information in argumentation-based reasoning. In: 5th Brazilian Conference on Intelligent Systems. pp. 31–26 (2016)

31. Panisson, A.R., Meneguzzi, F., Fagundes, M., Vieira, R., Bordini, R.H.: Formal semantics of speech acts for argumentative dialogues. In: 13th International Conference on Autonomous Agents and Multiagent Systems. pp. 1437–1438 (2014)
32. Panisson, A.R., Meneguzzi, F., Vieira, R., Bordini, R.H.: Towards practical argumentation in multi-agent systems. In: Brazilian Conference on Intelligent Systems. (2015)
33. Panisson, A.R., Sarkadi, S., McBurney, P., Parsons, S., Bordini, R.H.: Lies, bullshit, and deception in agent-oriented programming languages. In: Proc. of the 20th International Trust Workshop pp. 50–61 (2018)
34. Parsons, S., Sklar, E., McBurney, P.: Using argumentation to reason with and about trust. In: Argumentation in multi-agent systems, pp. 194–212 (2012)
35. Parsons, S., Tang, Y., Sklar, E., McBurney, P., Cai, K.: Argumentation-based reasoning in agents with varying degrees of trust. In: The 10th International Conference on Autonomous Agents and Multiagent Systems. pp. 879–886 (2011)
36. Rahwan, I., Cebrian, M.: Machine behavior needs to be an academic discipline (2018), http://nautil.us/issue/58/self/machine-behavior-needs-to-be-an-academic-discipline
37. Rao, A.S.: AgentSpeak(L): BDI agents speak out in a logical computable language. In: Proceedings of the 7th European Workshop on Modelling Autonomous Agents in a Multi-Agent World : agents breaking away: agents breaking away. pp. 42–55. MAAMAW '96, Springer-Verlag New York, Inc., Secaucus, NJ, USA (1996)
38. Rienstra, T., Thimm, M., Oren, N.: Opponent models with uncertainty for strategic argumentation. In: International Joint Conference on Artificial Intelligence. pp. 332–338 (2013)
39. Rosenschein, J.S.: Rational interaction: cooperation among intelligent agents (1986)
40. Sarkadi, S.: Deception. In: IJCAI. pp. 5781–5782 (2018)
41. Searle, J.R.: Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press (1969)
42. Thimm, M.: Strategic argumentation in multi-agent systems. KI-Künstliche Intelligenz **28**(3), 159–168 (2014)
43. Vieira, R., Moreira, A., Wooldridge, M., Bordini, R.H.: On the formal semantics of speech-act based communication in an agent-oriented programming language. J. Artif. Int. Res. **29**(1), 221–267 (Jun 2007)
44. Wooldridge, M.: Semantic issues in the verification of agent communication languages. Autonomous agents and multi-agent systems **3**(1), 9–31 (2000)
45. Wooldridge, M.: An introduction to multiagent systems. John Wiley & Sons (2009)