Distributed Theory of Mind in Multi-Agent Systems

Heitor Henrique da Silva[1], Michele Rocha[1], Guilherme Trajano[1], Analúcia Schiaffino Morales[1],
Stefan Sarkadi[2] and Alison R. Panisson[1]

[1] *Department of Computing, Federal University of Santa Catarina, Santa Catarina, Brazil*

[2] *Department of Informatics, King's College London, London, U.K.*

Keywords: Theory of Mind, Multi-Agent Systems, Distributed Theory of Mind.

Abstract: Theory of Mind is a concept from developmental psychology which elucidates how humans mentalise. More specifically, it describes how humans ascribe mental attitudes to others and how they reason about these mental attitudes. In the area of Artificial Intelligence, Theory of Mind serves as a fundamental pillar in the design of intelligent artificial agents that are supposed to coexist with humans within a hybrid society. Having the ability to mentalise, these artificial agents could potentially exhibit a range of advanced capabilities that underlie meaningful communication, including empathy and the capacity to better understanding the meaning behind the utterances others make. In this paper, we propose a distributed theory of mind approach in multi-agent systems, in which agents and human users share evidence to reach more supported conclusions about each other's mental attitudes. We demonstrate our approach in a scenario of stress detection, in which personal agents infer whether their users are stressed or not according to the distributed theory of mind approach.

1 INTRODUCTION

Theory of Mind (ToM) plays a pivotal role in the field of Artificial Intelligence as it bridges crucial gaps between our understanding of human cognition and the development of intelligent autonomous systems. At its core, ToM refers to the capacity to comprehend and model the mental states of others, enabling us to attribute, for example, beliefs, desires, intentions, and emotions to them. This fundamental cognitive ability has far-reaching implications across various scientific disciplines, grounding phenomena related to empathy, social interaction, and communication skills between individuals. ToM has predominantly been studied in humans, particularly in the context of cognitive development (Goldman et al., 2012).

Particularly, in Multi-Agent Systems (MAS), ToM plays a fundamental role in enhancing communication, fostering collaboration, detecting deceptive behaviour, and facilitating sophisticated human-agent interaction. These phenomena demand sophisticated reasoning mechanisms using the mental attitudes of others. ToM is recognised as an indispensable component in systems within the context of Hybrid Intelligence (HI) and eXplainable AI (XAI) (Akata et al., 2020).

Our work introduces an innovative approach to distributed ToM, empowering agents not only to model the mental attitudes of fellow agents, whether they are software agents or human users, but also to model ToM of those agents. Our approach incorporates a mechanism enabling agents to associate uncertainty with the mental models within their ToM. Furthermore, it enables the sharing of ToM and the aggregation of diverse models to arrive at more robust insights into the mental attitudes of other agents. In the context of HI, we have developed a natural language interface capable of inferring both users' mental attitudes and users' ToM. Our implementation is based on the JaCaMo Framework (Boissier et al., 2013), and we provide a case study in the domain of stress detection.

2 DISTRIBUTED ToM IN MAS

In this paper, we explore the distributed nature of ToM in multi-agent systems, considering not only the phenomena in which an agent is able to ascribe mental attitudes to other agents (software agents and human users) within the system but also the phenomena in which these models can be communicated by agents. This approach aims to achieve a more robust individual ToM, supported by multiple lines of evidence.

da Silva, H., Rocha, M., Trajano, G., Morales, A., Sarkadi, S. and Panisson, A.
Distributed Theory of Mind in Multi-Agent Systems.
DOI: 10.5220/0012563400003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 1*, pages 451-460
ISBN: 978-989-758-680-4; ISSN: 2184-433X

To instantiate the ToM model proposed in this work, we will utilise first-order predicates, similar to those employed in knowledge representation in Agent-Oriented Programming Language (AOPL). For instance, we will use `stressed(pietro)` to denote that ‘pietro is stressed,’ and similarly, we will use `likes(pietro,ice_cream)` to indicate that ‘pietro likes ice cream.’

Furthermore, in this paper, we will employ the following notation to represent ToM, including a degree of certainty, drawing inspiration from (Panisson et al., 2018; Sarkadi et al., 2019):

- $Bel_{ag}(\varphi)_{[\gamma]}$ means an agent *ag* believes an information φ, with a degree of certainty denoted by γ. For example, $Bel_{alice}(\texttt{stressed(pietro)})_{[0.8]}$ means that *alice* believes *pietro* is stressed with a degree of certainty of 0.8.
- $Des_{ag}(\varphi)_{[\gamma]}$ means an agent *ag* desires φ with a degree of certainty denoted by γ. For example, $Des_{alice}(\texttt{take\_day\_off(pietro)})_{[0.6]}$ means that *alice* desires *pietro* to take the day off with a degree of certainty of 0.6.

When employing a higher order of ToM, φ will be instantiated with one of the previously modelled mental attitudes from the ToM. For example, $Bel_{ag_i}(Bel_{ag_j}(\varphi))_{[\gamma]}$ represents that an agent ag_i believes that another agent ag_j believes in information φ, with γ indicating the degree of certainty regarding this information[1].

Communication between agents is grounded in the speech act theory (Austin, 1975). In the context of Agent Communication Languages (ACL), a message, in its basic form, consists of a pair that includes a performative and content (Mayfield et al., 1995; FIPA, 2008), in which the performative provides the sender’s intention in the communication. In this paper, we adopt the following notation to represent communication: $\langle ag_i, ag_j, pfm, \varphi \rangle$, where ag_i represents the sender, and ag_j represents the message’s target (receiver), *pfm* indicates the chosen performative, and φ the content of the message.

2.1 Ascribing ToM

Communication is a natural method for acquiring and updating a ToM regarding other individuals. This principle has also been applied to agent (software) communication (Panisson et al., 2019). It is a fundamental principle. When someone provides us with some piece of information, such as they liking for ice cream, we can construct a ToM in which that piece of information is present, keeping in mind that the person believes they like ice cream. However, it is worth noting that individuals may not always be completely honest in their speech, so there is a degree of certainty associated with modelling mental attitudes from communication. This degree of certainty can be inferred from various contexts, including the level of trust we place in the person or their expertise in the subject they are discussing. Some valuable insights in this area can be found in the literature on argumentation-based reasoning, as explored in works by (Parsons et al., 2012; Walton et al., 2008; Melo et al., 2016; Melo et al., 2017). This is why we incorporate γ into our ToM model, representing a degree of certainty an agent assigns to the information it infers about others’ mental attitudes.

For example, when an agent named ag_i receives a message $\langle ag_j, ag_i, tell, \varphi \rangle$, i.e., a message from agent ag_j, with the $tell$ performative and the content φ, ag_i is able to infer that ag_j believes on what it is telling, adding $Bel_{ag_j}(\varphi)_{[\gamma]}$ to its ToM model. We use the following semantics for agent communication:

- $\langle ag_i, ag_j, tell, \varphi \rangle \models Bel_{ag_i}(\varphi)_{[\gamma]}$, meaning that when an agent ag_j receives a message from another agent ag_i, with the performormative $tell$ and a content φ, it will model that ag_i believes on φ, i.e., $Bel_{ag_i}(\varphi)_{[\gamma]}$.
- $\langle ag_i, ag_j, achieve, \varphi \rangle \models Des_{ag_i}(\varphi)_{[\gamma]}$, meaning that when an agent ag_j receives a message from another agent ag_i, with the performormative *achieve* and a content φ, it will model that ag_i desires φ, i.e., $Des_{ag_i}(\varphi)_{[\gamma]}$.

When considering human-computer (agent) communication, γ may also be associated with the certainty of the natural language interface correctly identifying what the human user has communicated or even combining various uncertainties related to that information. For example, this might involve assessing how much trust we place in the person communicating that piece of information and the uncertainty associated with the natural language interface’s comprehension of that person’s statements.

At some point, an agent may possess in its ToM a sequence or trace of inferred mental models about other agents. For instance, it may model that agent *ag* believes in φ, denoted as $Bel_{ag}(\varphi)_{[\gamma]}$ but at different moments and with different degrees of certainty. This consideration takes into account the different instants it has interacted with that agent and when it has modelled their mental attitudes through interactions. These time-related (meta)information are represented here by the timestamps $t_1, \ldots, t_n$ annotated in the pieces of information aggregated by the generic operator $\oplus$, as defined in equation (1).

[1] We consider that γ incorporates the degree of certainty from the nested predicates, as we will discuss later.

$$Bel_{ag}(\varphi)_{[\gamma]} = Bel_{ag}(\varphi)_{[\gamma_1,t_1]} \oplus \ldots \oplus Bel_{ag}(\varphi)_{[\gamma_n,t_n]} \quad (1)$$

There are numerous ways to implement the operator $\oplus$, taking into account not only the application domain but also various agent profiles. These profiles may also reflect domain-specific requirements. For instance, there could be agents that only consider the most recent information, or those that take into account the entire history or 'trace' of information.

In this paper, as we explore different possible instantiations of $\oplus$, we will demonstrate two of the most straightforward choices in our examples.

Definition 2.1 (Time-Concerned Agent)**.** A time-concerned agent prioritises the most up-to-date information extracted from the traces of a particular mental attitude from its ToM, ignoring multiples occurrences of that information, as defined in Equation (2):

$$Bel_{ag}(\varphi)_{[\gamma]} = \underset{t}{\operatorname{argmax}}\ Bel_{ag}(\varphi)_{[\gamma,t]} \quad (2)$$

in which the agent will consider the most recently updated model, specifically, the last information it has modelled.

Definition 2.2 (Trace-Concerned Agent)**.** A trace-concerned agent prioritises the higher degree of certainty of a mental attitude extracted from the trace from its ToM, as defined in equation (3):

$$Bel_{ag}(\varphi)_{[\gamma]} = \underset{\gamma}{\operatorname{argmax}}\ Bel_{ag}(\varphi)_{[\gamma,t]} \quad (3)$$

in which the agent will consider the mental attitude with higher degree of certainty in the trace.

After understanding a particular modelled mental attitude and calculating a degree of certainty associated to it, agents can combine multiple modelled mental attitudes that support the same piece of information. In other words, agents can integrate different theories of mind by checking if a piece of information is consistent with the majority of the agents they have interacted with. For example, an agent ag updates its own ToM about a piece of information, denoted as $Bel_{ag}(\varphi)_{[\gamma]}$, based on other theories of mind modelled from different agents. These include the set of modelled attitudes $Bel_{ag_1}(\varphi)_{[\gamma_1]}, \ldots, Bel_{ag_n}(\varphi)_{[\gamma_n]}$ from agents $ag_1, \ldots, ag_n$. Equation 4 defines the general operator $\otimes$ that can be instantiated according to the specific interests of the application domain or when defining different agent profiles.

$$Bel_{ag}(\varphi)_{[\gamma]} = Bel_{ag_1}(\varphi)_{[\gamma_1]} \otimes \ldots \otimes Bel_{ag_n}(\varphi)_{[\gamma_n]} \quad (4)$$

In this paper, we will consider the instantiation provided in Equation (5), which aims to penalize the degree of certainty when conflicting models are found but rewards it when no conflicting model is present, as follow:

$$Bel_{ag}(\varphi)_{[\gamma]} \mid \gamma = \delta + (1-\delta) \cdot \kappa \cdot \frac{|S^+_\varphi| - |S^-_\varphi|}{|S^+_\varphi| + |S^-_\varphi|} \quad (5)$$

with

$$\delta = \frac{\sum_{s \in S^+_\varphi} \gamma \mid Bel_s(\varphi)_{[\gamma]}}{|S^+_\varphi| + |S^-_\varphi|}$$

in which $S^+_\varphi = \{s_1, ..., s_n\}$ is the set of n different agents that believe φ and S^-_φ is the set different agents that believe $\overline{\varphi}$ (i.e., its complement). In Equation (5), $\kappa \cdot \frac{|S^+_\varphi|-|S^-_\varphi|}{|S^+_\varphi|+|S^-_\varphi|}$ rewards (increases) the certainty regarding a particular piece of information when there is more evidences in $|S^+_\varphi|$. A larger value of κ will make the degree of certainty more sensitive to the difference between positive and negative occurrences, while a smaller value of κ will make it less sensitive.

In Figure 1, we can observe the behaviour of Equation (5). We have fixed $\gamma = 0.8$ for all evidences. On the x-axis (horizontal), we show the range of evidences for $|S^+_\varphi|$, varying from 0 to 10, and simultaneously, the range of evidence in $|S^-_\varphi|$, which varies from 0 to 2. In other words, when there is 5 pieces of evidence in $|S^+_\varphi|$, there is 1 piece of evidence in $|S^-_\varphi|$, and when there is 10 pieces of evidence in $|S^+_\varphi|$, there are 2 evidence at $|S^-_\varphi|$. On the y-axis (depth), we display the range of κ values, ranging from 0.1 to 0.9. The z-axis (vertical) shows the final degree of certainty. We can observe that the degree of certainty decreases rapidly with contrary evidences. However, it is possible to compensate for this effect by using a higher value for κ, making it more sensitive to the difference between positive and negative evidences.

2.2 Distributing ToM

In this section, we introduce an approach that agents can employ to share mental attitude from their theories of mind, facilitating the creation of a distributed ToM. To enable agents to share their ToM, we have introduced a new performative named *share_tom*, and its semantics is provided below:

- $\langle ag_i, ag_k, share\_tom, Bel_{ag_j}(\varphi)_{[\gamma]}\rangle \models Bel_{ag_j}(\varphi)_{[\gamma]}$, meaning that when an agent ag_k receives message from another agent ag_i, with the performative *share_tom* and the content $Bel_{ag_j}(\varphi)_{[\gamma]}$, it will add that information to its own ToM.

When striving to establish a distributed ToM across all agents within a multi-agent system, one approach involves creating a shared ToM for the entire

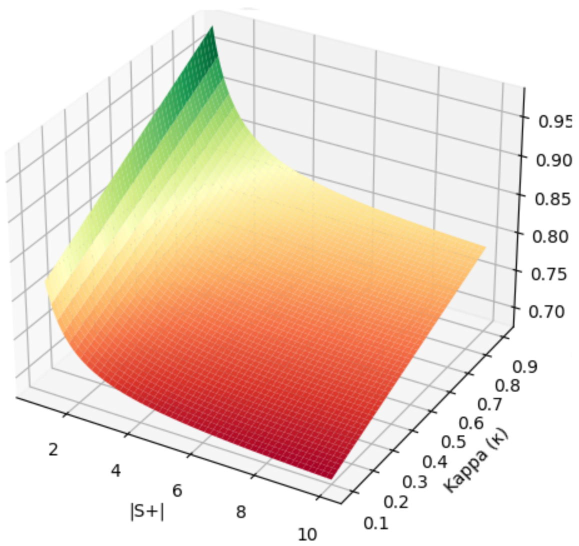


Figure 1: Equation (5) behaviour with γ fixed at 0.8 for all evidences, $|S^+_\varphi|$ ranging from 0 to 10 evidences and $|S^-_\varphi|$ ranging from 0 to 2 evidences (parallelly to $|S^+_\varphi|$).

system. This is achieved by having agents broadcast their mental attitudes from their private ToM using the performative *share_tom*. With the proposed semantics, all agents should converge to a state[2] of ToM, effectively establishing a common distributed ToM for the system. However, this approach can be highly inefficient in systems where agents continuously add and update mental attitudes from others to their individual ToMs, requiring they continually broadcast this new and updated information. Additionally, other agents need to aggregate this new information into their own ToMs, leading to a cycle of continuous updates and broadcasts.

To address this efficiency concern, we introduce the concept of Relevant Distributed ToM (RDToM). RDToM represents a distributed ToM containing information that is specifically relevant to a particular agent. Each agent maintains its RDToM, filtering and retaining only the information that is pertinent to its context. Agents also selectively share information, transmitting only the data that is relevant to other specific agents. Then, the RDToM is utilised by the agent for reasoning and decision-making.

3 CASE STUDY

In this section, we present a case study in the domain of stress detection. There are various professions, and each encounters unique challenges that can cause stress. For instance, office workers engage in demanding knowledge work that requires formal training, high productivity, and creativity, and a stressful condition can affect professional productivity (Awada et al., 2023). Employees with jobs requiring significant mental or physical effort are susceptible to stress, leading to poor performance, mental health issues, and disrupted sleep (Masri et al., 2023). Health workers, in particular, carry a heavy workload and are at a higher risk of infection, especially during events such as those experienced recently with the COVID-19 pandemic (Morales et al., 2022b). Several factors can contribute to occupational stress, such as increased anxiety, frequent injuries, insomnia, and environmental stressors, which are often associated with the workplace. Recently, studies have explored different perspectives for stress measurement, including physiological (biomarkers) (Morales et al., 2022c), psychological, and behavioral aspects (Masri et al., 2023). Differentiating between positive stress (eustress) and negative stress (distress) can be quite challenging because their physical symptoms may seem similar. However, the main difference lies in the emotional and psychological response (Betti et al., 2018). A subjective method of measuring stress is self-reported stress or perceived stress. These instruments typically involve questionnaires and scoring systems to identify stress or similar disorders. Examples include the Perceived Stress Scale, Relative Stress Scale, Brief Symptom Inventory, and others (Sharma and Gedeon, 2012). However, it is worth noting that using questionnaires, especially in a workplace setting, can interrupt the user's workflow. The impact of these instruments on outcomes is discussed in more detail in (Masri et al., 2023).

Numerous studies have been conducted to identify and diagnose stress. For example, researchers have explored the use of biomarkers associated with machine learning and deep learning algorithms to diagnose data collected from wearable devices. To monitor mental health and capture social signals, some wearable devices must be equipped with multiple sensors that work continuously (Morales et al., 2022a). The difference between positive and negative stress has been also investigated due to the complexity of stress detection considering the emotional, physical, and behavioral markers (Pluut et al., 2022). Among the most commonly reported symptoms of stress in these studies are difficulty sleeping, rapid heartbeat, sweating, and mood changes. Various physiological measures have been utilised to detect stress, including skin conductance, heart rate, skin and body temperatures, electrocardiogram signals, and electroen-

[2] They should converge to this state aggregating all information exchanged using the Equation (5), considering all agents share their private ToM.

cephalograms (Giannakakis et al., 2022). It is not possible to detect stress conditions with a single automated approach, so it is necessary to combine multiple approaches, such as sensors for physiological information and multi-agents for emotional information, to identify stress conditions. A precise approach should take physiological, psychological, emotional, and behavioral factors into account. Moreover, individual differences in stress reactions must also be considered. According to the data, accurate identification of stress conditions requires a comprehensive and individualized approach.

In this context, we have applied our approach to the domain of stress detection, where a multi-agent system interacts with a group of people working together. Each individual has a personal agent with whom they interact through a natural language interface implemented using chatbot technologies. Using the chatbot, each user can share their opinions about whether their co-workers are stressed. In addition, the personal agent has the capacity to recognize the stress level of its user. It accomplishes this by combining its user's theory (whether the user is stressed or not) with theories from other personal agents that have shared information about its user. Mary's personal agent may receive information from Paul's personal agent, indicating that Mary is stressed based on Paul's observations. These shared theories contribute to the personal agent's overall understanding. It is important to note that our approach, while primarily considering explicit user interactions regarding stress, also allows agents to consider other inputs. For instance, a computer vision agent can inform personal assistant agents when it believes someone in the group is stressed, with a degree of certainty extracted from its precision in identifying that information.

3.1 Ascribing ToM to Users Using Chatbot Technologies

To facilitate interaction with their users, the assistant agents in our case study are equipped with a natural language understanding interface provided by chatbot technologies, specifically using the open-source Rasa framework[3]. Chatbot technologies have been proposed in the literature as a promising and practical approach to implementing natural language interfaces within multi-agent systems. For example, integrating JaCaMo Framework (Boissier et al., 2013) with Dialogflow[4] (Engelmann et al., 2021). In this work, we follow a similar approach by using the JaCaMo framework to implement multi-agent systems, utilising CArtAgO artifacts (Ricci et al., 2011) integrated with Rasa framework, and a chatbot technology. This integration serves to create a natural language interface between agents and users, enabling one form of scalable agent-agent interoperability (Sarkadi et al., 2022).

In essence, a natural language unit was trained to identify the user's intention during their interaction and extract relevant entities from these communications. In this particular case study, the agents are interested in identifying two user intention, named:

- **Inform Self Stress:** when a user informs their personal agent that they are stressed.
- **Inform Other's Stress:** when a user informs their personal agent that a coworker is stressed.

While identifying the intention of the user behind their interactions is sufficient to inform self stress, when the user intends to inform others about stress, the natural language unit also extracts the coworker's name as an entity. Subsequently, this extracted information is made available to the user's personal agent, according to the semantics of the $tell$ performative introduced in Section 2, inferring what the user believes regarding their own and their coworker's stress. For example, when Mary's personal agent identify that Mary is informing self stress, it adds the belief that Mary is stressed to its ToM, i.e., Mary's personal agent will include the belief $Bel_{mary}(\texttt{stressed(mary)})_{[0.9]}$ in its ToM. Here, as an example, we instantiate γ with the precision returned from the natural language unit's classification of the user's intention[5], in this particular ex-

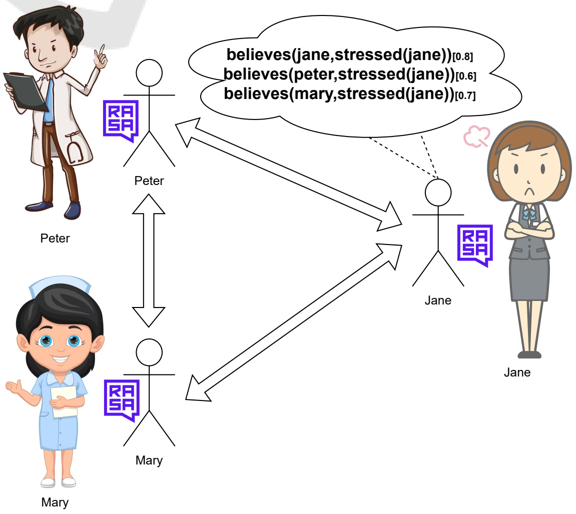
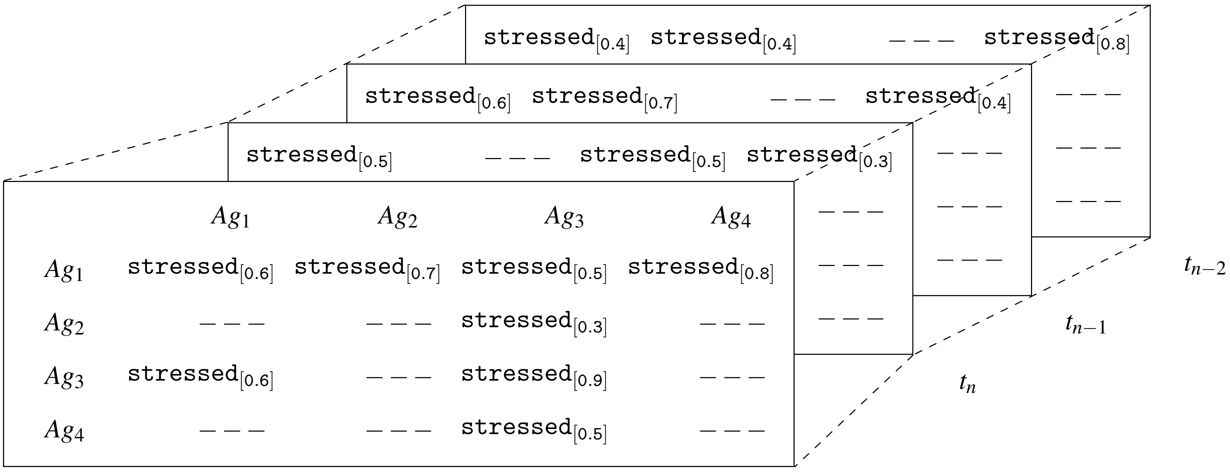

believes(jane,stressed(jane))[0.8]
believes(peter,stressed(jane))[0.6]
believes(mary,stressed(jane))[0.7]
Peter
Peter
Jane
Jane
Mary
Mary

Figure 2: Scenario.

[3] https://rasa.com/

[4] https://cloud.google.com/dialogflow

[5] However, as described in previous sections, a more so-


$\texttt{stressed}_{[0.4]}$ $\texttt{stressed}_{[0.4]}$ — — — $\texttt{stressed}_{[0.8]}$
$\texttt{stressed}_{[0.6]}$ $\texttt{stressed}_{[0.7]}$ — — — $\texttt{stressed}_{[0.4]}$
$\texttt{stressed}_{[0.5]}$ — — — $\texttt{stressed}_{[0.5]}$ $\texttt{stressed}_{[0.3]}$
— — —
t_{n-2}
t_{n-1}
t_n

	Ag_1	Ag_2	Ag_3	Ag_4
Ag_1	$\texttt{stressed}_{[0.6]}$	$\texttt{stressed}_{[0.7]}$	$\texttt{stressed}_{[0.5]}$	$\texttt{stressed}_{[0.8]}$
Ag_2	— — —	— — —	$\texttt{stressed}_{[0.3]}$	— — —
Ag_3	$\texttt{stressed}_{[0.6]}$	— — —	$\texttt{stressed}_{[0.9]}$	— — —
Ag_4	— — —	— — —	$\texttt{stressed}_{[0.5]}$	— — —

Figure 3: Distributed ToM Progress.

ample, $\gamma = 0.9$. Furthermore, when Mary's personal agent identifies that Mary is reporting that Jane is stressed, it adds the belief that Mary believes Jane is stressed to its ToM, i.e., Mary's personal agent adds $Bel_{mary}(\texttt{stressed(jane)})_{[0.7]}$ in its ToM, where $\gamma = 0.7$. In our implementation, this information is represented using first-order predicates as follow: `believes(mary,stressed(mary))[0.9]` and `believes(mary,stressed(jane))[0.7]`, respectively, as it also can be observed in Figure 2.

3.2 Scenario

To evaluate our approach to distributed ToM, we conducted various experiments simulating a group of four individuals working together, named Ag_1, Ag_2, Ag_3, and Ag_4. These individuals interacted with their personal assistants and shared information about their own stress levels and stress levels of their colleagues. The personal agents are named according to their respective users, meaning that the personal agent for Ag_1 is also named Ag_1. The multi-agent system models only the mental attitudes of the users.

Furthermore, agents exclusively share mental attitudes about the users with the users' personal assistants. In other words, when a specific agent, such as Ag_2, models information in its ToM regarding Ag_1, i.e., information about Ag_1 as its user, it sends that mental attitude to Ag_1. However, if Ag_2 models information concerning its own user, it does not share that information. In this scenario, users' information is relevant only to their respective personal assistants. Personal assistants aggregate the distributed ToM in order to reach a more supported conclusion about user's stress.

For example, in Figure 3, we demonstrate a scenario in which four agents interact with their users in order to identify their own or their coworkers stress levels, sharing that information with other agents when relevant. In Figure 3, γ represents the precision returned by the natural language interface in understanding the user's messages. The 4 tables (in depth) in the figure represent discrete point in time (three different timestamp), resulting in the final distributed ToM model at forefront in Figure 3. To simplify the representation in Figure 3, rows in the tables represent the agents (and their respective user), and columns represent the target users (and their respective agents) of the ToM model. For example, in the front table of Figure 3, the entry in the first row and first column is $\texttt{stressed}_{[0.6]}$, signifying that Ag_1 believes that Ag_1 is stressed, with a degree of certainty of 0.6, i.e., $Bel_{Ag_1}(\texttt{stressed(Ag}_1\texttt{))}_{[0.6]}$ in our formal model.

In this particular scenario, agent Ag_1 has interacted with its user (also referred to as Ag_1), adding models about the user being stressed to its ToM at each interaction, i.e., $Bel_{Ag_1}(\texttt{stressed(Ag}_1\texttt{))}_{[\gamma]}$. Using these models, the agent can infer a final model $Bel_{Ag_1}(\texttt{stressed(Ag}_1\texttt{))}_{[0.6]}$, in which this degree of certainty is derived from trace of interactions, as represented by the tables at the back. In this particular example, agents are trace-concerned agents, according to definition 2.2, utilising Equation (2), with $\gamma = 0.6$ calculated as follow:

$$Bel_{Ag_1}(\texttt{stressed(Ag}_1\texttt{))}_{[0.6]} =$$

$$\underset{\gamma}{\text{argmax}} \begin{Bmatrix} Bel_{Ag_1}(\texttt{stressed(Ag}_1\texttt{))}_{[0.4,t_{n-2}]} \\ Bel_{Ag_1}(\texttt{stressed(Ag}_1\texttt{))}_{[0.6,t_{n-1}]} \\ Bel_{Ag_1}(\texttt{stressed(Ag}_1\texttt{))}_{[0.5,t_n]} \end{Bmatrix}$$

phisticated degree of certainty can be implemented to suit the application's needs.

Additionally, Ag_1's user has mentioned, at some point, that their coworkers, named Ag_2, Ag_3, and Ag_4, are stressed as well. As a result, Ag_1 has the following models in its ToM: $Bel_{Ag_1}(\texttt{stressed(Ag}_2\texttt{))}_{[0.7]}$, $Bel_{Ag_1}(\texttt{stressed(Ag}_3\texttt{))}_{[0.5]}$, $Bel_{Ag_1}(\texttt{stressed(Ag}_4\texttt{))}_{[0.8]}$. Similarly, Ag_1 has aggregated the multiple evidences its user has provided about the coworkers using Equation (2), for example, inferring the Ag_2's stress, as follow:

$$Bel_{Ag_1}(\texttt{stressed(Ag}_2\texttt{))}_{[0.7]} =$$

$$\operatorname*{argmax}_{\gamma} \begin{Bmatrix} Bel_{Ag_1}(\texttt{stressed(Ag}_2\texttt{))}_{[0.4,t_{n-2}]} \\ Bel_{Ag_1}(\texttt{stressed(Ag}_2\texttt{))}_{[0.7,t_{n-1}]} \end{Bmatrix}$$

In Figure 3, we also observe that Ag_2 has informed that Ag_3 is stressed, Ag_3 has informed Ag_1 and itself are stressed, and Ag_4 has informed that Ag_3 is stressed. All pieces of evidence are aggregated for these agents according to Equation (2), resulting on the distributed ToM shown at the front table of Figure 3.

Of course, as mentioned in Section 2.1, different agents profiles could be considered, according to the application needs. For example, when using Equation (3) instead of Equation (2), i.e., agents care more about the most updated information from the trace, we would have $Bel_{Ag_1}(\texttt{stressed(Ag}_1\texttt{))}_{[\gamma]}$ with $\gamma = 0.5$ calculated as follow:

$$Bel_{Ag_1}(\texttt{stressed(Ag}_1\texttt{))}_{[0.5]} =$$

$$\operatorname*{argmax}_{t} \begin{Bmatrix} Bel_{Ag_1}(\texttt{stressed(Ag}_1\texttt{))}_{[0.4,t_{n-2}]} \\ Bel_{Ag_1}(\texttt{stressed(Ag}_1\texttt{))}_{[0.6,t_{n-1}]} \\ Bel_{Ag_1}(\texttt{stressed(Ag}_1\texttt{))}_{[0.5,t_n]} \end{Bmatrix}$$

After agents aggregating those information modelled in their ToM, they are capable of sharing those models with other agents, as introduced in Section 2.2. In our case study, agents will share those information with agents for whom their users are the subjects of the model. In other words, agents represented by the rows in the from table in Figure 3 will share these models with agents represented by the columns in the front table of Figure 3. When all agents share information related to other agents, they collectively reach a RDToM that corresponds to the columns in the front table of Figure 3. For example, when all agents share with Ag_3 what Ag_3's coworkers think about whether they are stressed or not, agent Ag_3 will have the following[6] RDToM:

$$RDToM_{Ag_3} = \begin{Bmatrix} Bel_{Ag_1}(\texttt{stressed(Ag}_3\texttt{))}_{[0.5]} \\ Bel_{Ag_2}(\texttt{stressed(Ag}_3\texttt{))}_{[0.3]} \\ Bel_{Ag_3}(\texttt{stressed(Ag}_3\texttt{))}_{[0.9]} \\ Bel_{Ag_4}(\texttt{stressed(Ag}_3\texttt{))}_{[0.5]} \end{Bmatrix}$$

Subsequently, Ag_3 aggregates these models from its RDToM using the Equation (5), reaching $Bel_{dtom}(\texttt{stressed(Ag}_3\texttt{))}_{[0.82]}$, with $\kappa = 0.6$. Similarly, Ag_1 reaches $Bel_{dtom}(\texttt{stressed(Ag}_1\texttt{))}_{[0.84]}$, Ag_2 reaches $Bel_{dtom}(\texttt{stressed(Ag}_2\texttt{))}_{[0.88]}$, and Ag_4 reaches $Bel_{dtom}(\texttt{stressed(Ag}_2\texttt{))}_{[0.92]}$

4 PROPERTIES

An inherent property of our approach is the ability of agents to combine both software agents' and humans users' ToM. When an agent directly interacts with a human user, it can model a ToM about that user, for example, $Bel_{user1}(\texttt{stressed(user1))}_{[0.8]}$. When the user informs other members of the working group, their personal agent can model its user's ToM about other individuals, i.e., $Bel_{user1}(\texttt{stressed(user2))}_{[0.8]}$. By sharing this information with other assistants, they can combine their ToM about their user with the ToM of other users about their user. For instance, `user1`'s personal agent models $Bel_{user1}(\texttt{stressed(user1))}_{[0.8]}$ and receives $Bel_{user2}(\texttt{stressed(user1))}_{[0.7]}$, indicating that another user believes `user1` is stressed. This information can then be combined using Equation (5). Also, our approach allow agents to reach (Relevant) Distributed ToM.

Agents reach a distributed ToM. When a group of agents $\{ag_1, ag_2 \ldots, ag_n\}$ have N mental attitudes in their respective ToM about another agent ag_j, they can collectively reach a distributed ToM about ag_j executing broadcast messages using the performative *share_tom*, as defined in its semantics in Section 2.2. □

Agents reach a Relevant Distributed ToM. When a group of agents $\{ag_1, ag_2 \ldots, ag_n\}$ have N mental attitudes in their ToM about another agent ag_j, and these mental attitudes are relevant to a particular agent ag_i, ag_i can reach a RDToM about ag_j receiving N messages from $\{ag_1, ag_2 \ldots, ag_n\}$ with the performative *share_tom*, as defined in its semantics in Section 2.2. □

[6] Corresponding to the column labelled as Ag_3 in Figure 3.

5 RELATED WORK

A concise overview of how ToM has been applied in agent-based modelling and multi-agent systems is given in (Rocha et al., 2023).

There are works representing ToM in AOPL, such as (Cantucci and Falcone, 2020; Cantucci and Falcone, 2022) representing ToM in JaCaMo (Boissier et al., 2013), (Harbers et al., 2011) representing ToM in 2APL (Dastani, 2008), and (Chang and Soo, 2008) representing ToM in JADE (Bellifemine et al., 2005). Additionally, the work by (Montes et al., 2022; Montes et al., 2023) introduces an abductive reasoning approach for argumentation and ToM in AOPLs, while (Mosca et al., 2020; Mosca and Such, 2022) emphasises the need for ToM in generating social explanations based on decisions reached through abduction and value-based argumentation in multi-agent scenarios. Also, there are works using ToM to represent emotions, which is close to our case study (Feng et al., 2019; Reisenzein et al., 2013). Furthermore, there are works that use propositional logic and text to represent ToM (Gebhard et al., 2018; Walton, 2019; Husemann et al., 2022).

Our approach is based on the idea of a collective Theory of Mind (Shteynberg et al., 2023). The work closest to our approach is that of (Westby and Riedl, 2023) who used an approach for developing a network of Bayesian agents that collectively model the mental states of teammates from the observed communication. (Westby and Riedl, 2023) calibrate their model on human experiments to show how humans model themselves and their mental state as a collective.

Our work distinguishes itself from all of the above because we propose an approach for distributed ToM in which agents can model and aggregate not only the mental attitudes of software or AI agents but also those of human users in the same multi-agent system. We model the Theory of a Collective Mind of a Hybrid Society, e.g. a society where both humans and machines are socially interactive agents (Sarkadi et al., 2021; Sarkadi, 2023).

6 CONCLUSION

In this work, we have presented an approach for distributed ToM within MAS. Specifically, we introduced the concept of Relevant Distributed ToM, where agents selectively share information that is relevant to other agents within the system, and only with those agents for whom the information should be relevant.

We have demonstrated our approach through a case study focused on stress detection. In this case study, agents were capable to aggregate multiple mental models from the distributed ToM, allowing them to draw more robust conclusions about users' stress levels. The case study incorporates a natural language interface implemented using chatbot technologies. By interacting with users through this interface, agents not only model the mental attitudes of the users but also the users' ToM regarding their coworkers. Agents then share this ToM to support personal agents in making inferences about their users' stress levels.

While our case study focuses on simulating user interactions to demonstrate the proposed approach, our future work aims to conduct evaluations in real-life scenarios where human users directly interact with the system.

REFERENCES

Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., et al. (2020). A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28.

Austin, J. L. (1975). *How to do things with words*, volume 88. Oxford university press.

Awada, M., Becerik-Gerber, B., Lucas, G., and Roll, S. C. (2023). Predicting office workers; productivity: A machine learning approach integrating physiological, behavioral, and psychological indicators. *Sensors*, 23(21).

Bellifemine, F., Bergenti, F., Caire, G., and Poggi, A. (2005). Jade—a java agent development framework. *Multi-agent programming: Languages, platforms and applications*, pages 125–147.

Betti, S., Lova, R. M., Rovini, E., Acerbi, G., Santarelli, L., Cabiati, M., Ry, S. D., and Cavallo, F. (2018). Evaluation of an integrated system of wearable physiological sensors for stress monitoring in working environments by using biological markers. *IEEE Transactions on Biomedical Engineering*, 65(8):1748–1758.

Boissier, O., Bordini, R. H., Hübner, J. F., Ricci, A., and Santi, A. (2013). Multi-agent oriented programming with jacamo. *Science of Computer Programming*, 78(6):747–761.

Cantucci, F. and Falcone, R. (2020). Towards trustworthiness and transparency in social human-robot interaction. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pages 1–6. IEEE.

Cantucci, F. and Falcone, R. (2022). Collaborative autonomy: Human–robot interaction to the test of intelligent help. *Electronics*, 11(19):3065.

Chang, H.-M. and Soo, V.-W. (2008). Simulation-based story generation with a theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence* *and Interactive Digital Entertainment*, volume 4(1), pages 16–21.

Dastani, M. (2008). 2apl: a practical agent programming language. *Autonomous agents and multi-agent systems*, 16:214–248.

Engelmann, D., Damasio, J., Krausburg, T., Borges, O., Colissi, M., Panisson, A. R., and Bordini, R. H. (2021). Dial4jaca–a communication interface between multi-agent systems and chatbots. In *Int. conference on practical applications of agents and multi-agent systems*, pages 77–88. Springer.

Feng, D., Carstensdottir, E., El-Nasr, M. S., and Marsella, S. (2019). Exploring improvisational approaches to social knowledge acquisition. In *Int. Conference on Autonomous Agents and MultiAgent Systems*.

FIPA, T. (2008). Fipa communicative act library specification. *Foundation for Intelligent Physical Agents, http://www. fipa. org/specs/fipa00037/SC00037J. html (30.6. 2004)*.

Gebhard, P., Schneeberger, T., Baur, T., and André, E. (2018). Marssi: Model of appraisal, regulation, and social signal interpretation. In *International conference on Autonomous agents and multi-agent systems*.

Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., and Tsiknakis, M. (2022). Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, 13(1):440–460.

Goldman, A. I. et al. (2012). Theory of mind. *The Oxford handbook of philosophy of cognitive science*, 1.

Harbers, M., van den Bosch, K., and Meyer, J.-J. C. (2011). Agents with a theory of mind in virtual training. In *Multi-Agent Systems for Education and Interactive Entertainment: Design, Use and Experience*, pages 172–187. IGI Global.

Husemann, S., Pöppel, J., and Kopp, S. (2022). Differences and biases in mentalizing about humans and robots. In *IEEE International Conference on Robot and Human Interactive Communication*, pages 490–497.

Masri, G., Al-Shargie, F., Tariq, U., Almughairbi, F., Babiloni, F., and Al-Nashash, H. (2023). Mental stress assessment in the workplace: A review. *IEEE Transactions on Affective Computing*, pages 1–20.

Mayfield, J., Labrou, Y., and Finin, T. W. (1995). Evaluation of kqml as an agent communication language. In Wooldridge, M., Müller, J. P., and Tambe, M., editors, *ATAL*, volume 1037, pages 347–360. Springer.

Melo, V. S., Panisson, A. R., and Bordini, R. H. (2016). Argumentation-based reasoning using preferences over sources of information. In *International Conference on Autonomous Agents & Multiagent Systems, 2016, Cingapura.*

Melo, V. S., Panisson, A. R., and Bordini, R. H. (2017). Meta-information and argumentation in multi-agent systems. *iSys-Brazilian Journal of Information Systems*, 10(3):74–97.

Montes, N., Luck, M., Osman, N., Rodrigues, O., and Sierra, C. (2023). Combining theory of mind and abductive reasoning in agent-oriented programming. *Autonomous Agents and Multi-Agent Systems*, 37(2):36.

Montes, N., Osman, N., and Sierra, C. (2022). Combining theory of mind and abduction for cooperation under imperfect information. In *European Conference on Multi-Agent Systems*, pages 294–311. Springer.

Morales, A., Barbosa, M., Morás, L., Cazella, S. C., Sgobbi, L. F., Sene, I., and Marques, G. (2022a). Occupational stress monitoring using biomarkers and smartwatches: A systematic review. *Sensors*, 22(17).

Morales, A. S., de Oliveira Ourique, F., Morás, L. D., Barbosa, M. L. K., and Cazella, S. C. (2022b). *A Biomarker-Based Model to Assist the Identification of Stress in Health Workers Involved in Coping with COVID-19*, pages 485–500. Springer.

Morales, A. S., de Oliveira Ourique, F., Morás, L. D., and Cazella, S. C. (2022c). *Exploring Interpretable Machine Learning Methods and Biomarkers to Classifying Occupational Stress of the Health Workers*, pages 105–124. Springer International Publishing, Cham.

Mosca, F., Sarkadi, Ş., Such, J. M., and McBurney, P. (2020). Agent expri: Licence to explain. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9–13, 2020, Revised Selected Papers 2*, pages 21–38.

Mosca, F. and Such, J. (2022). An explainable assistant for multiuser privacy. *Autonomous Agents and Multi-Agent Systems*, 36(1):10.

Panisson, A., Sarkadi, S., McBurney, P., Parsons, S., and Bordini, R. (2018). Lies, bullshit, and deception in agent-oriented programming languages. In *Proc. of the 20th International Trust Workshop*, pages 50–61.

Panisson, A. R., Sarkadi, S., McBurney, P., Parsons, S., and Bordini, R. H. (2019). On the formal semantics of theory of mind in agent communication. In *Agreement Technologies: 6th International Conference, AT 2018, Bergen, Norway, December 6-7, 2018, Revised Selected Papers 6*, pages 18–32. Springer.

Parsons, S., Atkinson, K., Haigh, K. Z., Levitt, K. N., McBurney, P., Rowe, J., Singh, M. P., and Sklar, E. (2012). Argument schemes for reasoning about trust. *COMMA*, 245:430.

Pluut, H., Curșeu, P. L., and Fodor, O. C. (2022). Development and validation of a short measure of emotional, physical, and behavioral markers of eustress and distress (meds). *Healthcare*, 10(2).

Reisenzein, R., Hudlicka, E., Dastani, M., Gratch, J., Hindriks, K., Lorini, E., and Meyer, J.-J. C. (2013). Computational modeling of emotion: Toward improving the inter-and intradisciplinary exchange. *IEEE Transactions on Affective Computing*, 4(3):246–266.

Ricci, A., Piunti, M., and Viroli, M. (2011). Environment programming in multi-agent systems: An artifact-based perspective. *Autonomous Agents and Multi-Agent Systems*, 23(2):158–192.

Rocha, M., da Silva, H. H., Morales, A. S., Sarkadi, S., and Panisson, A. R. (2023). Applying theory of mind to multi-agent systems: A systematic review. In *Brazilian Conference on Intelligent Systems*, pages 367–381. Springer.

Sarkadi, Ş., Panisson, A. R., Bordini, R. H., McBurney, P., and Parsons, S. (2019). Towards an approach for modelling uncertain theory of mind in multi-agent systems. In *Agreement Technologies: 6th International Conference, AT 2018, Bergen, Norway, December 6-7, 2018, Revised Selected Papers 6*, pages 3–17. Springer.

Sarkadi, Ş., Rutherford, A., McBurney, P., Parsons, S., and Rahwan, I. (2021). The evolution of deception. *Royal Society open science*, 8(9):201032.

Sarkadi, Ş., Tettamanzi, A.G.B. and Gandon, F. (2022). Interoperable AI: Evolutionary Race Toward Sustainable Knowledge Sharing. *IEEE Internet Computing*, 26(6):25-32.

Sarkadi, Ş. (2023). An Arms Race in Theory-of-Mind: Deception Drives the Emergence of Higher-level Theory-of-Mind in Agent Societies. In *Proc. of 2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)*, pages 1–10.

Sharma, N. and Gedeon, T. (2012). Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer Methods and Programs in Biomedicine*, 108(3):1287–1301.

Shteynberg, G., Hirsh, J. B., Wolf, W., Bargh, J. A., Boothby, E. B., Colman, A. M., Echterhoff, G., and Rossignac-Milon, M. (2023). Theory of collective mind. *Trends in Cognitive Sciences*.

Walton, D. (2019). Using argumentation schemes to find motives and intentions of a rational agent. *Argument & Computation*, 10(3):233–275.

Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation schemes*. Cambridge University Press.

Westby, S. and Riedl, C. (2023). Collective intelligence in human-ai teams: A bayesian theory of mind approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6119–6127.