HHAI 2024: Hybrid Human AI Systems for the Social Good
F. Lorig et al. (Eds.)

doi:10.3233/FAIA240201

Common Ground Provides a Mental Shortcut in Agent-Agent Interaction

Ramira VAN DER MEULEN [a,1], Rineke VERBRUGGE [b] Max VAN DUIJN [a]
[a] *Leiden Institute of Advanced Computer Science, Leiden University*
[b] *Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen*

Abstract. With the growing integration of chatbots, automated writing tools, game AI and similar applications into human society, there is a clear demand for artificially intelligent systems that can successfully collaborate with human partners. This requires overcoming not only physical and communicative barriers, but also those of fundamental understanding: Machines do not see and understand the world in the same way as humans do. We introduce the concept of 'Common Ground' (CG) as a possible solution. Using a model inspired on a collaborative card game known as 'The Game', we study agents that are instantiated to use different strategies, i.e., they each 'see' the model world in a different way. Agents work towards a joint goal that is easy to understand but complex to attain, requiring them to constantly anticipate their partner, which is classically seen as a task requiring active perspective modelling using a form of Theory of Mind. We show that agents achieving Common Ground increase their joint performance, while the need to actively model each other decreases. We discuss the implications of this finding for interaction between computational agents and humans, and suggest future extensions of our model to study the benefits of CG in hybrid human-agent settings.

Keywords. Human-AI Collaboration, Common Ground, Theory of Mind, Agent-based Models, Explainable AI

1. Introduction

With the digital complexity of human society reaching an all-time high, many voices advocate for increased human control over technology [1,2,3] – a cause that seems ever more challenging as the amount of data and the complexity of systems to be managed keeps growing. This necessitates a hybrid approach: humans and Artificial Intelligence *collaborating* towards common goals [4].

Human-AI collaboration has its unique challenges to overcome: Machines do not perceive, memorise, reason, etc., in the same way humans do. Here we address the question how entities that differ in how they 'see' the world can successfully interact, collaborate, and achieve joint goals. In humans, successful collaboration is often credited to actively modelling the other party's perspective, using a form of 'Theory of Mind' (ToM), the ability to take someone else's perspective and make estimations of their knowledge, beliefs, desires, and intentions [5]. Similarly, it has been argued that AI systems collaborating with

[1] Corresponding Author: Ramira van der Meulen, a.van.der.meulen@liacs.leidenuniv.nl

humans have to be able to understand their partners' perspectives, through a 'machine ToM' [6,7,8]. However, such active perspective modelling is cognitively expensive and prone to errors [9,10]. We introduce an alternative perspective offered by Clark [11], who argues that collaboration does not by default rely on active perspective modelling, but rather on what he refers to as 'Common Ground' (CG). Active modelling, then, is needed only in cases that *deviate* from the default created by CG [12].

We use this notion of CG to study how it can support agents in collaboration towards a joint goal, in spite of their initially different 'views' of a model world. We do this by implementing a version of the card game known as 'The Game', in which agents need to jointly play as many cards as possible, in counting order. Our agents differ in how they are initialised, such as in eagerness to play their own cards versus being more accommodating to their partners. This causes variance in their strategic repertoires – they initially have no 'strategic Common Ground'. The game is set up such that the agents will need to coordinate their actions in each turn. This coordination can be achieved either through constantly modelling the other's view on the game state, requiring a form of ToM, or through achieving CG on various aspects. We hypothesise that (I) accounting for the other's perspective using a form of ToM increases performance of the agents on a collaborative task and that (II) establishing CG will retain performance, while decreasing the need for agents to actively model the other's perspective.

We start by providing a brief background on CG and ToM, including previous work that formalised and modelled these phenomena. We then detail our computational agent-based model, and give an overview of our experimental setup. After this, we explain 'The Game' and how its implementation affected our model considerations. This is followed by a report and explanation of the results of our experiment, focusing on agent performance and how this is affected by ToM and CG. We discuss the implications for interaction between agents and humans, and conclude by suggesting a future extension of our model to study the benefits of CG in hybrid human-agent settings.

2. Background

2.1. Defining Common Ground and Theory of Mind

We define Common Ground (CG) as *information two or more individuals have in common about a given scenario*, and *access to the knowledge that they both have this information* [11,13], i.e., 'reflexive shared knowledge' ($CG_\phi : \phi,\ K_p\phi,\ K_q\phi,\ \forall_{p,q}(K_{p,q}K_{p,q}\phi)$). Part of CG roots in humans having the same physical structure: They have the same biology and senses as any friend *or* stranger. They know that they are all embodied humans, who need food and sleep to survive, and have hopes, desires, and intentions that drive them. CG is further shaped through life experiences in a process referred to as *grounding* [14,15]. If two individuals grow up in the same country, they share knowledge about its history, its stories, its people, and its conventions. Moreover, if two individuals share an interest, they can rely on their knowledge about that interest to model each other. Even in completely different parts of the world, two people with a love for musical theater are both familiar with recent Broadway musicals, and once they learn about the other's love, they can immediately assume that knowledge of each other.

The concept of Theory of Mind was popularised by Premack and Woodruff when they researched whether chimpanzees have similar perspective-taking capabilities as humans

[16]. Research in the ensuing decades showed that ToM is a complex phenomenon that emerged over millions of years of evolutionary time [17,18] and takes until at least late adolescence to fully develop in an individual human's lifetime [19,20,21,22]. The most well-known evaluation method is the 'Sally-Anne' task, in which pre-school children are asked to evaluate the mental model of a fictional character who had different or partial access to information in a short story [23]. As humans grow up, their ToM is greatly shaped by their culture [24,25], which influences how they interpret an event [26,27] and whether they may view someone's actions as egocentric or collective-oriented [28,29,30]. ToM is also influenced by receptiveness to cultural differences [31,32]. In this sense, one's CG informs their use of ToM, which should factor into models involving ToM and the establishment of CG.

There is debate over the question whether grounding relies on iteratively modelling the other's perspective, which would imply that CG itself relies on higher-order ToM [33].[2] In Clark's view, which we adopt here, this is not the case. Consider the following: People living in the same country know on which side of the road to drive and its traffic rules, and assume this of each other, making it safe to engage in traffic. Here, actions are successfully coordinated with little active thought. CG comes so naturally that it goes unnoticed – until there is a hitch. When someone suddenly starts to drive against the direction of the traffic, potential mismatches in CG come to the surface as people start to make sense of the situation: "What is this person *doing*? Do they not *know* on which side of the road one needs to drive here? Do they not *want* themselves and others to be safe?" The illusion is shattered – not because an active model has been violated, but because a passive assumption has failed. CG does not result from iteratively modelling others' perspectives, but forms the basis on which such modelling can take place if needed [12].

2.2. Modelling Theory of Mind and Common Ground

Previous research has shown that agent models can be greatly beneficial to expressing ToM in a shared setting, be it competitive, collaborative, or a mix thereof [34,35,36, 37,38]. We ground our model in 'Simulation-ToM' [39]: Agents implicitly predict their partner's perspective in their own behaviour, rather than explicitly representing it in their reasoning ('Theory-ToM'[3]). Work on the 'Tacit Communication Game' has shown that collaboration can be achieved through non-verbal, game action, communication [42]. Our setting follows the same idea: Agents do not communicate beyond playing their cards. We are not aware of any existing models in which agents attempt to form CG by our definition.

3. Methods and Model

We have developed a simulation in which agents play a collaborative counting game by observing and adjusting their behaviour based on their fellow player's moves. Agents implicitly assume that their partner is modelling them in response (inspired by Simulation-

[2]The argument is that for realising one has CG with the other, one needs to *know* that the other *knows* that one *knows*, etc., up to the fifth order.

[3]For a discussion of this classic opposition in the literature on ToM see [40,12] Alternatives to this debate include narrative practice [41] and multi-systems [5] approaches.

ToM [39]). They are initialised to represent different views on the world – they can be more egocentric, and vary in how receptive they are towards adjusting in the direction of agents who act differently than they do. We use an existing collaborative card game, known as 'The Game' [43], in which players take turns to try to play every card from their hand and the deck, in the right order. The goal is to exhaust the full set of cards. Players are not aware of the order of the cards in a shuffled deck [44] and are not allowed to share which cards they drew. This makes it an imperfect-information game.

3.1. Experimental Setting and Procedure

Each player takes turns to play at minimum two cards on any of four central 'piles', to either count up (from '1'), or down (from '100'). Cards that deviate from the top card by '10' may always be played, regardless of counting direction (an example setup can be found in Figure 1a). We use the base conditions balanced by the game designers, with an additional caveat that verbal communication is not allowed: Our setting consists of two players, who start with 7 cards in hand (each game uses a deck of 98 cards – ranging from 2 to 99). After playing 2 to 7 cards on any of the central piles, a player fills their hand back up to 7 and passes on their turn. It is important to coordinate actions to ensure that cards are played before it is too late (Figure 1b). This continues until at minimum one player can no longer play a card.[4] The final score is based on the played number of cards, so players benefit from modelling the actions and strategies of their partners.

To win 'The Game', players establish communication protocols between each other to ensure they play every card, without knowing each other's hand. In practice, this results in player protocols revealing their intentions and considerations: a 'CG'-based communication strategy specific to their group. We evaluate the establishment of these protocols by communication-through-play. This allows us to study CG formation through observation, rather than active discourse about intentions.[5]

(a) Example in 'The Game'. There are increasing and decreasing piles – and playing a card with a difference of '10' can reset the counting process.

(b) Player turns: Purple has played '53', '50' on yellow's '57'. Yellow then plays '60', '56', closing the window for purple's '59'.

Figure 1. Example situations in 'The Game', displaying two different game states.

3.2. Model Assumptions and Features

We do not simulate perfectly rational agents [45], as this would predefine a (CG) protocol and would not emulate human-realistic play [46,47]. Our models learn to score as high as

[4] If the deck is exhausted, players are allowed to play one card per turn (and they no longer draw cards).
[5] While purposefully kept out of scope here, such active discourse will be included in future research.

possible based on a few (unique) starting principles and iterated joint play. We initialised three settings, for 'Low', 'Medium' and 'High' skill agents, based on how well their interactions perform at the start of their gameplay. Every agent turn is observed by the other, who uses its own knowledge about the game to estimate the play-through strategy. If they see the other agent eagerly playing a lot of cards at the same time, they themselves likely start doing so as well – likewise, if the other agent plays fewer cards than the agent would expect, they are likely to act less eager as well. The agents in our model use three scalable features, based on interviews on human play in 'The Game'.

- **Self-benefit** to playing more than the required two cards, based on proximity to the top card ('Do I await more information, or is it better game-wise to play now?')
- **Eagerness** to play more than two cards given the current partner ('Do I want to get rid of my cards before I run the risk of being unable to play them? Do I give them more room to play cards?)
- **Cooperativeness** towards setting ('Does collaboration in its current form work?').

When an agent plays a card, it finds a balance between the learned self-benefit to play the card (sB_a) and the drive to play cards in a way that benefits collaboration with partner i ($eager_i$), informed by difference (dif_c) between the card and the numerically closest card pile (difference '2' is played more often than one of '7'). This results in equilibrium p_c (Eq. (1)). If the resulting collaboration improves performance, agents raise cooperation $coop_i$ (and vice versa). At peak value, self-benefit represents a fully egocentric approach: play the minimum number of cards, as new information is always better for oneself – it is raised and lowered based on the score average over 10 rounds. Eagerness both helps the agent realise that it should play more cards before its partner renders them unplayable, and helps it realise that it plays cards so often that its partner's cards become unplayable. Eagerness is normalised and adjusted with a $coop_i$ in both scenarios (up for 5% more cards - down for 5% fewer cards).

$$p_c : sB_a * totalNumCards > eager_i * dif_c \tag{1}$$

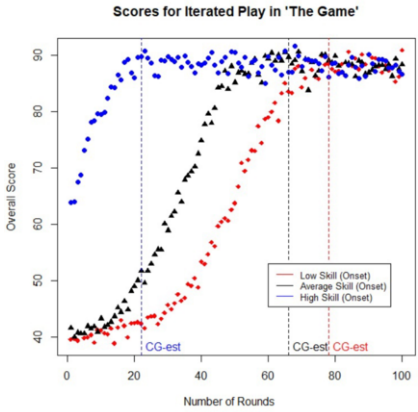
Modelling in our model is implicit: Agents observe the actions of their partner, and slightly adjust their behaviour based on those actions, using ToM to know their partners do the same. We express this in 'Eagerness' ($eager_i$). Successful collaboration entails correctly estimating when their partner will play a card in a given scenario, and when they themselves should. This does not mean that each agent has to respond exactly the same as the other, but it is crucial to align on what to do in a given situation. CG in turn comes from a state of mutual self-reflection. With the ToM implementation, each agent continuously updates its model based on observations of its partner. CG is assumed when an agent both (1) no longer notices a change in its partner's behaviour (update stop $eager_i$) and (2) no longer notices a change in the game score resulting from its own behaviour (update stop sB_a). With both of these conditions fulfilled, the agent 'believes' everyone is on the same wavelength about the current joint successful collaboration strategy.

The interactions in our model take place over the course of 100 rounds of 50 games. As each game consists of one randomly shuffled deck, some games are a lot harder to do well at than others. Using 50 games per round allows us to determine a decent average for evaluating our agents' overall performance. Once an agent detects that CG exists, it stops updating its behaviour – resulting in it no longer using ToM to strategically align with its partner, nor adjusting its own self-benefit sB_a - we call this CG-established ($CG\text{-}est$).

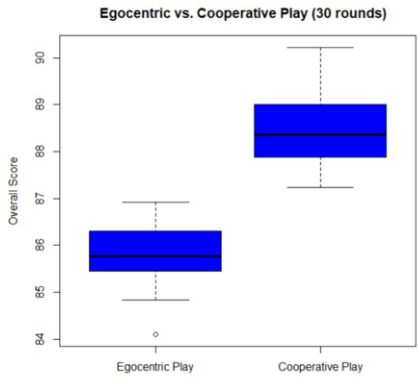
4. Results

Running our simulation shows that agents are able to solve the task by observing the game states and adapting to each other's playstyles. This leads to them obtaining scores in the range of 87 - 90 (Figure 2a). As discussed, we have run the simulation under various settings, resulting in a 'Low', 'Medium' and 'High' skill onset. Agents stabilise towards successful interaction, despite not communicating beyond playing the cards. We observe that post-convergence strategies continue to be effective despite every game differing. The agents themselves do *not* 'become' the same: A successful solution strategy can, f.e., involve one agent acting more eagerly to account for its partner's actions than its partner acts towards them. We highlight our results in light of our hypotheses below.

Hypothesis I: Accounting for the other's perspective using a form of ToM increases performance of the agents on a collaborative task We have compared multiple settings of initialised agents and their performance at the task. The agents always adapt their playstyle to account for their agent-partner successfully, and informed changes in eagerness ($eager_i$) increase the score (Figure 2a). Doing this in harmony with their partner results in collaboration: If an agent notices themselves becoming too eager (i.e., starts playing too many cards compared to their partner), they actively become less eager, to give their partner space. While the initial scores are generally low, the gameplay after learning to successfully collaborate results in 87 - 90 points. This is a significant improvement over an egocentric, self-benefit heavy, approach to the problem, where agents default to a strategy involving only self-centered play. A paired t-test comparison for 30 rounds using a collaboration strategy and 30 rounds using an egocentric strategy (Figure 2b) yields strongly significant results at $p < 0.0001$ ($t = 14.9265$, $df = 29$, $sed = 0.174$).



(a) Scores over 100 game rounds. Vertical lines indicate the point at which Common Ground is established from the perspective of both agents (CG-est).



(b) A comparison for egocentric vs. collaborative game performance in 'The Game'.

Figure 2. Figure (a) depicts the score increases over rounds for collaborative agent-agent play; Figure (b) displays the results of egocentric vs collaborative play.

Hypothesis II: Common Ground will retain model performance, allowing the agents to decrease their use of ToM Our models show that the agents can agree on a joint strategy, once they (1) observe that the behaviour of their agent-partner no longer changes, and (2) decide that their individual behaviour no longer yields a higher score. This results in them 'locking' their playstyle. Afterwards, the scores remain stable, despite every game being randomly shuffled (Figure 2a): After establishing CG, agents no longer update their behaviour by actively modelling their opponent after every round, while the score does not decline. We can accept Hypothesis II under the condition that the CG is *genuine*: While the agents in Figure 2a have all converged to an equilibrium, agents who falsely assume CG will eventually decline into an egocentrical strategy, losing model performance (this happens when *CG-est* does not overlap between agents). We elaborate on this in the Discussion.

5. Discussion

Our model shows that if there is sufficient CG, the agents no longer need to actively anticipate the playstyle of their partner (without losing performance). Computationally, the CG becomes a shortcut to skip the reasoning (update) steps that the agents would otherwise have to perform. This helps with resolving our aforementioned notion that actively modelling others is costly [9,48,10]. Forming CG may be a means of decreasing cognitive workload in human-human and hybrid human-machine interaction as well.

Establishing CG *does* occasionally fail: If agents only *think* they have found a solution strategy that works for both of them, the collaboration approach will eventually fall back into, or develop towards, egocentrism. This seems realistic, given that even rigid CG can fail ('Which side of the road to drive on', Background section). As agents cannot communicate about each other's intentions when an 'uncharacteristically' complex situation arises, establishing CG without communication is quite fragile. Such fragility especially occurs when the agents perform well in one round, and then perform badly at a high number of games in the next round.[6] The only method for an agent to break the slow adoption of egocentric strategies is to actively start modelling their partner again.

This drift to egocentrism is illustrated in Table 1. If both agents have established the two model conditions for CG at the same time, the collaboration remains stable. The more accurately CG was pinpointed, the longer the collaboration strategy stays in use.[7] This is, however, not guaranteed: e.g., Agent a can falsely assume CG, and stops modelling agent b. Agent b continues to model agent a, and either fixes a balance that misaligns with agent a, or never 'finds' CG at all. This reinforces the importance of 'grounding' whether there actually *is* CG — which *does* rely on ToM.

Table 1. Post-training belief B_i of $Agent_i$ in CG with $Agent_j$ ($CG_{i,j}$) – and whether that is actually the case.

	$B_b(CG_{a,b}),\ CG_b$	$B_b(CG_{a,b}),\ \neg CG_{a,b}$
$B_a(CG_{a,b}),\ CG_a$	Successful collaboration	Slow decline towards solo-strategy
$B_a(CG_{a,b}),\ \neg CG_a$	Slow decline towards solo-strategy	Rapid decline towards solo-strategy

[6] This almost happens for the 'High Skill (Onset)' (blue) in Figure 2a - rounds 9 to 11. CG is not falsely established here, because agents are still learning from each other (through $eager_i$). If one of two agents had concluded that its partner was no longer changing its behaviour as well, the equilibrium would have collapsed.

[7] Close pinpoint: assuming CG when the partner was about to 'lock' its behaviour, but had not done so yet.

CG in our definition ultimately results from a collaboratively agent-written script [49]. This *can* result in imperfect solutions if the situation is more complex than the agents have managed to perceive. As the agents have stopped modelling each other, they will never realise that there was a better possible outcome. This fits with current literature that shows ToM to be mostly useful in more complex scenarios [8]. The CG shortcut results in a successful but suboptimal outcome that can only be overcome with cognitively complex and demanding reasoning. We hypothesise that giving our agents the ability to explicitly communicate about a strategy works as a cost-effective solution to resolve this issue: If they signal that new experience has taught them there may be a better long-term strategy, they can break the cycle by agreeing to both adapt. We leave this aspect to be explored in our future work.

6. Conclusion and Follow-up

We have modelled a counting task known as 'The Game' to study how agents that differ in their initial view on the model world, can reach a successful agent-agent collaboration, by taking the other's perspective into account using a form of ToM. We highlight that an implicit mutual reflection on the success of this collaboration allows for the solution strategy to be 'fixed' into a joint strategy – finding Common Ground, which decreases the need to actively model one's partner. This is an important step for formalising the relationship between CG and ToM. Understanding this relationship will in turn help reasoning about alignment in both agent-agent and agent-human collaborative environments.

The next step is modelling a more nuanced representation of differences between agents, both on a fundamental (architecture) and acquired (nurture) level. We wish to see whether our hypotheses still hold if we ask our agents to collaborate with humans of different backgrounds. Additionally, our agents have a shared goal, but there is no explicit shared intent [50]: Each agent figures out what their partner is doing, but this happens implicitly. To speak about true collaboration, the agents have to actively state their intentions and discuss before one of them takes an action that affects the interaction, instead of responding based on observations only. This once again seems to indicate that we need to introduce explicit communication to further develop the impact of our model.

Lastly, we wish to experiment with the formation of CG in other scenarios. This includes other game settings, but also research tasks more grounded in real-world practice. 'The Game' is useful to study the formation of CG, but is limited in its action space: CG there mostly concerns how many cards to play in any given situation. Social dilemmas are more complex than optimising strategic counting, and we wish to show that our models are similarly capable of handling such situations. One might consider resource allocation, which is heavily influenced by ToM even in early infancy under both explicit [51] and hidden conditions [52]. Additional avenues include teaching-related scenarios: Reaching CG in a teaching scenario through explicit reinforcements in one's behaviour [53], instructions about conformity [54], or peer-to-peer teaching [55]. In such ways, we wish to contribute to a society in which ever-advancing technologies stay aligned with humans.

Acknowledgments This research is part of the Hybrid Intelligence gravitation programme – number 024.004.022, financed by the Netherlands Organisation for Scientific Research (NWO).

References

[1] Bryson JJ, Theodorou A. How society can maintain human-centric artificial intelligence. In: Toivonen M, Saari E, editors. Human-Centered Digitalization and Services. Springer; 2019. p. 305-23.

[2] Shneiderman B. Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human–Computer Interaction. 2020;36(6):495-504.

[3] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; 2021. p. 610-23.

[4] 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(8):18-28.

[5] Apperly IA, Butterfill SA. Do humans have two systems to track beliefs and belief-like states? Psychological Review. 2009;116(4):953.

[6] Baker C, Saxe R, Tenenbaum J. Bayesian theory of mind: Modeling joint belief-desire attribution. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 33; 2011. .

[7] Rabinowitz N, Perbet F, Song F, Zhang C, Eslami SA, Botvinick M. Machine Theory of Mind. In: International Conference on Machine Learning. PMLR; 2018. p. 4218-27.

[8] De Weerd H, Verbrugge R, Verheij B. Higher-order theory of mind is especially useful in unpredictable negotiations. Autonomous Agents and Multi-Agent Systems. 2022;36(2):30.

[9] Lewis PA, Birch A, Hall A, Dunbar RIM. Higher order intentionality tasks are cognitively more demanding. Social Cognitive and Affective Neuroscience. 2017 03;12(7):1063-71.

[10] Wilson R, Hruby A, Perez-Zapata D, van der Kleij SW, Apperly IA. Is recursive "mindreading" really an exception to limitations on recursive thinking? Journal of Experimental Psychology: General. 2023.

[11] Clark HH. Using Language. Cambridge University Press; 1996.

[12] Van Duijn MJ. The Lazy Mindreader. A Humanities Perspective on Mindreading and Multiple-Order Intentionality; 2016.

[13] Verhagen A. Grammar and cooperative communication. In: Dąbrowska E, Divjak D, editors. Handbook of Cognitive Linguistics. Berlin, München, Boston: De Gruyter Mouton; 2015. p. 232-52.

[14] Clark EV. Common ground. In: The Handbook of Language Emergence. Wiley Online Library; 2015. p. 328-53.

[15] Geurts B. Convention and common ground. Mind & Language. 2018;33(2):115-29.

[16] Premack D, Woodruff G. Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences. 1978;1(4):515-26.

[17] Dunbar RI, Shultz S. Understanding primate brain evolution. Philosophical Transactions of the Royal Society B: Biological Sciences. 2007;362(1480):649-58.

[18] Tomasello M, Vaish A. Origins of human cooperation and morality. Annual Review of Psychology. 2013;64:231-55.

[19] Baron-Cohen S, Leslie AM, Frith U. Does the autistic child have a "theory of mind"? Cognition. 1985;21(1):37-46.

[20] Meinhardt-Injac B, Daum MM, Meinhardt G. Theory of mind development from adolescence to adulthood: Testing the two-component model. The British Journal of Developmental Psychology. 2020;38:289–303.

[21] Samson D, Apperly IA, Braithwaite JJ, Andrews BJ, Bodley Scott SE. Seeing it their way: Evidence for rapid and involuntary computation of what other people see. Journal of Experimental Psychology: Human Perception and Performance. 2010;36(5):1255.

[22] Apperly I. Can theory of mind grow up? Mindreading in adults, and its implications for the development and neuroscience of mindreading. In: Understanding Other Minds: Perspectives from developmental social neuroscience. Oxford University Press Oxford; 2013. p. 72-92.

[23] Wimmer H, Perner J. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition. 1983;13(1):103-28.

[24] Vygotskij LS, John-Steiner V. Mind in society: The development of higher psychological processes. Harvard University Press; 1979.

[25] Tomasello M, Moll H. The gap is social: Human shared intentionality and culture. In: Mind the Gap: Tracing the Origins of Human Universals. Springer; 2010. p. 331-49.

[26] Spaulding S. Do you see what I see? How social differences influence mindreading. Synthese. 2018;195:4009-30.

[27] Lavelle JS. The impact of culture on mindreading. Synthese. 2021;198(7):6351-74.
[28] Wellman HM, Liu D. Scaling of theory-of-mind tasks. Child Development. 2004;75(2):523-41.
[29] Wellman HM, Fang F, Liu D, Zhu L, Liu G. Scaling of theory-of-mind understandings in Chinese children. Psychological Science. 2006;17(12):1075-81.
[30] Shahaeian A, Nielsen M, Peterson CC, Slaughter V. Iranian mothers' disciplinary strategies and theory of mind in children: A focus on belief understanding. Journal of Cross-Cultural Psychology. 2014;45(7):1110-23.
[31] Perez-Zapata D, Slaughter V, Henry JD. Cultural effects on mindreading. Cognition. 2016;146:410-4.
[32] Kim LR, Jetten J, Pekerti A, Slaughter V. Mindreading across cultural boundaries. International Journal of Intercultural Relations. 2023;93:101775.
[33] Scott-Phillips T. Speaking Our Minds: Why Human Communication is Different, and how Language Evolved to Make it Special. Bloomsbury Academic; 2015.
[34] 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;31:250-87.
[35] Baker C, Saxe R, Tenenbaum J. Bayesian models of human action understanding. Advances in Neural Information Processing Systems. 2005;18.
[36] Baker C, Tenenbaum J. Modeling human plan recognition using Bayesian theory of mind. In: Sukthankar G, Geib C, Bui HH, Pynadath D, Goldman RP, editors. Plan, Activity, and Intent Recognition: Theory and Practice. vol. 7. Waltham, MA: Morgan Kaufmann; 2014. p. 177-204.
[37] 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.
[38] Foerster J, Song F, Hughes E, Burch N, Dunning I, Whiteson S, et al. Bayesian action decoder for deep multi-agent reinforcement learning. In: International Conference on Machine Learning. PMLR; 2019. p. 1942-51.
[39] Gallese V, Goldman A. Mirror neurons and the simulation theory of mind-reading. Trends in Cognitive Sciences. 1998;2(12):493-501.
[40] Gallagher S. Empathy, simulation, and narrative. Science in Context. 2012;25:355-81.
[41] Gallagher S, Hutto D. Understanding others through primary interaction and narrative practice. In: The Shared Mind: Perspectives on Intersubjectivity. John Benjamins; 2008. p. 17-38.
[42] 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.
[43] Benndorf S. The Game. White Goblin Games;. Accessed: 2024-04-12. https://boardgamegeek.com/boardgame/173090/the-game.
[44] Fisher RA, Yates F. Statistical tables for biological, agricultural and medical research. Hafner Publishing Company; 1953.
[45] Aumann RJ. Agreeing to disagree. Annals of Statistics. 1976;4:1236-9.
[46] McKelvey RD, Palfrey TR. An experimental study of the centipede game. Econometrica: Journal of the Econometric Society. 1992:803-36.
[47] Pulford BD, Krockow EM, Colman AM, Lawrence CL. Social value induction and cooperation in the centipede game. PLoS One. 2016;11(3):e0152352.
[48] Frith C, Frith U. Theory of mind. Current Biology. 2005;15(17):R644-5.
[49] Schank RC, Abelson RP. Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Psychology Press; 2013.
[50] Dunin-Keplicz B, Verbrugge R. Collective intentions. Fundamenta Informaticae. 2002;51(3):271-95.
[51] Mulvey KL, Buchheister K, McGrath K. Evaluations of intergroup resource allocations: The role of theory of mind. Journal of Experimental Child Psychology. 2016;142:203-11.
[52] Li L, Rizzo MT, Burkholder AR, Killen M. Theory of mind and resource allocation in the context of hidden inequality. Cognitive Development. 2017;43:25-36.
[53] Alibali MW, Nathan MJ, Church RB, Wolfgram MS, Kim S, Knuth EJ. Teachers' gestures and speech in mathematics lessons: Forging common ground by resolving trouble spots. ZDM – Mathematics Education. 2013;45:425-40.
[54] Tomasello M. The ontogeny of cultural learning. Current Opinion in Psychology. 2016;8:1-4.
[55] Ziv M, Solomon A, Strauss S, Frye D. Relations between the development of teaching and theory of mind in early childhood. Journal of Cognition and Development. 2016;17(2):264-84.