Computational Theory of Mind for Human-Agent Coordination (Full)

Emre Erdogan[1], Frank Dignum[1,2], Rineke Verbrugge[3], and Pınar Yolum[1]

[1] Utrecht University, Utrecht, Netherlands
{e.erdogan1,p.yolum}@uu.nl
[2] Umeå University, Umeå, Sweden
dignum@cs.umu.se
[3] University of Groningen, Groningen, Netherlands
l.c.verbrugge@rug.nl

Abstract. In everyday life, people often depend on their theory of mind, i.e., their ability to reason about unobservable mental content of others to understand, explain, and predict their behaviour. Many agent-based models have been designed to develop computational theory of mind and analyze its effectiveness in various tasks and settings. However, most existing models are not generic (e.g., only applied in a given setting), not feasible (e.g., require too much information to be processed), or not human-inspired (e.g., do not capture the behavioral heuristics of humans). This hinders their applicability in many settings. Accordingly, we propose a new computational theory of mind, which captures the human decision heuristics of reasoning by abstracting individual beliefs about others. We specifically study *computational affinity* and show how it can be used in tandem with theory of mind reasoning when designing agent models for human-agent negotiation. We perform two-agent simulations to analyze the role of affinity in getting to agreements when there is a bound on the time to be spent for negotiating. Our results suggest that modeling affinity can ease the negotiation process by decreasing the number of rounds needed for an agreement as well as yield a higher benefit for agents with theory of mind reasoning.

Keywords: social cognition · communication · affinity · abstraction · heuristics · negotiation · human-inspired computational model.

1 Introduction

Theory of Mind (ToM) is the ability of reasoning about the mental content of other people, such as their beliefs and desires, making it possible to understand and predict their behaviour [26, 9, 24]. Being an important part of social cognition, the capability of ToM develops early in life and bestows on humans a plethora of social skills such as negotiating, teaching, and tricking. Recursively employing ToM provides a direct path to reason about how others use ToM, which is widely known as "higher-order ToM" (e.g., "I believe that Alice does

not know that Bob is planning a baby shower for her"), and is particularly helpful for adapting to the complex dynamics of social life.

Agent-based computational models have previously been used to analyze the effectiveness of ToM in competitive [11, 13] and cooperative [14] games and mixed-motive situations [20, 12, 15] in which the models are based on recursive reasoning and behaviourally limited by the complexity and rules of the games. Baker *et al.* [5] model ToM within a Bayesian framework using partially observable Markov decision processes and test its performance in a simple spatial setting. Osten *et al.* [25] propose a multiagent ToM model that extends the model described in [11] and evaluates its performance in a multiplayer stochastic game. Winfield [34] shows how robots can use a ToM model in improving their safety, making decisions based on simple ethical rules, and imitating other robots' goals. In most of the studies around computational ToM models, the results are generally promising and demonstrate that the use of ToM leads to better outcomes for the studied tasks. Still, the existing models have not been widely adopted as a computational tool in many real-life settings. We argue that for a ToM model to be applicable, it needs to adhere to the following criteria:

Generic: Most of the existing models (see [25, 31, 11–15]) are built for a specific game-theoretic setting in mind. The models thus are based on the rules of the game as well as interpreted semantics of the strategies. This creates a limitation because it is not straightforward to use these models outside of these settings. We argue that real-life social interaction is generally more complex and for a more comprehensive model of ToM, agents should take into account a variety of both context-dependent and context-independent information such as traits, as well as social frames of reference such as roles, norms, and values [29, 6]. Ideally, a computational ToM model should be *generic*; i.e., independent of the particular setting to which it is applied so that it can be used in a variety of settings.

Feasible: In general terms, ToM is about beliefs and knowledge an agent has or can derive about the mental attitudes of other agents. Without a proper control, the number of elements in an agent's belief and knowledge set can increase rapidly over time. This has two immediate disadvantages. First, it will not be clear to the agent which beliefs about the other agents would be useful to consider in a given context, leading to complex decision processes. Second, the volume of information will make it more difficult for the agent to make fast and accurate inferences about others. On the other hand, the agent can benefit from a control mechanism which can sort out the relevant and important information according to the context that the agent operates in. Thus, for a more efficient computational model, it is necessary to ensure that the agent can abstract from existing information to yield *feasible* computation of ToM.

Human-Inspired: In various social contexts, humans are known to rely on social skills that are based on more automatic and fast-working heuristics and require less conscious effort, such as repetition (i.e., repeating behaviours that yield desirable results), imitation (i.e., mimic others) [19], and stereotypes [16]. These agile mechanisms can be especially helpful for humans in social interactions where the time spent on reasoning and/or the cognitive resources allocated

are vital concerns. For an agent to better explain the behaviour of humans that it interacts with, its ToM should be *human-inspired*, such that it should be able to capture and interpret the heuristics that humans use in every day dealings.

An important area where ToM could be of particular use is hybrid intelligence [1], where an agent can coordinate 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 wearable physical activity monitor agent à la Fitbit that works with a human to ensure that the human establishes a healthy life. Typical interactions with such devices take the form of information passing, such as that the device periodically informs the human what more she has to do (e.g., "take another 200 steps"), milestones she has achieved (e.g., "you received a Tiger badge"); or it requests information (e.g., "enter the foods consumed today"). Take the first type of interaction. This necessitates the human to take an action that is not easy to do and thus requires nudging from the agent. Ideally, if the agent could have a ToM for the human, it could create strategies as to how to proceed with such requests. The long-term goal of our research is thus to design and develop a generic, feasible, and human-inspired ToM that could be applicable in such settings to improve human-agent coordination and thus to facilitate hybrid intelligence.

As an initial step towards this goal, we develop an abstraction framework for ToM over which we construct an abstraction heuristic. The underlying idea is to employ an agent's belief and knowledge set to produce a more abstract, complex *interaction state* that can be readily used by the agent. To investigate the principle of abstraction we use the concrete concept of *affinity* that summarizes how we relate to someone based on many things we know about that person and our history of interactions. Computational affinity captures how humans use affinity in their interactions and can be used in tandem with ToM reasoning when designing agent models. To demonstrate its usage and power in human-agent interaction, we employ it in two-agent negotiation. Our results show that capturing affinity improves agent-agent coordination and agents who perform ToM reasoning obtain outcomes that are better than agents who do not.

The rest of this paper is organized as follows. Section 2 describes abstraction heuristics and computational affinity. Section 3 explains our framework and how we integrate ToM with affinity in negotiation. Section 4 evaluates our proposed model over two-agent simulations. Section 5 discusses our results, addresses related research in the literature, and points to future research directions.

2 Abstraction Heuristics and Affinity

Humans are known to use behavioural simplification mechanisms in their decision-making processes (e.g., stereotypes, biases) [33]. Inspired by this idea, we envision an abstraction-guided ToM agent paradigm that simplifies its beliefs and knowledge into compact representations that can serve for heuristics. Computationally, what we call an "abstraction mechanism" is an agent apparatus that does the following (Figure 1):

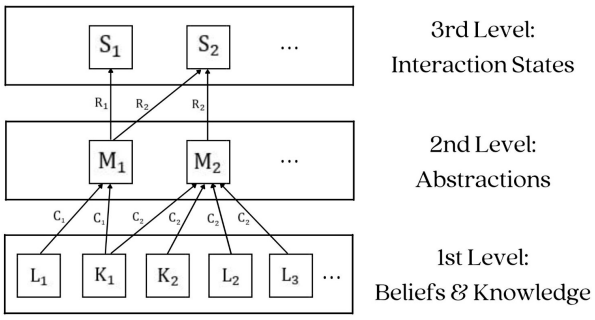


Fig. 1: Abstraction procedure: Individual beliefs (L_i) and knowledge (K_j) are used to create abstractions M_k that are then used in interactions S_n.

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

We claim that such an abstraction procedure embedded in a ToM agent should produce interaction states that can be used in a variety of settings, are simple enough to easily mesh with the agent's decision-making processes, and capture and interpret the related human behaviour. Figure 1 shows our layered approach to such an abstraction mechanism. The first layer holds the set of beliefs and knowledge about others that could come from different sources, such as observations or explicitly stated information from others. While the agent can keep this set, it does not operate at that level but instead 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 change. The abstractions in the second level influence how the agents operate in the third level. One can think of the third level as pertaining to the application in question. Figure 1 also shows that beliefs and knowledge can have multiple characteristics C_k, C_l, etc. which guide the production of the corresponding abstractions M_k, M_l etc.; multiple abstractions can be used to produce an interaction 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. Here, we do not intend to provide a full-fledged abstraction model that addresses and gives possible solutions to all kinds of challenges a ToM agent may face during its lifetime. We will discuss some important points that can help us further develop our abstraction mechanism in Section 5.

2.1 Computational Affinity

We propose that in principle, this abstraction approach can be used with complex human notions. We demonstrate our intuition in a specific type of abstraction mechanism, which captures *affinity*: "a feeling of closeness and understanding that someone has for another person because of their similar qualities, ideas, or interests" [23]. People are inclined to get along with and gravitate to others that are similar to them [21, 32]. One practical outcome of this feeling is generating generous behaviour: People tend to do favors for others they like [2]. We claim that affinity can be captured within an abstraction mechanism in which one can merge many beliefs and pieces of knowledge (e.g., "I believe that he leads a healthy life *like me*.") into a more abstract belief that shares the same characteristic ("I believe that he is *very similar to me*.") and then to an interaction state ("I feel a strong affinity towards him because I believe that he is very similar to me.") which can be more effectively used within a rule set when making decisions ("I feel a strong affinity towards him. I can do small favors to people that I feel strong affinity towards. Thus, I will do a small favor for him.").

Observing a similarity is essential for affinity [7]. In our computational framework, we limit similarity to interacting agents having the same opinions on a subject and use opinions as comparable tokens that are Boolean in nature (e.g., healthy living is important: yes/no). Moreover, we limit observation to communication, meaning that opinions are private and unobservable unless the agent shares them with another agent. Following this intuition, we provide three different definitions of computational affinity that pertain to how it is brought about. Note that the aim is not to come up with the most precise definition but with reasonable, alternative definitions that an agent might adopt.

All definitions are based on the agents exchanging opinions. Thus, we consider each agent A to have a set of fixed opinions on various subjects.

Definition 1. *For an agent A to have a **type-1 affinity** towards another agent B, at least one of the opinions B tells A must match with that of A.*

The most important aspect of this form of affinity is that it is static, meaning that after A establishes affinity towards B, even if B later tells its opinion on another subject that does not match with that of A, A does not lose its affinity. However, in real life, affinity is not always static; thus, we define another type of affinity to capture its dynamic nature:

Definition 2. *For an agent A to have a **type-2 affinity** towards another agent B, the most recent opinion B tells A must match with that of A.*

Still, affinity does not have to depend only on the latest matching opinion. For example, agents can do multiple comparisons before establishing affinity. Thus, we give another, more concrete way to define affinity computationally:

Definition 3. *For an agent A to have a **type-3 affinity** towards another agent B, the majority of opinions B tells A must match with those of A (i.e., the number of matching opinions is bigger than zero and not smaller than the number of non-matching opinions).*

Note that the abstraction mechanism that the agents employ is relatively simple: Comparing just one pair of opinions is enough to produce (or change) the abstract belief of similarity which agents further employ to decide whether to establish affinity or not. Even with this simple mechanism, we observe that computational affinity as an abstract entity that has a life-cycle: it is born, lives, and dies (and can be brought from the dead again). It can be active or passive, subject to the situation the agent is in. Plus, it holds basic information about the relationship between agents. Depending on the communication history of the agents, it can be reciprocal or not since both agents should tell each other their opinions for both of them to have affinity towards the other. Other features that we have not incorporated here can include the duration (e.g., how long it affects the agent's decision) and strength (e.g., how strongly it affects the agent's decision). Next we demonstrate how even a simple abstraction mechanism as described can be useful in human-agent interactions.

2.2 Computational Affinity and ToM

In its core, we observe that a person who has affinity towards another person can act in ways that would be helpful to the second person. This can mean different things depending on the context; here we define it in a two-agent setting and as generic as possible. In simplest terms, an agent A that has an affinity towards another agent B can do *a thing* that is more favorable to B than *the thing* A normally does when it does not have affinity towards B. For example, in a negotiation, a seller A with an affinity towards a buyer B can make an offer that is more favorable for B than the offer A makes when it does not have affinity.

In addition, we want our agents to not just establish affinity but also attribute it as a mental state to others, as people do. Essentially, we also design agents that have ToM about other agents and reason whether another agent has an affinity towards a certain agent or not (e.g., "I believe B has an affinity towards me"). The reasoning mechanism shall rely on basic perspective-taking and the condition that both agents share their opinions (remember that affinity is not inherently mutual). Later on, we will explain how such an agent with ToM can also use this affinity attribution mechanism to its benefit.

In this body of work, we call the ToM agents that can have type-x affinity towards others "type-x affinity agents with 1^{st} order ToM" or shortly "A_1^x agents". Similarly, we denote the agents that do not have ToM as "type-x affinity agents with 0^{th} order ToM" or shortly "A_0^x agents".

3 Negotiation with Computational Affinity and ToM

Now, we discuss how our proposed model can be used for human-agent negotiation. We return to our example in Section 1 where a wearable physical activity monitor agent is working with a human to increase the number of steps the human takes. As the underlying mechanism, we choose two-agent negotiation,

because it is a robust mixed-motive setting that also provides a good context for exploring behavioural capabilities of ToM agents.

To make our setup concrete, we define an **agent** as an autonomous entity which can either be an **activity monitor agent** or a **human agent**, where the former is working to increase the number of steps taken by the human while the human is reluctant to walk. To achieve its goal, each agent can **make an offer** or **accept an offer** made by the other agent. Furthermore, an agent has fixed **opinions** on various subjects (e.g., healthy living is important: yes/no). It can **tell** the other agent its opinions, including those about the negotiation outcomes themselves, and **compare** a told opinion with its own opinion (same subject).

3.1 Negotiation Framework

The subject matter of negotiation is agreeing on the number of steps to be taken. The negotiation protocol can be briefly described as alternating monotonic concession with communication, a variant of the monotonic concession protocol [30]. Basically, it is a rule set for two agents to **negotiate** and **communicate** in alternating rounds. An agent can both negotiate and communicate with the other agent in the same round in which it can either make a new offer or accept the latest offer made by the other agent (negotiation part) and can tell the other agent its opinions (communication part). Furthermore, negotiations should be done in the form of **monotonic concession**: No agent can make an offer that is less preferred by the other agent than an earlier offer that it made. Lastly, a negotiation **ends** when an agent accepts the latest offer made or a fixed number of rounds pass without an agreement (e.g., 10 total rounds).

Negotiating agents' offers and counter-offers are generally governed by their strategies: a prepared plan of action to achieve a goal under conditions of uncertainty. The negotiation literature is rich with sophisticated strategies [27] [4]. In order to focus only on the effects of computational affinity and ToM, we opt for a simple strategy for agents such that each agent makes an offer and adjusts the number of steps with a constant, predetermined value until it goes beyond the **reserve value** (or **reserve price**). For example, a human agent starts the offer at 5000 and increases it with 100 every round until it goes beyond 5500[1]. We call this value of 100 the **unit increment/decrement value** of agents and make all agents use this strategy as the baseline strategy when making offers.

3.2 Negotiating with Affinity and ToM

Agent A that has affinity towards another agent B can give an offer that can be more favorable for B than the offer A gives when it does not have affinity towards B, as we have stated earlier. More specifically, we utilize computational affinity as a regulator for unit increment/decrement values that agents use when making offers. As a design decision, we make reserve values not affected by affinity in our framework. Here, we give an example.

[1] In this case, for example the activity monitor agent could start with an offer of 5700 and decrease it with 100 until it goes beyond 5300.

Table 1: Four negotiation scenarios Sc_1, Sc_2, Sc_3 and Sc_4 are given (Example 1).

(a) Opinions do not match in Sc_1, but match in Sc_2. Only A tells its opinion.

Sc_1			Sc_2		
R	A	B	**R**	A	B
1	1500		1	1500	
1	O_{yes}		1	O_{yes}	
2		2100	2		2100
3	1600		3	1600	
4		2000	4		1990
5	1700		5	1700	
6		1900	6		1880
7	1800		7	1800	
8		Accepts	8		Accepts

(b) Opinions match and A (resp. B) starts in Sc_3 (resp. Sc_4). Both tell opinions.

Sc_3			Sc_4		
R	A	B	**R**	A	B
1	1500		1		2100
1	O_{yes}		1		O_{yes}
2		2100	2	1500	
2		O_{yes}	2	O_{yes}	
3	1610		3		1990
4		1990	4	1610	
5	1720		5		1880
6		1880	6	1720	
7	1830		7		1770
8		Accepts	8	Accepts	

Example 1. A human agent A of type A_0^1 and an activity monitor agent B of type A_0^1 are negotiating. The reserve values of A and B are set to be 1850 and 1750, respectively. Their respective unit increment/decrement values are both 100 and affinity increases it with 10. Each agent has a Boolean opinion on the same subject O: It can be either O_{yes} or O_{no}. In Tables 1a and 1b, we give four different scenarios (Sc_1, Sc_2, Sc_3, and Sc_4).

Example 1 depicts two crucial points. First, affinity does not always produce a different result (e.g., Sc_1 and Sc_2) and second, either agent can benefit from the result when affinity is reciprocal (e.g., Sc_3 and Sc_4), since the final situation depends on other factors as well (e.g., the reserve values, the starting agent). Additionally, one can see that although reciprocal affinity introduces variance in the agreements (e.g., Sc_3 and Sc_4 in which the accepted offers are 1770 and 1830, respectively), it stays the same on average (e.g., 1800) due to the symmetry in the provided benefits for both agents.

In the previous section, we have noted that an A_1^x agent can use its ToM ability to its benefit when making offers. In particular, when an A_1^x agent concludes that there is a mutual affinity, it can change its unit increment/decrement value so that its offer *adjustments* (not offers themselves) are not as generous as its opponent's adjustments. For example, if an A_1^x activity monitor agent decides that there is a mutual affinity and observes that its opponent's current increment value (i.e., the difference between the latest two offers of the opponent) is 110, it can change its own to a value lower than 110, say 105. With this improvement, it is guaranteed that a reciprocal affinity between an A_1^x and an A_0^y will result in an offer that A_1^x prefers more than A_0^y. Here, we give an illustrating example.

Example 2. A human agent A of type A_1^1 and an activity monitor agent B of type A_0^1 are negotiating. The reserve values of A and B are set to be 1850 and

Table 2: Two negotiation scenarios Sc_5 and Sc_6 are given for A and B. Both tell their opinions in both scenarios (Example 2).

Sc_5: Opinions match, A starts.			Sc_6: Opinions match, B starts.		
R	A	B	**R**	A	B
1	1500, O_{no}		1		2100, O_{no}
2		2100, O_{no}	2	1500, O_{no}	
3	1610		3		1990
4		1990	4	1605	
5	1715		5		1880
6		1880	6	1710	
7	1820		7		1770
8		Accepts	8	Accepts	

1750, respectively. Their respective unit increment/decrement values are both 100. Affinity increases it by 10 but mutual affinity increases it only by 5. Each agent has a Boolean opinion on the same subject O: It can be either O_{yes} or O_{no}. In Table 2, two different scenarios (Sc_5 and Sc_6) are given.

One can notice in Example 2 that A_1^x is designed to limit its own affinity-induced generousness using ToM. The superiority of A_1^x over A_0^x can be seen in the newly introduced asymmetrical variance in the agreements (e.g., Sc_5 and Sc_6 in which the accepted offers are 1770 and 1820, respectively) and the new average (e.g., $1795 < 1800$), benefiting A_1^x agent A more than A_0^x agent B.

4 Experiments and Results

We are interested in understanding the role of affinity in getting to agreements when there is a bound on the time spent for negotiating. To answer this general question in detail, we have created an experimental setup with four simulation experiments. We configure our negotiation framework (including the reserve values, starting offers, and unit increment and decrement values) so that an agreement can be achieved in a maximum of 12 rounds, even without affinity. In all simulations, activity monitor agents' and human agents' starting offers are set to 2000 and 1000 and reserve values are 1450 and 1550, respectively. Unit increment and decrement values are both set to 100 at the beginning and it is common knowledge that agents do not decrease these values below 100 (agents can increase them in case of affinity). The worst offer an agent can make for itself is with its reserve value. A negotiation begins with two newly created agents, namely, an activity monitor agent and a human agent, where every opinion of agents is created randomly: it can be a "yes" or "no" with the same probability. One of the agents is randomly chosen to start the process and the other agent continues accordingly. In the first two rounds, each agent gives its starting offer.

There are two additional restrictions in the protocol we use. First, an agent tells all of its opinions in the negotiation process. Second, opinions are told in a

pre-arranged order (i.e., subject 1, subject 2, subject 3...) where an agent tells only one opinion per round in a conversational flow. This is because we intend to keep the communication as simple as possible and do not want to analyze how different communication patterns affect the life-cycle of affinity. We also want to ensure that affinity can be formed reciprocally in the negotiations.

Every round, an A_0^x agent first checks if the latest offer is acceptable. If yes, it accepts and ends the negotiation. If not, it compares the shared opinion(s) to check whether affinity ensues or not, following the criteria of its affinity definition. If it does not establish affinity, it gives an offer that is 100 higher (resp. lower) than its previous offer, if it is a human (resp. activity monitor) agent. On the other hand, if the agent establishes affinity, it changes 100 to a multiple of 5 between 100 and 150 (including the boundaries) and makes an offer accordingly. Then, it ends its turn by telling one of its opinions according to the sharing order until all are shared. We introduce this randomness into A_0^x agent's offer-making mechanism to make it more dynamic. It is worth to note that this can also indirectly change the agent that gives the final offer.

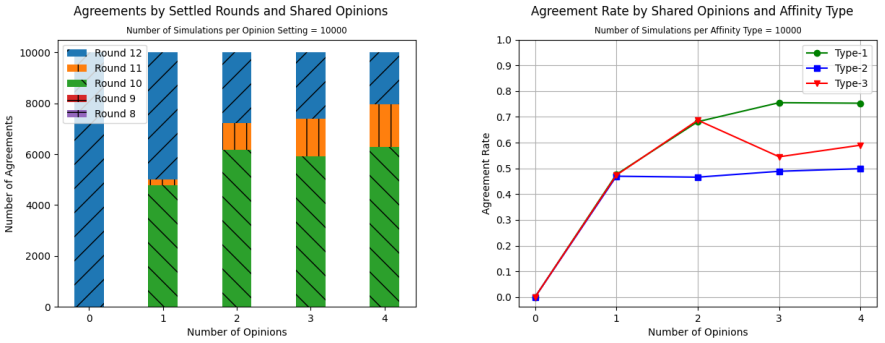
Every round, an A_1^x agent also checks if the latest offer is acceptable. If yes, it accepts and ends the negotiation. If not, it compares the shared opinion(s) to check whether affinity ensues or not, following the criteria of its affinity definition. Additionally, it also decides whether the other agent has established affinity or not. If the A_1^x agent does not establish affinity or decides that its opponent does not have affinity, it gives an offer that is 100 higher (resp. lower) than its previous offer, if it is a human (resp. activity monitor) agent, like A_0^x agents. Otherwise, it changes 100 to a multiple of 5 between 100 and X (including X) and makes an offer accordingly, where X is equal to the difference between the latest two offers of its opponent (i.e., the opponent's currently observed unit increment/decrement value). It ends its turn by telling one of its opinions according to the sharing order until all are shared. Again, we introduce this opponent-dependent randomness into the offer-making mechanism of an A_1^x agent to make it more dynamic and limit the agent's own affinity-induced generousness.

There are four different experimental variations in which we use only A_m^1 (V1), only A_m^2 (V2), only A_m^3 (V3), and all types of agents (V4), where $m \in \{0, 1\}$ unless told otherwise. Every experimental variation consists of four different opinion settings: In the n-opinion setting, every agent has n opinion(s) on the same n subjects, where $n \in \{1, 2, 3, 4\}$. Per setting, we perform simulations with $10,000$ different agent pairs where every agent negotiates once.

4.1 The Effect of Affinity on Agreements

In the first experiment, our aim is to find how affinity affects the number of agreements made when A_0^1, A_0^2, and A_0^3 agents negotiate with each other (V4). An agent is created as an A_0^1, A_0^2, or A_0^3 agent with the same probability. We limit the maximum number of rounds of negotiation to 12. Through the simulation, we also keep track of the final rounds in which agreements are settled.

The stacked bars in Figure 2a show the number of successful negotiations that are done by A_0^x in the simulation. All different opinion settings are given



(a) Affinity conceives early agreements. (b) Agreements and affinity types (all A_0^x).

Fig. 2: Affinity helps coordination.

in the x-axis (i.e., 0-4), while the y-axis shows the total number of achievable agreements; colors and hatches together represent the final round information of the agreements (i.e., 8-12).

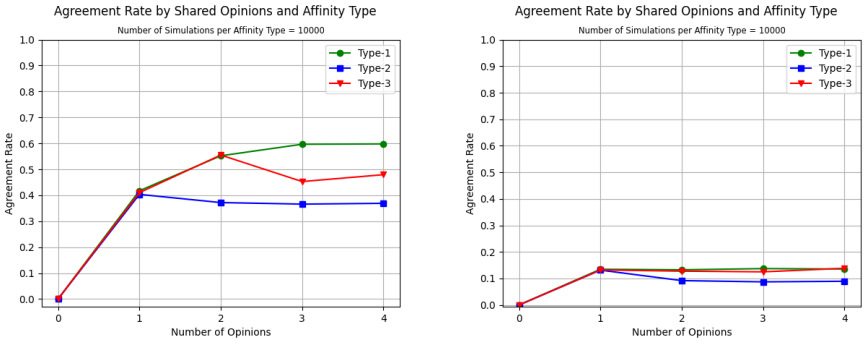
When no opinion is shared, all 12 rounds are necessary for reaching an agreement in all simulations. However, even sharing one opinion makes a big difference. We can see in Figure 2a that nearly half of the agreements are done in 10 rounds in the 1-opinion setting. Other settings also show similar results: The number of agreements that need 12 rounds decreases when the number of shared opinions increases. Hence, we can conclude that when A_0^x agents negotiate, the number of agreements that are settled on earlier than 12 rounds increases with the number of shared opinions. This shows that by modeling affinity explicitly, the agents can reduce the number of interactions needed to agree.

4.2 Affinity Types and Agreement Rates

In the second experiment, our aim is to find how affinity type and number of shared opinions affect the number of agreements made when A_0^x agents negotiate. The experiment consists of the first three variations V1, V2, and V3. We limit the maximum number of rounds of negotiation to 10 to get a better understanding of how different affinity types get to early agreements.

The line plots in Figure 2b show the percentage of successful negotiations that are achieved by A_0^x in 10 rounds over all negotiations per affinity type.

When no opinion is shared, the number of agreements that can be achieved in 10 rounds is zero. Figure 2b shows that for the experiment's V1 variation with 1-opinion setting, we can see that agents sharing just one opinion makes a significant difference in the number of agreements. When A_0^1 agents negotiate, nearly 0.50 of all simulations end with an agreement. The number increases to 0.68 and 0.75 for 2-opinion and 3-opinion settings. This increase can be explained by the fact that when the agents exchange more opinions, the probability of finding a negotiating agent pair that has at least one matching opinion increases.



(a) A_0^x vs. A_1^x. (b) All A_1^x.

Fig. 3: Agreement rates depend on both ToM and affinity type.

In the 4-opinion setting, however, it does not go higher since we set hard limits on the unit increment/decrement values and also due to the overall randomness in the agent creation and offer-making procedures. Thus, it shows that Type-1 affinity affects the agents in such a way that the number of agreements made increases more slowly when the number of rounds is fixed.

When A_0^2 agents negotiate, we see a different pattern. For every opinion setting of the experiment's V2 variation, nearly 0.50 of all simulations end with an agreement. This is mainly because Type-2 affinity is not static like Type-1 affinity and every agent can lose its affinity during the opinion comparison process. Thus, Type-2 affinity affects and changes the average unit increment/decrement value that an agent uses before reaching an agreement, but not as much as Type-1 affinity. On the other hand, A_0^3 agents generate a different pattern that is a mixture of the previous ones. Excluding the 0-opinion setting, the agreement rate in the experiment's V3 variation is on average greater than 0.5 but not as much as the average we see in V1. Hence, we can say that Type-2 and Type-3 affinity types do not create agreements as much as Type-1 affinity.

4.3 Roles of ToM and Affinity in Agreements

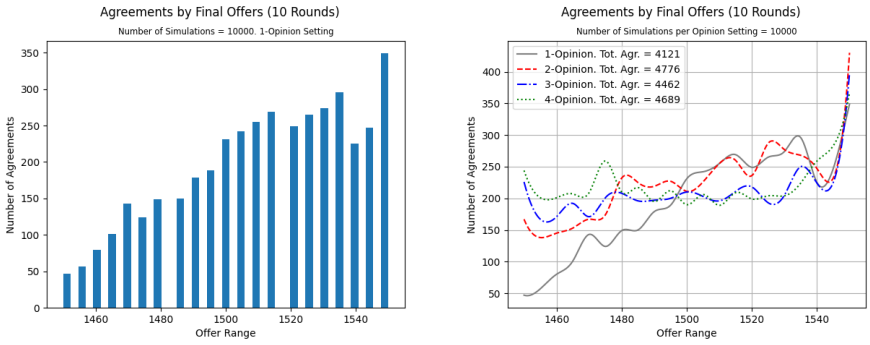
In the third (resp. fourth) experiment, our aim is to find how ToM reasoning and affinity together affect the number of agreements made when human agents of type A_0^x (resp. A_1^x) and activity monitor agents of type A_1^y negotiate. Both experiments consist of variations V1, V2, and V3, similar to the second experiment. The maximum number of rounds is set to 10.

The line plots in Figure 3a (resp. Figure 3b) show the percentage of successful negotiations that are achieved by A_0^x (resp. A_1^x) human agents and A_1^y activity monitor agents in 10 rounds over all negotiations per affinity type.

Comparing with Figure 2b, Figure 3a shows a general decrease in the agreement rates by shared opinions and affinity types. For example, when A_0^x human agents negotiate with A_1^y activity monitor agents, nearly 0.40 of all simulations

end with an agreement in the 1-opinion setting, instead of 0.50. This number increases up to 0.60 for the 4-opinion setting which is lower than the corresponding agreement rate given in Figure 2b (0.76). The drop in the agreement rates is drastic when A_1^x human agents negotiate with A_1^x activity monitor agents, as plotted in Figure 3b. This is on par with what we have expected from the negotiating behaviour of ToM agents since it is affected by opponents' offer-making behaviour as well: ToM can have a relatively negative effect in the number of agreements when the number of rounds is fixed.

We have done additional simulations to provide more depth to the negotiations in which ToM agents negotiate with agents that do not have ToM. In Figure 4a, we analyze A_0^x-A_1^y negotiations where all affinity types are used and only one opinion is shared. The x-axis shows the number of agreements done in 10 rounds and the y-axis shows the agreeable offer range (1450-1550). We can see more agreements on the right side of the figure (> 1500) than the left side (< 1500), implying that A_1^y activity monitor agents end up with offers that are on average better for them (the average offer is approximately equal to 1512). In Figure 4b, we analyze how an increase in the number of shared opinions changes this asymmetrical benefit. Every line plot shows how number of agreements correlates with the final offers in a specific opinion setting. We can see that the A_0^x human agents in many-opinion settings end up with better offers on average than the A_0^x human agents in few-opinion settings (still not better than their A_1^y monitor agent counterparts). It shows that when more opinions are shared, the superiority of ToM agents over non-ToM agents decreases in negotiations where we explicitly model affinity. This emergent phenomenon reminds us that it is not so easy to develop and maintain affinity with sheer communication (i.e., it also needs a strategy) and it is even harder to benefit from it (i.e., ToM's advantage diminishes).



(a) A_1^y benefit more than A_0^x. (b) ToM's benefit decreases with opinions.

Fig. 4: ToM with affinity benefits agents.

5 Discussion and Future Work

Within our computational ToM framework, founded on the abstraction mechanism defined in Section 2, we propose a human-inspired heuristic called computational affinity for agents to improve coordination in hybrid interactions. We use agents to simulate a human-agent negotiation in the context of activity monitoring. Our findings demonstrate that explicitly modeling affinity can ease the agreement process. We show how sharing more information can also help the activity monitoring agents forge more agreements, albeit depending on the agent's affinity type. Our results indicate that when negotiating with human agents that do not have a ToM, activity monitoring agents that have a ToM end up with agreements that is more favorable to them than to their opponents. Although the communication part of negotiations needs further analysis and strategies on its own [27], the results provide the motivation to develop more sophisticated ToM agents that can generate affinity and benefit from it, and test them in real-life negotiations to see if and how they can improve human-agent coordination.

Research on computational ToM models suggests that ToM reasoning benefits agents in different ways and even more in the higher orders. De Weerd *et al.* [11] show that agents benefit from higher-order ToM reasoning in competitive game-theoretic settings, although with diminishing returns beyond third-order ToM. Further, they investigate how higher-order ToM can be beneficial for agents in a strictly cooperative game [14] and show that communication can be set up more quickly when agents beyond zero-order ToM play the game. De Weerd *et al.* [15] determine to what extent agents benefit from higher-order ToM reasoning in a mixed-motive situation called the "Colored Trails". The results indicate that there is a considerable benefit in using second-order ToM; however, first-order ToM has a limited effectiveness. Kröhling and Martínez [20] investigate the role of ToM in single-issue negotiations between "context-aware" agents where the negotiation context is modeled by two variables, summarized as necessity and risk. Görür *et al.* [17] propose a ToM agent model for estimating humans' intentions in a shared human-robot task. Brooks and Szafir [8] show how robots can create second-order ToM models by using humans' actions in spatial settings.

Observing and communicating are crucial components of human social behaviour. Our long-term goal is to design socially intelligent agents that can understand how humans "tick" and work with them in synergy. Computationally modeling ToM ability with the abstraction heuristics that we defined in Section 2 is a first step toward this goal. Unlike the studies we mention above, we design our human-inspired abstraction procedure to be as generic as possible and generate interaction states which emulate how humans develop and maintain the mental states they experience through their lives. The procedure also provides a useful simplification technique for abstracting information for social agents to yield feasible ToM models of humans they interact with. Affinity, which is essentially based on abstracting observed and communicated similarities, is one particular interaction state we use in this paper. It presents a good starting point, being a human mental state which is also a valuable heuristic in decision-making, and inspires us to computationally formalize other useful interaction states as well.

As a follow-up work, we aim for a more complete model that captures the ways humans abstract their beliefs and knowledge. We will start with a formalization from tip to toe (i.e., beliefs, abstractions, procedure etc.). For that, we need to answer a couple of fundamental questions such as which beliefs to use when abstracting, when to stop the procedure, what to do in case of a belief update, and which interaction states to activate after abstracting. In addition to these issues, a ToM agent should also be able to correctly attribute this abstraction process to others. As we aim to design higher-order ToM agents that can also take into account how their own artificial minds are perceived by others, we plan to benefit from *mind perception theory* [18, 22] when investigating the roles of observation and communication in recursive ToM reasoning. Additionally, we consider benefiting from *value-based reasoning* [28, 10, 3] to develop agents that takes others' values into account when doing ToM reasoning. With a more comprehensive, formalized model, we will further analyze how affinity can be used within other negotiation and communication protocols and strategies as well as get a broader view of its effects in multi-issue negotiations.

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. Akata, Z., Balliet, D., De Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., et al.: A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer **53**(08), 18–28 (2020)
2. Aronson, E., Akert, R.M., Wilson, T.D.: Social Psychology. Upper Saddle River, Prentice Hall, 7th edn. (2010)
3. Atkinson, K., Bench-Capon, T.: Taking account of the actions of others in value-based reasoning. Artificial Intelligence **254**, 1–20 (2018)
4. Baarslag, T., Hendrikx, M.J.C., Hindricks, K.V., Jonker, C.M.: Learning about the opponent in automated bilateral negotiation: A comprehensive survey of opponent modeling techniques. Autonomous Agents and Multi-Agent Systems **30**(5), 849–898 (09 2016)
5. Baker, C.L., Saxe, R.R., Tenenbaum, J.B.: Bayesian theory of mind: Modeling joint belief-desire attribution. Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society **33**(33) (01 2011)
6. Baksh, R.A., Abrahams, S., Auyeung, B., MacPherson, S.E.: The Edinburgh Social Cognition Test (ESCoT): Examining the effects of age on a new measure of theory of mind and social norm understanding. PloS ONE **13**(4), e0195818 (2018)
7. Bell, R.A., Daly, J.A.: The affinity-seeking function of communication. Communications Monographs **51**(2), 91–115 (1984)
8. Brooks, C., Szafir, D.: Building second-order mental models for human-robot interaction. arXiv preprint arXiv:1909.06508 (2019)

9. Carruthers, P., Smith, P.K.: Theories of theories of mind. Cambridge University Press (1996)
10. Cranefield, S., Winikoff, M., Dignum, V., Dignum, F.: No pizza for you: Value-based plan selection in BDI agents. In: IJCAI. pp. 178–184 (2017)
11. 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 **199-200**, 67–92 (2013)
12. 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, pp. 213–224 (2014)
13. 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. pp. 128–136. Springer (2014)
14. De Weerd, H., Verbrugge, R., Verheij, B.: Higher-order theory of mind in the tacit communication game. Biologically Inspired Cognitive Architectures **11**, 10–21 (2015)
15. 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 **31**(2), 250–287 (2017)
16. Fiske, S.T.: Stereotyping, prejudice, and discrimination. In: Gilbert, D.T., Fiske, S.T., Lindzey, G. (eds.) Handbook of Social Psychology, vol. 2, pp. 357–411. McGraw-Hill, New York, 4 edn. (1998)
17. Görür, O.C., Rosman, B.S., 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)
18. Gray, H.M., Gray, K., Wegner, D.M.: Dimensions of mind perception. Science **315**(5812), 619–619 (2007)
19. Heyes, C.M.: Imitation, culture and cognition. Animal Behaviour **46(5)**, 999–1010 (1993)
20. 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). pp. 180–193 (2019)
21. Lazarsfeld, P.F., Merton, R.K.: Friendship as a social process: A substantive and methodological analysis. Freedom and Control in Modern Society **18**(1), 18–66 (1954)
22. Lee, M., Lucas, G., Gratch, J.: Comparing mind perception in strategic exchanges: Human-agent negotiation, dictator and ultimatum games. Journal on Multimodal User Interfaces **15**(2), 201–214 (06 2021)
23. Merriam-Webster: Affinity. In Merriam-Webster.com dictionary ((nd)), https://www.merriam-webster.com/dictionary/affinity
24. Michlmayr, M.: Simulation Theory Versus Theory Theory: Theories Concerning the Ability to Read Minds. Master's thesis, Leopold-Franzens-Universität Innsbruck (2002)
25. Osten, F.B.V.D., Kirley, M., Miller, T.: The minds of many: Opponent modeling in a stochastic game. In: IJCAI. pp. 3845–3851. AAAI Press (2017)
26. Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences **1**(4), 515–526 (1978)
27. Raiffa, H.: The art and science of negotiation. Harvard University Press (1982)

28. Rangel, A., Camerer, C., Montague, P.R.: A framework for studying the neurobiology of value-based decision making. Nature Reviews Neuroscience **9**(7), 545–556 (2008)
29. Rosati, A., Knowles, E., Kalish, C., Gopnik, A., Ames, D., Morris, M.: What theory of mind can teach social psychology: Traits as intentional terms. In: Malle, B.F., Moses, L.J., Baldwin, D.A. (eds.) Intentions and Intentionality: Foundations of Social Cognition, pp. 287–303. MIT Press Cambridge, MA (2003)
30. Rosenschein, J.S., Zlotkin, G.: Designing conventions for automated negotiation. AI Magazine **15**(3), 29 (09 1994)
31. Stevens, C., Taatgen, N.A., Cnossen, F.: Metacognition in the prisoner's dilemma. In: 13th Annual International Conference on Cognitive Modeling. p. 112 (04 2015)
32. Suls, J., Martin, R., Wheeler, L.: Social comparison: Why, with whom, and with what effect? Current Directions in Psychological Science **11**(5), 159–163 (2002)
33. Tversky, A., Kahneman, D.: Judgment under uncertainty: Heuristics and biases. Science **185**(4157), 1124–1131 (1974)
34. Winfield, A.F.T.: Experiments in artificial theory of mind: From safety to storytelling. Frontiers in Robotics and AI **5**, 75 (2018)