HHAI2022: Augmenting Human Intellect
S. Schlobach et al. (Eds.)

doi:10.3233/FAIA220199

Abstracting Minds: Computational Theory of Mind for Human-Agent Collaboration

Emre ERDOGAN [a] and Frank DIGNUM [b] and Rineke VERBRUGGE [c] and Pınar YOLUM [a]

[a] *Utrecht University, Utrecht, Netherlands*
[b] *Umeå University, Umeå, Sweden*
[c] *University of Groningen, Groningen, Netherlands*

Abstract. Theory of mind refers to the human ability to reason about mental content of other people such as beliefs, desires, and goals. In everyday life, people rely on their theory of mind to understand, explain, and predict the behaviour of others. Having a theory of mind is especially useful when people collaborate, since individuals can then reason on what the other individual knows as well as what reasoning they might do. Realization of hybrid intelligence, where an agent collaborates with a human, will require the agent to be able to do similar reasoning through computational theory of mind. Accordingly, this paper provides a mechanism for computational theory of mind based on abstractions of single beliefs into higher-level concepts. These concepts can correspond to social norms, roles, as well as values. Their use in decision making serves as a heuristic to choose among interactions, thus facilitating collaboration on decisions. Using examples from the medical domain, we demonstrate how having such a theory of mind enables an agent to interact with humans efficiently and can increase the quality of the decisions humans make.

Keywords. social cognition, abstraction, heuristics, collaboration, communication, human-inspired computational model

1. Introduction

Hybrid intelligence requires human-agent collaboration, where a human and a computational agent complement each other in the tasks that they achieve and interactions require a mixed initiative. In order to realize successful collaborations, agents need to be empowered with capabilities that humans use on an everyday basis. One of these crucial capabilities is the modeling of Theory of Mind (ToM). Put simply, this capability enables a human to reason about other humans, making it possible to understand and predict their behaviour [1,2,3]. It is even possible for humans to use higher-order ToM reasoning to infer how others employ ToM (e.g., I believe that Alice does not know that I am an expert on this topic). Due to this capability of ToM, humans exhibit what are called social skills to carry out tasks effectively and efficiently and allowing human social interactions to create added value to all parties.

To understand how ToM works, various computational models have been developed. An important line of research analyzed its use in game settings where the rules of the game are well-defined and possible behaviours are limited [4,5,6,7,8,9,10]. Experiments in competitive, cooperative, as well as mixed-motive settings show that agents equipped with ToM reasoning achieve better results compared to the agents without them. Various techniques to model ToM exist. For example, Baker *et al.* [11] model ToM within a Bayesian framework using partially observable Markov decision processes. Their evaluation in a simple spatial setting is promising. Winfield [12] shows how robots use a ToM model by imitating other robots' actions. Using simple ethical rules, they show that ToM helps to improve robots' safety.

There has been a lot of research on human-machine collaboration in various domains such as negotiation [13], planning [14], and behavioral support systems [15]. However, the use of computational ToM in human-machine collaboration is relatively novel. Hiatt *et al.* [16] describe a ToM robot model based on the ACT-R cognitive architecture [17] to account for human behavioral variability in human-robot teams. Devin and Alami [18] develop a ToM agent framework for collaborative task achievement. Their system takes mental states regarding the goals, plans, and actions of humans into account when executing human-robot shared plans. Görür *et al.* [19] propose a ToM agent model for estimating humans' intentions in a shared human-robot task. Buehler and Weisswange propose a ToM-based communication framework for human-agent cooperation [20]. They combine Bayesian inference with planning under uncertainty to evaluate the effect of ToM-based communication on joint performance in an illustrative scenario. Lim *et al.* [21] design a Bayesian ToM-inspired [22] agent model and investigate the performance of humans with agents with and without a ToM in a collaborative setting. The results of these studies around computational ToM models are also generally promising and collectively suggest that the use of ToM can have positive impacts on human-agent collaboration.

An important area where ToM could be of particular use is hybrid intelligence [23], where an agent can collaborate with a human towards a particular goal, where the agent would have varying capabilities that could complement those of the human to yield the goal. As an example, consider a computational agent doctor that is designed to collaborate with a human doctor. Such an agent doctor's capabilities can include cooperating with surgeons in operations [24] as well as providing assistance to improve medical diagnosis processes [25]. For a more complete human-agent collaboration to take place, an ideal agent doctor should not only function as a medicinal support tool but also be able to understand the doctor's behaviour, communicate well with her, and continuously learn from their shared experience as well. Thus, we argue that the agent doctor would benefit from having a functional ToM for the human doctor in achieving their collective goals in such hybrid settings.

Realizing such a ToM model is useful but difficult. Most of the existing models start by modeling individual beliefs about others and build a ToM model based on that. However, in a complicated setting as described above, there will be too many beliefs that the agent acquires over time, where some of these will be applicable in certain situations, whereas others will be useful in others. Storing, maintaining, and using these individual beliefs will be ineffective over time. To remedy this, we propose an abstraction framework for ToM through which a set of individual beliefs can be aggregated to produce abstraction concepts. The underlying idea is to employ an agent's belief and knowledge set to produce more abstract, complex concepts for the agent to benefit from when in-

teracting with humans. These concepts can correspond to various social norms, human values as well as emotions among individuals. Collectively, they serve as human-inspired heuristics for the agent to make effective decisions.

To investigate the principle of abstraction, we start with an example featuring a setting in which an agent doctor and a human doctor collaborate towards a medical diagnosis. To make it more concrete, we computationally model several human decision-making heuristics and show how ToM reasoning can be efficiently used within our abstraction procedure. We subsequently indicate the importance of social roles, norms, and values with respect to the interaction context and extend our example to illustrate how these can be integrated naturally into our framework. Integrating roles, norms, and values helps the agent to choose among different actions to yield a result that would fit the current situation better.

The rest of this paper is organized as follows. Section 2 sets up our working example. Section 3 describes our abstraction heuristics. Section 4 illustrates how a computational agent can combine abstraction with ToM reasoning in its decision-making mechanism. Section 5 explains how we integrate social roles, values, and norms in our abstraction framework. Section 6 provides an outline of our current work and and points to future research directions.

2. Working Example

Our setup consists of an agent doctor A and a human doctor B that work together towards the diagnosis of a patient C's health problem. In this setting, the core objective of A is to use its capabilities to complement those B. Thus, they share the workload according to their strengths during the diagnostic process [26]. For example, B can perform the patient interview and the physical examination processes, while A can work on the diagnostic testing (e.g., analyzing MRI scan results [27]).

Although Artificial Intelligence (AI) research in health continues to progress [28, 29], the usual paradigm suggests that AI agents as well as robots and software applications are treated as supporting tools that doctors can use. Doctors have the final say in the medical procedure and can neglect the information that such agents may provide altogether. However, within our working example, we give equal rights to both agents and humans in the diagnostic process; thus, we have both an "agent doctor" and a "human doctor". Essentially, our example provides a collective decision-making process in which A and B can share their findings with each other, assess each other's work, and agree on the diagnosis together in an interactive manner. Although we do not explicitly discuss this point, the interaction can include that the human doctor explains her decision to the agent doctor. This will in itself also be a good check for the human doctor on the correctness of that decision.

Now, suppose that a difference of opinion has arisen between A and B during their discussion for the diagnosis of C's health problem. For instance, B may say that the clinical interview R_1 and the physical examination results R_2 (provided by B) together point to a specific disease D_1 but A may say that it can be another disease D_2 according to the diagnostic testing results R_3 (provided by A). B may further add that they should give low importance to the diagnostic testing results R_3 because the disease is nearly always D_1 when similar physical examination results are observed. In this case, a simple thing

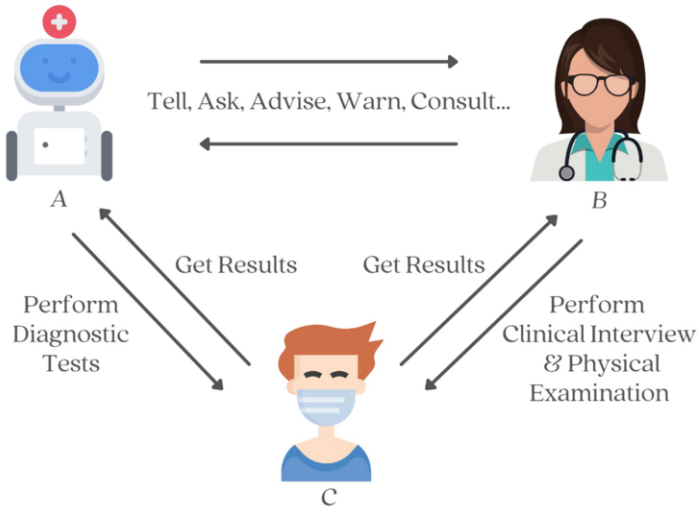

Figure 1. Hybrid Collaboration in Medicine: A computational agent doctor (A) and a human doctor (B) are working together towards the diagnosis of a patient's (C) health problem. Each doctor has different set of capabilities that would be useful for the diagnosis.

for *A* to do can be checking whether it should insist on its own diagnosis decision and elaborate on its findings or simply accept *B*'s decision, say, because of time constraints.

Compared to a medical support tool, one can see that the possible ways that *A* can utilize its beliefs and knowledge are too many. In our scenario, the set of possible actions that *A* can do includes *doing interactive reasoning* to check whether a diagnostic result is of good quality, *warning B* about poor quality results, *advising B* to put more emphasis on one result rather than another, *consulting* another doctor *E*, *telling B* its beliefs and knowledge, and *asking B*'s opinion on a subject that is relevant to *C*'s health problem.

In the remainder of this paper, we will refer to *A* as "it", *B* as "her", and *C* as "him" for practical purposes when necessary. Figure 1 outlines the interaction that takes place among the agent doctor *A*, the human doctor *B*, and the patient *C* during the diagnostic process.

3. Abstraction Framework

In this section, we lay the foundations of our computational ToM framework. At heart, what we envision is an agent that can simplify its beliefs and knowledge into more abstract, compact representations that can serve for heuristics in its decision-making processes. Computationally, an "abstraction mechanism" is an agent instrument that does the following (Figure 2):

1. It takes a set of beliefs and knowledge as input.
2. Using a shared prominent characteristic of such input, it produces an intermediate output in the form of a simple yet more abstract belief or piece of knowledge, or simply an *abstraction*, which shares the same characteristic and is stable over some period of time independent of (small) changes in the beliefs of that agent.

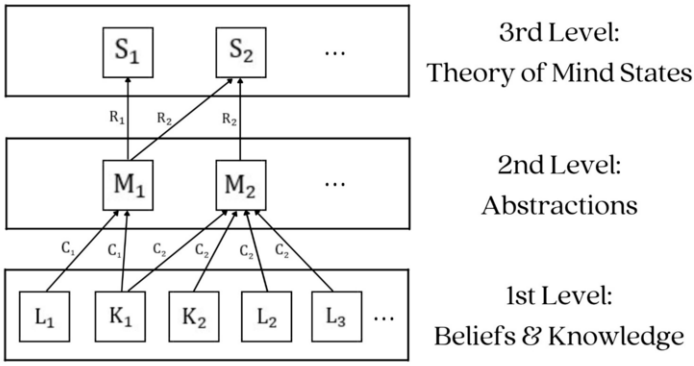

Figure 2. Abstraction procedure: Individual beliefs (L_i) and knowledge (K_j) are used to create abstractions M_k which are ultimately transformed into Theory of Mind states S_n.

3. Applying rules that govern the role of the intermediate output, it produces ToM states for the agent to operate in.

It is well-known that humans use behavioural heuristics in their decision-making processes [30]. In order to capture these heuristics, we make use of an abstraction procedure embedded in a ToM agent. By computing various ToM states, an agent can then make use of the heuristics in its own decision-making processes. Figure 2 illustrates our three-level approach to such an abstraction mechanism. The first level holds the set of beliefs and knowledge about others that could come from different sources, such as observations or explicitly stated information. With this set, the agent creates abstractions in the second level. The first level influences the second level; thus, if the agent observes more information at the first level, the abstractions in the second level might be updated. However, the general idea is that this update does not have to be done frequently. E.g., if the abstraction models that an agent is considered to be trustworthy and this is based on taking a weighted average score of observed actions fulfilling or violating a commitment, then the trustworthiness can be updated frequently at the start, but will not change much after a thousand observations. The abstractions in the second level influence how the agent operates in the third level. Again, changes in the second level will influence the third level. One can think of the third level as a set of beacons for the agent to follow with respect the context of interaction it is in. Figure 2 also shows that beliefs and knowledge can have multiple characteristics C_k, C_l, and so on, which guide the production of the corresponding abstractions M_k, M_l, and so on; multiple abstractions can be used to produce a ToM state S_n with respect to the corresponding rule R_n.

Note that abstractions are not designed to prevent agents from using their beliefs and knowledge directly. Instead, abstractions act as additions that require low maintenance and that are used whenever possible to avoid having to use too much information. In the remainder of this paper, we will illustrate how our abstraction mechanism can be utilized efficiently in various ways to foster successful human-agent collaboration, starting with a simple example to explain both the concept of abstraction and the mechanism more concretely.

4. Abstraction and Theory of Mind

We propose that in principle, our abstraction approach can take into account complex human notions. For our dispute example, we demonstrate our intuition in a specific type of abstraction mechanism, which captures *respect*: "a feeling of deep admiration for someone or something elicited by their abilities, qualities, or achievements" [31]. For the sake of the example, we can summarize it to "excellence-induced respect". Using the abstraction mechanism accordingly to assess *B*'s (medicinal) proficiency, *A* can computationally capture respect by following the steps below:

(1) *A* first takes the following beliefs and knowledge as input:

(1.a) "I know that *B* is a *distinguished* alumna of a top-tier medical school."
(1.b) "I believe that she has a *perfect* diagnostic track record."
(1.c) "I know that she recently won an *award* for being the *best* neurosurgeon in Europe."

(2) Using a shared characteristic of input (i.e., "excellence"), it produces a simple yet more abstract intermediate output in the form of another belief: "I believe *B* is a very *excellent* human doctor".
(3) Applying a rule that governs the role of the intermediate output, it produces a ToM state for the agent's use: "I believe *B* is a very *respectable* doctor because of her many *accomplishments* in the medicinal domain (i.e., I respect *B*)".

An agent can use such a state of respect within its decision-making processes when making further decisions. In our case, *A* can give up insisting on its own decision due to its respect towards *B*, in case of a disagreement with *B*. Note that the *respect* can be revoked when *A* observes *B* making a clear and big mistake.

What makes this abstraction heuristic more valuable, however, is that *A* can also *attribute the whole procedure to B* with the help of recursive ToM reasoning. We theorize that such a ToM agent can then interpret how people feel respect towards others. We illustrate such an ideal behaviour of *A* below:

(1.a) "I know that *B* knows that I (i.e., the agent *A*) am a *distinguished* alum of a top-tier medical school for Artificially Intelligent Robots."
(1.b) "I believe that she believes that I have a *perfect* diagnostic track record."
(1.c) "I believe that she knows that I recently won an *award* for being the *best* computational agent doctor in Europe."
(2) "Thus, I believe that *B* believes that I am a very *excellent* agent doctor".
(3) "I believe that *B* believes that I am a very *respectable* agent doctor because of my *excellence* in the medicinal domain (i.e., I believe that *B* respects me)."

By attributing the abstraction mechanism for respect to *B*, *A* can now form this new high-level belief that *B* respects it. Tweaking our simple dispute case a bit, now suppose that *A* also decides that it does not respect *B* (for instance, due to *B* being an inexperienced doctor). By using these two ToM states, *A* can choose to insist on its diagnostic judgment instead of accepting *B*'s own and nudge *B* to take into account the results of the diagnostic tests as well and have another go at it before making her decision.

5. Abstraction Concepts: Roles, Norms, and Values

In this section, we will illustrate how social roles, norms, and values can be incorporated into our abstraction paradigm. These concepts are important for capturing social interactions. Thus, being able to represent and reason on them would enable an agent to produce effective interactions with the human.

5.1. Roles

A "role" is a socially expected set of behavior that is determined by an individual's status or position in society. Humans, as well as agents, can have multiple social roles in the groups to which they belong [32]. Having a capable ToM that can properly differentiate among these roles is essential to understand how the role-governance dynamics (e.g., norms, goals, emotions, beliefs, etc.) might vary with the corresponding roles.

We have shown in the previous abstraction example, in which both *A* and *B* have the role of "collaborator", that computationally capturing and interpreting respect can be helpful in case of a conflict between collaborators. Looking from another perspective, *A* and *B* also have the role of "doctor" as they are in a doctor-patient relation with *C*. In the patient interviewing process, founding a good relationship between a doctor and the patient is deemed important since it can affect the quality of the acquired clinical history and hence the determination of the diagnosis [33]. It is known that people are inclined to get along with others that they like [34,35] because they can feel comfortable with the people that they feel affinity towards, share more about themselves, and are more eager to listen [36,37]. The patient-doctor relationship is a good example in which a mutual affinity between *B* and *C* can be beneficial in determining the correct diagnosis for *C*.

We claim that *A* can computationally capture affinity within an abstraction mechanism as well. To analyze whether *C* feels affinity towards *B* (and vice versa), *A* can utilize its knowledge and beliefs about the interaction that takes place between *B* and *C*. *A* can check the length of *C*'s answers to *B*'s questions, the amount of sensitive information that *C* shares, *C*'s past experience with *B* (if it exists), and *B*'s notes on *C*'s overall demeanor (e.g., shy, aggressive, cautious etc.) to understand whether *C* feels comfortable with *B* and opens up to her during the interview. Furthermore, *A* can compare the duration of the interview, the number of *B*'s questions, and the completeness of *C*'s clinical history with the standards that need to be met within the corresponding medical sub-domain. These points should help *A* to arrive at a more compact judgment on the quality of interaction between *B* and *C*. During their discussion, *A* can further assess *B*'s communicative effort, which can induce a feeling of closeness in *C* (e.g., demonstrating empathy, acknowledging *B* and *C*'s shared similarities, explicitly taking *C*'s needs, values, and preferences into account etc.). Using these abstract judgments, *A* can then decide if there is a shared affinity between *B* and *C* or not.

Now suppose that *A* arrives at the conclusion that there is a lack of affinity that could have produced a patient interview of poor quality. *A* can then warn *B* about that in their discussion and advise her to use the interview information cautiously. *A* can further support its argument by providing accompanying reasons (e.g., shortness of the duration, lack of detailed questions/answers, lack of empathy, etc.) and suggest putting more emphasis on the diagnostic testing results R_3 rather than the interview R_1 and/or seeking consultation with another doctor *E*. Below, we give an exemplary reasoning that *A* can make in this case (Table 1):

$$
\begin{array}{l}
DoctorPatient(B,C) \rightarrow Doctor(B) \wedge Patient(C) \\
(DoctorPatient(B,C) \wedge \neg GoodInteraction(B,C)) \rightarrow \neg Affinity(B,C) \\
(\neg Affinity(B,C) \wedge Result(B,C,R_1)) \rightarrow \neg GoodResult(B,C,R_1) \\
(\neg Affinity(B,C) \wedge \neg GoodResult(B,C,R_1)) \rightarrow Warn(A,B,R_1) \\
(\neg Affinity(B,C) \wedge \neg GoodResult(B,C,R_1) \wedge Doctor(E)) \rightarrow Consult(A,B,E) \\
(Result(A,C,R_3) \wedge \neg GoodResult(B,C,R_1)) \rightarrow Advise(A,B,R_3,R_1) \\
\hline
DoctorPatient(B,C) \\
Doctor(E) \\
Result(B,C,R_1) \\
Result(A,C,R_3) \\
\neg GoodInteraction(B,C) \\
\hline
\therefore \neg GoodResult(B,C,R_1) \wedge Warn(A,B) \wedge Advise(A,B,R_3,R_1) \wedge Consult(A,B,E)
\end{array}
$$

Table 1.: Exemplary ToM reasoning for affinity.

Note that some of this reasoning is inductive rather than deductive. I.e., we assume that a bad interaction is the result of a lack of affinity. But there might be other reasons of course. Here, we just show this simplified deliberation to illustrate our points. In principle, abstraction should aid agents in modeling other role-dependent high-level concepts that are relevant to the interaction context the agent is in, such as "trust" and "confidence". Collaborating agents may need to model multiple ToM states when they deal with conflicts in more complex scenarios. Especially in patient-doctor relationships with a shared past, in which several roles can be intertwined with one another, several ToM states can be beneficial to the agents in capturing even higher-level concepts like "rapport" [38,39] to use in the conflict resolution process.

5.2. Norms

In order to properly incorporate social dynamics into our abstraction-guided ToM agent framework, it is essential to understand the guiding principles behind human social behaviour. "Social norms", which are defined as commonly known standards of acceptable social behaviour, represent one of these key principles [40]. Below, we will further explore our conflict scenario in the light of social norms to illustrate how ToM agents can benefit from them.

Normally, a doctor-patient relationship is expected to be built on trust, communication, and a common understanding of both sides' needs [33]. In our case, B needs C to share all relevant information whereas C trusts B to keep this information to herself and not disclose it to others. Although adhering to these medico-social norms is expected from both parties, they may choose to not follow them. Suppose that C chooses to keep some sensitive information about himself and/or lie about his health conditions out of mistrust, shame, embarrassment, or other personal reasons. With an abstract reasoning similar to the one that we previously described in the social role example (Table 2), A can step in to examine whether a possible norm-defying behaviour is the root cause of the conflict by assessing the quality of the patient interview (or the physical examination):

$$
\begin{array}{l}
DoctorPatient(B,C) \rightarrow Doctor(B) \wedge Patient(C) \\
(DoctorPatient(B,C) \wedge Lie(C,B)) \rightarrow \neg Trust(C,B) \\
\neg Trust(C,B) \rightarrow \neg NormAdherence(C) \\
(\neg NormAdherence(C) \wedge Result(B,C,R_1)) \rightarrow \neg GoodResult(B,C,R_1) \\
(\neg NormAdherence(C) \wedge DoctorPatient(B,C)) \rightarrow Warn(A,B) \\
(Result(A,C,R_3) \wedge \neg GoodResult(B,C,R_1)) \rightarrow Advise(A,B,R_3,R_1) \\
\hline
DoctorPatient(B,C) \\
Result(B,C,R_1) \\
Result(A,C,R_3) \\
Lie(C,B) \\
\hline
\therefore \neg GoodResult(B,C,R_1) \wedge Warn(A,B) \wedge Advise(A,B,R_3,R_1)
\end{array}
$$

Table 2.: Exemplary ToM reasoning for norm adherence.

Disputes in diagnostic processes in the medical domain can also arise from different normative influences that shape doctors' decision-making behaviours. Coupled with contextual ambiguity (e.g., obscure personality traits, symptoms shared with other health problems, symptom variability through life stages etc. [41]), two doctors holding divergent theoretical orientations towards the health problem can arrive at different conclusions [42]. Back to our example, suppose that A decides that the quality of the diagnostic process is not poor but needs further analysis. A can then check whether normative influences such as opposing schools of thought play a role in its disagreement with B. For instance, A can ask B for elaboration on her deduction process (e.g., which symptoms she deemed important and why, whether she considered alternative explanations, etc.). This should further improve the quality of the collaborative discussion process. Below, we give an example that depicts how A can discuss with B about their disagreement with respect to the diagnosis by taking an argumentation approach:

- B : R_1 and R_2 suggest that we have D_1 in C's case. D_1 is a common disease and explains most of the symptoms of C.
- A : R_3 suggests that D_2 also explains C's health problem. D_2 also covers many of the observed symptoms and the MRI test results show that there is an 80 percent probability of D_2 when we see *this* specific anomaly in *this* part of the brain.
- B : I also consider D_2 but it is very uncommon in this part of the world. Also, C reported little to no fatigue which is considered an important indicator of D_2.
- A : I see in your notes that C lives a relatively healthy life (e.g., regularly jogs, eats very little junk food, lives in the countryside, etc.), so lack of fatigue might be irrelevant in our case.
- B : I find that the 20 percent probability (of not being D_2) is still very high. There is another diagnostic test that we can do to be more certain about D_2. It is an invasive procedure and may negatively affect C's health in other ways.
- A : My data shows that D_1 is a rapidly-progressing disease. I believe it is safer to go with D_1 as the running diagnosis and start the treatment immediately. Since D_2 is a slowly-progressing disease, we can take a watchful waiting approach for now.
- B : I agree. We should check with C regularly. If the treatment fails and C starts showing more symptoms of D_2, we can then do the additional test for D_2.

5.3. Values

In the social sciences, "values" denote a person's set of preferences that determine appropriate courses of action in their lives. Values tend to influence social behaviour [43] and can serve as guiding principles in people's lives [44]. Similar to roles and norms, we consider values as abstract concepts that can guide agents in interpreting behavioral patterns of humans when properly captured within abstraction mechanisms. Building on "The Theory of Basic Human Values" [45], we now extend our dispute example below by incorporating a couple of high-level concepts regarding basic human values.

Schwartz recognizes ten universal human values which can be organized in four higher-order groups [45]. One of these groups is called "self-transcendence" and characterized by benevolence, altruism, and universalism. Suppose that our patient C places high importance on self-transcendence and acts accordingly to preserve and enhance the welfare of others. For example, C may take an active part in helping less fortunate people, which requires traveling frequently to different places and interacting with a lot of people (e.g., participating in a humanitarian aid event). These endeavors of C may play a crucial part in his health condition (e.g., resulting in fatigue, stomach problems, contagious diseases). Although it is a doctor's duty to be inquisitive and eager to learn more information that can be relevant to the case, B may fail to do so. If B fails to grasp the values on which C puts high importance, she may forget or omit to ask more questions about his life. Consequently, C may not share much about his travels and endeavors because he may see them as irrelevant to his condition, resulting in a lack of important communication.

Another one of Schwartz's higher-order groups regarding basic human values is called "conservation", which is characterized by security, conformity, and tradition [45]. Now, suppose that our human doctor B places high importance on conservation, which partially affects B's behaviour in her patient-doctor relationship with C. In line with the customs that her culture provides, for example, B may choose to omit asking C sensitive personal questions and restrain herself from doing specific actions in the physical examination. B may also be generally cautious and biased towards alternative diagnostic explanations, especially if these alternatives are very uncommon. Moreover, B can be sceptical about an artificially intelligent agent's diagnostic testing capabilities, which may also hinder her relationship with A. Furthermore, A may already have formed a belief that B demonstrates conservative behaviour (e.g., A may have deduced it from past experience).

We argue that our agent doctor A can use abstraction to its benefit by computationally capturing and interpreting the values that B and C may hold as important. Using its knowledge and beliefs about B and C accordingly, A can check whether B and/or C's values have indirectly affected the disagreement between A and B. For instance, A can ask B whether she may have missed other cultural, societal, or personal factors that can play a role in C's health condition. A can also use the beliefs that it formed before (e.g., "B shows conservative behaviour") and further explain its findings and the mechanisms behind them to convince B, which may help it establish rapport with B in the long run. Below (Table 3), we draw a sketch in which A abstracts its knowledge about B and C by using epistemic logic and accordingly ask B for further elaboration on her findings ($K_A\phi$ stands for "A knows that ϕ" and $L_B\phi$ stands for "B believes that ϕ"):

$(K_A(AidEvent(F)) \wedge K_A(TakePart(C,F))) \rightarrow (K_A(Travel(C,F)) \wedge L_A(Benevlnt(C)))$
$(K_A(Fatigue(C)) \wedge K_A(Stomachache(C)) \wedge K_A(Travel(C,F))) \rightarrow L_A(Traveller(C))$
$(L_A(Benevlnt(C)) \wedge L_A(Traveller(C))) \rightarrow L_A(SelfTrns(C))$
$L_A(Conservative(B)) \rightarrow L_A(\neg GoodInteraction(B,C))$
$L_A(\neg GoodInteraction(B,C)) \rightarrow L_A \neg L_B(SelfTrns(C))$
$(L_A(SelfTrns(C)) \wedge (L_A \neg L_B(SelfTrns(C)))) \rightarrow Tell(A,B,L_A(SelfTrns(C)))$
$(L_A(SelfTrns(C)) \wedge (L_A \neg L_B(SelfTrns(C)))) \rightarrow Ask(A,B,Traveller(C))$
$(L_A(SelfTrns(C)) \wedge (L_A \neg L_B(SelfTrns(C)))) \rightarrow Ask(A,B,Benevlnt(C))$

$K_A(AidEvent(F))$
$K_A(TakePart(C,F))$
$K_A(Fatigue(C))$
$K_A(Stomachache(C))$
$L_A(Conservative(B))$

$\therefore Tell(A,B,L_A(SelfTrns(C))) \wedge Ask(A,B,Traveller(C)) \wedge Ask(A,B,Benevlnt(C))$

Table 3.: Exemplary ToM reasoning for human values.

6. Conclusion and Future Work

Computationally modeling ToM ability with the abstraction heuristics that we defined in Section 3 is a first step towards our long-term goal of designing social agents that are capable of collaborating efficiently with human partners. With examples from the medical domain, we illustrated how abstracting beliefs and knowledge into higher-level concepts can be useful for an agent doctor in dealing with conflicts that can happen when doing collective decision-making with a human doctor towards the diagnosis of a patient's health problem. By explicitly taking into account the interaction context that the agent is in, we emphasized how social dynamics shaped by roles, norms, and values can play important parts in such hybrid settings. Furthermore, we sketched several ways with various reasoning tools that the agent doctor can employ these social dynamics within the abstraction mechanism to create various context-relevant ToM states and use them to resolve conflicts efficiently, suggesting the power and versatility of the proposed abstraction framework.

As a follow-up work, we aim for a more complete abstraction model that captures the ways humans abstract their beliefs and knowledge. We will start with formalizing the entities in the abstraction framework (i.e., beliefs, abstractions, ToM states, etc.). Because we aim for an interactive reasoning system which should be well-versed in the ways of social cognition, we plan to benefit from various methods and tools in logic, artificial intelligence, and cognitive sciences (e.g., ontologies, machine learning algorithms, belief-desire-intention (BDI) models [46], etc.). Another research direction can be to further investigate the role of human-agent communication in recursive ToM reasoning. For that purpose, "mind perception theory" [47,48] can be beneficial when designing higher-order ToM agents that can accurately infer how their own artificial minds are perceived and modeled by humans. With a more comprehensive ToM agent model, which is also equipped with mind abstraction abilities, we will further test our ToM agents in human-agent settings to evaluate their collaborative skills in dynamic environments.

Acknowledgements

This research was funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, grant number 024.004.022.

References

[1] Premack D, Woodruff G. Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences. 1978;1(4):515-26.
[2] Carruthers P, Smith PK. Theories of theories of mind. Cambridge University Press; 1996.
[3] Michlmayr M. Simulation theory versus theory theory: Theories concerning the ability to read minds [Master's thesis]. Leopold-Franzens-Universität Innsbruck; 2002.
[4] de Weerd H, Verbrugge R, Verheij B. How much does it help to know what she knows you know? An agent-based simulation study. Artificial Intelligence. 2013;199-200:67-92.
[5] de Weerd H, Verbrugge R, Verheij B. Theory of Mind in the Mod Game: An Agent-Based Model of Strategic Reasoning. In: European Conference on Social Intelligence. Springer; 2014. p. 128-36.
[6] de Weerd H, Verbrugge R, Verheij B. Higher-order theory of mind in the tacit communication game. Biologically Inspired Cognitive Architectures. 2015;11:10-21.
[7] Kröhling D, Martínez E. On integrating Theory of Mind in context-aware negotiation agents. In: XX Simposio Argentino de Inteligencia Artificial (ASAI 2019)-JAIIO 48 (Salta); 2019. p. 180-93.
[8] de Weerd H, Verbrugge R, Verheij B. Agent-based models for higher-order theory of mind. In: Advances in Social Simulation, Proceedings of the 9th Conference of the European Social Simulation Association. vol. 229; 2014. p. 213-24.
[9] Osten FBVD, Kirley M, Miller T. The Minds of Many: Opponent Modeling in a Stochastic Game. In: IJCAI. AAAI Press; 2017. p. 3845-51.
[10] de Weerd H, Verbrugge R, Verheij B. Negotiating with other minds: The role of recursive theory of mind in negotiation with incomplete information. Autonomous Agents and Multi-Agent Systems. 2017;31(2):250-87.
[11] Baker CL, Saxe RR, Tenenbaum JB. Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society. 2011 01;33(33).
[12] Winfield AFT. Experiments in Artificial Theory of Mind: From Safety to Story-Telling. Frontiers in Robotics and AI. 2018;5:75.
[13] Hindriks KV, Jonker C, Tykhonov D. Towards an open negotiation architecture for heterogeneous agents. In: International Workshop on Cooperative Information Agents. Springer; 2008. p. 264-79.
[14] Sycara K, Norman TJ, Giampapa JA, Kollingbaum MJ, Burnett C, Masato D, et al. Agent support for policy-driven collaborative mission planning. The Computer Journal. 2010;53(5):528-40.
[15] Shamekhi A, Bickmore T, Lestoquoy A, Gardiner P. Augmenting group medical visits with conversational agents for stress management behavior change. In: International Conference on Persuasive Technology. Springer; 2017. p. 55-67.
[16] Hiatt LM, Harrison AM, Trafton JG. Accommodating human variability in human-robot teams through theory of mind. In: Twenty-Second International Joint Conference on Artificial Intelligence; 2011. .
[17] Anderson JR. How can the human mind occur in the physical universe? Oxford University Press; 2009.
[18] Devin S, Alami R. An implemented theory of mind to improve human-robot shared plans execution. In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE; 2016. p. 319-26.
[19] Görür OC, Rosman BS, Hoffman G, Şahin Albayrak. Toward integrating Theory of Mind into adaptive decision-making of social robots to understand human intention. In: International Conference on Human-Robot Interaction. Workshop on the Role of Intentions in Human-Robot Interaction; 2017. .
[20] Buehler MC, Weisswange TH. Theory of mind based communication for human agent cooperation. In: 2020 IEEE International Conference on Human-Machine Systems (ICHMS). IEEE; 2020. p. 1-6.
[21] Lim TX, Tio S, Ong DC. Improving multi-agent cooperation using theory of mind. arXiv preprint arXiv:200715703. 2020.

[22] Baker CL, Jara-Ettinger J, Saxe R, Tenenbaum JB. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. Nature Human Behaviour. 2017;1(4):1-10.
[23] Akata Z, Balliet D, De Rijke M, Dignum F, Dignum V, Eiben G, et al. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer. 2020;53(08):18-28.
[24] Shademan A, Decker RS, Opfermann JD, Leonard S, Krieger A, Kim PC. Supervised autonomous robotic soft tissue surgery. Science translational medicine. 2016;8(337):337ra64-4.
[25] Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Ophthalmology. 2017;124(7):962-9.
[26] National Academies of Sciences, and Engineering, and Medicine and others. The Diagnostic Process. In: Balogh EP, Miller BT, Ball JR, editors. Improving Diagnosis in Health Care. National Academies Press; 2015. p. 31-80.
[27] Hazlett HC, Gu H, Munsell BC, Kim SH, Styner M, Wolff JJ, et al. Early brain development in infants at high risk for autism spectrum disorder. Nature. 2017;542(7641):348-51.
[28] Loh E. Medicine and the rise of the robots: a qualitative review of recent advances of artificial intelligence in health. BMJ leader. 2018:leader-2018.
[29] Briganti G, Le Moine O. Artificial intelligence in medicine: today and tomorrow. Frontiers in medicine. 2020;7:27.
[30] Tversky A, Kahneman D. Judgment under Uncertainty: Heuristics and Biases. Science. 1974;185(4157):1124-31.
[31] Lexico. Respect; (n.d.). In Lexico.com dictionary. Available from: https://www.lexico.com/definition/respect.
[32] Dastani M, Dignum V, Dignum F. Role-assignment in open agent societies. In: Proceedings of the second international joint conference on Autonomous agents and multiagent systems; 2003. p. 489-96.
[33] National Institutes of Health (U S). Talking with Your Older Patient: A Clinician's Handbook. NIH publication. Department of Health & Human Services, NIH, National Institute on Aging; 2016. Available from: https://books.google.nl/books?id=YyhCvgAACAAJ.
[34] Lazarsfeld PF, Merton RK. Friendship as a social process: A substantive and methodological analysis. Freedom and Control in Modern Society. 1954;18(1):18-66.
[35] Suls J, Martin R, Wheeler L. Social comparison: Why, with whom, and with what effect? Current Directions in Psychological Science. 2002;11(5):159-63.
[36] Aronson E, Akert RM, Wilson TD. Social Psychology. 7th ed. Upper Saddle River, Prentice Hall; 2010.
[37] Bell RA, Daly JA. The affinity-seeking function of communication. Communications Monographs. 1984;51(2):91-115.
[38] Tickle-Degnen L, Rosenthal R. The nature of rapport and its nonverbal correlates. Psychological inquiry. 1990;1(4):285-93.
[39] DiMatteo MR. A Social-Psychological Analysis of Physician-Patient Rapport: Toward a Science of the Art of Medicine. Journal of Social Issues. 1979 Jan;35(1):12-33.
[40] Dignum F. Autonomous agents with norms. Artificial intelligence and law. 1999;7(1):69-79.
[41] CHADD. Conflicting Diagnoses? Take the Time to Get It Right; (n.d.). Available from: https://www.chadd.org/adhd-weekly/conflicting-diagnoses-take-the-time-to-get-it-right/.
[42] Gerard NM. A diagnosis of conflict: theoretical barriers to integration in mental health services & their philosophical undercurrents. Philosophy, Ethics, and Humanities in Medicine. 2010;5(1):1-8.
[43] Bardi A, Schwartz SH. Values and behavior: Strength and structure of relations. Personality and social psychology bulletin. 2003;29(10):1207-20.
[44] Schwartz SH, Cieciuch J, Vecchione M, Davidov E, Fischer R, Beierlein C, et al. Refining the theory of basic individual values. Journal of personality and social psychology. 2012;103(4):663.
[45] Schwartz SH. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In: Advances in experimental social psychology. vol. 25. Elsevier; 1992. p. 1-65.
[46] Rao AS, Georgeff MP. Decision procedures for BDI logics. Journal of Logic and Computation. 1998;8(3):293-343.
[47] Gray HM, Gray K, Wegner DM. Dimensions of mind perception. Science. 2007;315(5812):619-9.
[48] Lee M, Lucas G, Gratch J. Comparing mind perception in strategic exchanges: Human-agent negotiation, dictator and ultimatum games. Journal on Multimodal User Interfaces. 2021 06;15(2):201-14.