Title
Egocentric Tendencies in Theory of Mind Reasoning:An Empirical and Computational Analysis

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 41(0)

Authors
P¨oppel, Jan
Kopp, Stefan

Publication Date
2019

Peer reviewed

Egocentric Tendencies in Theory of Mind Reasoning: An Empirical and Computational Analysis

Jan Pöppel (jpoeppel@techfak.uni-bielefeld.de) and **Stefan Kopp (skopp@techfak.uni-bielefeld.de)**
Social Cognitive Systems, CITEC, Bielefeld University
Bielefeld, Germany

Abstract

Humans develop an ability for Theory of Mind (ToM) by the age of six, which enables them to infer another agent's mental state and to differentiate it from one's own. Much evidence suggests that humans can do this in a presumably optimal way and, correspondingly, a Bayesian Theory of Mind (BToM) framework has been shown to match human inferences and attributions. Mostly, this has been investigated with specific, explicit mentalizing tasks. However, other research has shown that humans often deviate from optimal reasoning in various ways. We investigate whether typical BToM models really capture human ToM reasoning in tasks that solicit more intuitive reasoning. We present results of an empirical study where humans deviate from Bayesian optimal reasoning in a ToM task but instead exhibit egocentric tendencies. We also discuss how computational models can better account for such sub-optimal processing.

Keywords: Theory of Mind; Bayesian Modeling; Egocentric Tendencies; Bounded rationality

Introduction

An important ability of humans is to infer and reason about ones own as well as other's mental states such as intentions, (potentially false) beliefs, or emotions (Wellman & Liu, 2004). While the exact development of this so-called Theory of Mind (ToM) (Premack & Woodruff, 1978) is still not clear, there is a consensus that we acquire full ToM abilities around the age of six (Wellman & Liu, 2004). This allows us to make sense of our social environment, to learn more from the actions around us (Jara-Ettinger, Baker, & Tenenbaum, 2012) and to better understand or even manipulate others in cooperative or competitive interactions (Heyes & Frith, 2014).

Because of its importance for social interaction, there is a great interest in endowing artificial systems with similar capabilities. Recently, the most prominent approach has been the Bayesian Theory of Mind (BToM) framework (Baker, Saxe, & Tenenbaum, 2009). Building upon the rational agent assumption (Dennett, 1989) and inverse planning, the BToM framework constructs probabilistic generative models that relate hidden mental states to observable actions. These models can then be inverted to infer mental states from behavior, while accounting for inherent uncertainty. This framework has been shown to make inferences that correlate well with those made by humans in a wide range of different tasks, such as the inference of desires and beliefs (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017) or preferences (Jern, Lucas, & Kemp, 2017). It is also in line with the Bayesian Brain hypothesis positing that humans incorporate information similar to optimal observers (Knill & Pouget, 2004).

At the same time, humans do not always behave like optimal observers or reasoners. Instead, they exhibit a range of fallacies leading to systematic errors in different types of inference (Haselton, Nettle, & Murray, 2015). This has also been argued to hold for social interaction. For example, Keysar (2007) showed that adults fail to adjust correctly for different perspectives in communication tasks. This trait is often referred to as *egocentric tendency* and refers to the tendency to impute one's own mental perspective on others (Nickerson, 1999). Keysar, Lin, and Barr (2003) present an experiment in which even adults fail a false-belief test, showing that an egocentric tendency is not always effectively suppressed. In other words, we do not appear to always use our ToM capabilities to the fullest extent (cf. (De Weerd, Verbrugge, & Verheij, 2013)). This is often attributed to limited mental resources, such as working memory and processing time. Vul, Goodman, Griffiths, and Tenenbaum (2014) argue that many biases are actually optimal when seen as the result of the number of samples for inference being limited.

It is unclear how those limitations affect ToM reasoning in humans. While the BToM framework has been shown to correlate well with humans' explicit ToM reasoning, it has not been evaluated with regard to humans' intuitive or implicit inferences, i.e. when sophisticated ToM reasoning is not explicitly evoked. Recently, Nakahashi and Yamada (2018) showed that a full inverse planning approach based on the BToM framework overestimates the rationality of humans and that modified inference achieves better correlations with human judgments. We are interested in whether, in an intuitive setting, humans employ different kinds of ToM models as a function of, e.g., computational costs, available resources, or current task demands. We have argued elsewhere that employing different kinds of ToM models for "satisficing mentalizing" can be beneficial for artificial systems, where full Bayesian models often suffer from intractabilities (Pöppel & Kopp, 2018). Here, we study whether humans may also employ different simpler, non-optimal models depending on the given circumstances and, specifically, whether they may fail to realize or account for differences between one's own and another one's mental states. We thus focus on the extent to which humans employ mentalizing in a settings that is more implicit than those used in previous BToM research.

In the remainder of this paper, we present empirical evidence suggesting that humans exhibit different degrees of egocentric tendencies in a simple ToM reasoning task, thus deviating from rational optimal observers usually assumed in previous BToM models. The next section first describes the scenario we are looking at. Then, we present an empirical study we have carried out in this scenario to investigate intuitive human ToM reasoning. After this, we present different computational ToM models, partially based on the BToM framework, and report their correlations with our data.

Scenario

The scenario we chose for our empirical study is the inference of an agent's desire in a navigation task within a 2D maze. The maze has four exits, each of which leading to a distinct destination (denoted Red, Yellow, Blue or Orange). The agent has to find the exit that leads to one specific destination, which we consider to be the agent's desire. In previous work we already gathered behavioral data in the form of trajectories of human participants solving this navigation task in different mazes with differing amounts of information available (Pöppel & Kopp, 2018). Here, we consider the task of an additional observer, who watches the agent move around in the maze and has to infer the agent's desire – a perceptual and cognitive task humans solve frequently in everyday life. According to the ToM scale by Wellman and Liu (2004), this kind of inference is also among the first ones to be mastered by children.

In order to create a need for differentiating between the mental perspectives of the agent and the observer, we employ two conditions: in the first condition, participants acting as agents had full knowledge of the maze, the locations of all exits, and the destinations behind them. That is, they could take an optimal path in order to reach their desired destination. In the second condition, the acting participants knew about the locations of the exits in the maze, but had to discover the corresponding destinations themselves by establishing a line of sight with the exit. Thus participants had to search for the specific exit (one out of four) that leads to their desired destination, resulting in an exploration behavior. This scenario is similar to earlier work on BToM, e.g. (Baker et al., 2017), in that it involves navigating a simple grid-world to achieve a desired outcome with potential uncertainty about the true location of that outcome.

Figure 1 shows an example of the different stimuli that the acting participants received in the two conditions. In the bottom example belonging to the second condition, the exits are marked but covered. Note that in this situation the agent has moved to a position, where it could see the exit thus revealing its corresponding destination (Blue). In the present study, we use recordings of the online navigation behavior in these two conditions and let human participants play the role of the observer. In particular, their task is to identify the desired destination of the observed agent at different points on the recorded trajectory.

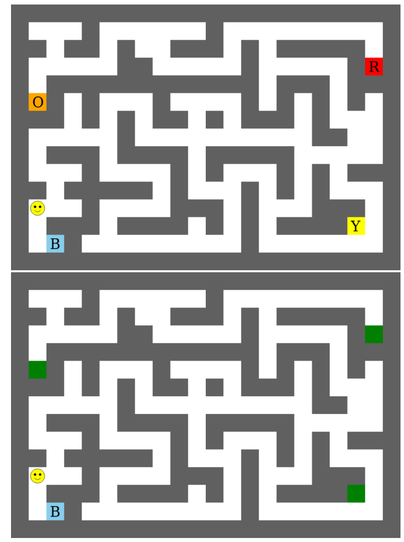

Figure 1: View of the navigating agents. Top: full knowledge about exits and destinations; Bottom: exits only reveal their destination once a line of sight is established.

Empirical Study: ToM Reasoning in Humans

Humans employ their ToM capabilities rarely to their fullest extent. However, it is still unclear what factors, apart from cognitive load, may influence the extent to which a person employs her ToM capabilities. Previous research has shown that explicit asking for likelihood ratings of all alternatives yields responses predicted by the BToM framework. However, we conjecture that this experimental design inherently evokes explicit reasoning about mental states in the participants, including the full consideration and comparison of all alternatives. This evocation may be part of the reason for the discrepancy between very good fittings in BToM research and findings of suboptimal behavior in other research. In contrast to previous research, we therefore deliberately chose not to ask for likelihood ratings for all possible desires, but instead ask for *soft forced-choice* responses in order to test for a more intuitive and natural ToM reasoning. We call it *soft* because we gave participants the additional option "I do not know".

We also included a second group of participants who were additionally prompted to self-assess their belief about the observed agent's knowledge. We included this group to test the effect of putting an agent's belief into focus of (more explicit) ToM reasoning, thus testing if different task demands influence the employed ToM models.

Participants We recruited two distinct groups of participants (first group 120; second group 65) each via an online platform called "figure-eight" (formerly crowdflower). All participants had a "contributor level three", which is advertised as "Highest Quality: Smallest group of most experienced, highest accuracy contributors". After completion of the study, participants were reimbursed with \$0.20 via the figure-eight system.

Stimuli For each of the two conditions mentioned above, we chose two typical trajectory recordings in two different mazes. The four trajectories are shown in figure 2. Participants could see the maze and, importantly, all destinations behind the different exits. That is, they always had full knowl-

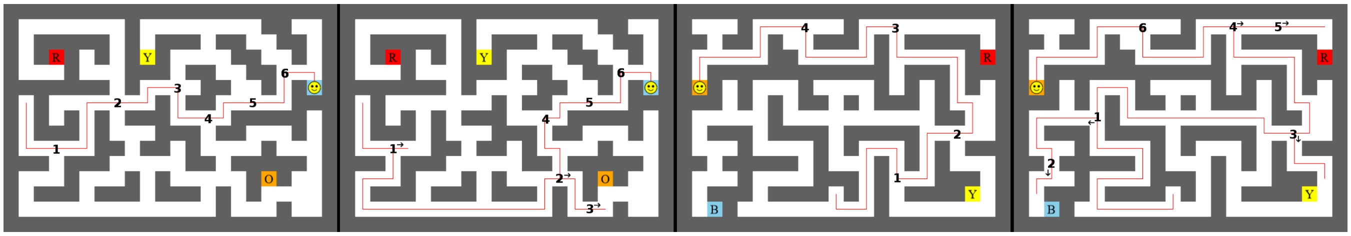

Figure 2: The recordings 1 through 4 used as stimuli, one from each condition (destinations known/unknown to the agent) in two different mazes. Numbers indicate the location of query points at which participants had to give their responses. Small arrows indicate the agent's next action after the query point when ambiguous.

edge about the destinations, while the agents they observed might or might not have had the same knowledge. Participants would see the agent navigating the maze according to the recordings, leaving behind a red trail as in figure 2, so that participants always knew where the agent moved. We chose to replay these recordings with fixed time intervals between two steps in order to remove potential noise within the trajectories (while also eliminating information about speed or possible hesitations of the agents).

Procedure All participants received the same initial instruction that they were going to watch the recordings of four different human players navigating a maze. They were informed that the maze had four exits, each leading to a different destination, and that each player had her own specific destination she had to reach as quickly as possible. We further explained that the four players were part of two different conditions. In one condition, they had full knowledge about the destinations behind the exits, while in the other they had to first discover which exit led to which destination. In order to make this clear, we provided participants with the images in figure 1 alongside the instructions. The instructions read: "*Now you will watch the agents follow their trajectories while you will be able to see which exit corresponds to which destination. At certain points in time you will be asked to tell the agent's desired destination (R,B,Y,O). You may also say that you do not know*." The query points are those shown in figure 2. We deliberately asked for the agent's *desired destination* instead of an exit to focus on the agent's desire instead of the exit locations close to the agent. The second group of participants were further instructed about the additional question regarding the agent's knowledge, which read "*Additionally, you will be asked to specify if you think the agent knows which exit leads to which destination*.".

Upon confirming these instructions, participants got to see the first maze as in figure 2 (without the query point numbers) with the agent at the beginning of its trajectory. After hitting a *Start* button the agent started to move leaving behind the red trail. The playback stopped at each QP and participants were asked to choose one out of the four possible goal destinations, or to signal that they cannot tell otherwise, which we will refer to as *Uncertain* (U) from here on. In order to avoid misinterpretations (such as having asked for current target location only), we instructed participants with: "*Please specify which destination you think the agent wants to reach after leaving the maze*." The second group of participants received an additional question before identifying the agent's destination: "*Do you think the agent you are watching currently knows which exit leads to which destination?*" Participants could respond with either "Yes" or "No". Once the agent reached its destination, participants could proceed to the next recording. In total, each of the 185 participant had to make 22 judgements (taking less than 400s on average for the first group and less than 485s for the second group).

For the first group, we counter-balanced the order in which participants saw the different recordings/mazes. We used a Fisher's exact test on the response frequencies in order to test whether or not the order in which the stimuli was presented had any effect on participants' responses. The test revealed no significant effects of the stimuli ordering for all but one responses (recording 1, QP 6). We thus concluded that the order of presentation of recorded trajectories/mazes did not influence participant's responses. We thus collapse the results of participants in the first group for the analyses in the remainder of this paper. Furthermore, we decided to use only one ordering for the second group in order to simplify the design.

For analyses, we excluded all participant's responses for a particular recording if participants always picked the same destination and if this destination was not the correct one within one recording. We further excluded responses if participants chose to predict a destination after the agent already turned away from it in recordings 2 and 4. We assume that these participants did not really pay attention to the actual trajectories as these are obvious errors. After excluding such participants, we had 110 participants in group 1 and 57 participants in group 2 remaining.

Results Figures 3 and 4 show normalized response frequencies for several interesting query points in the two groups. Note, however, that the reported tendencies also hold for the other recordings and query points.

For the first group of participants who only had to identify the likely destination of the agent, we find a strong bias

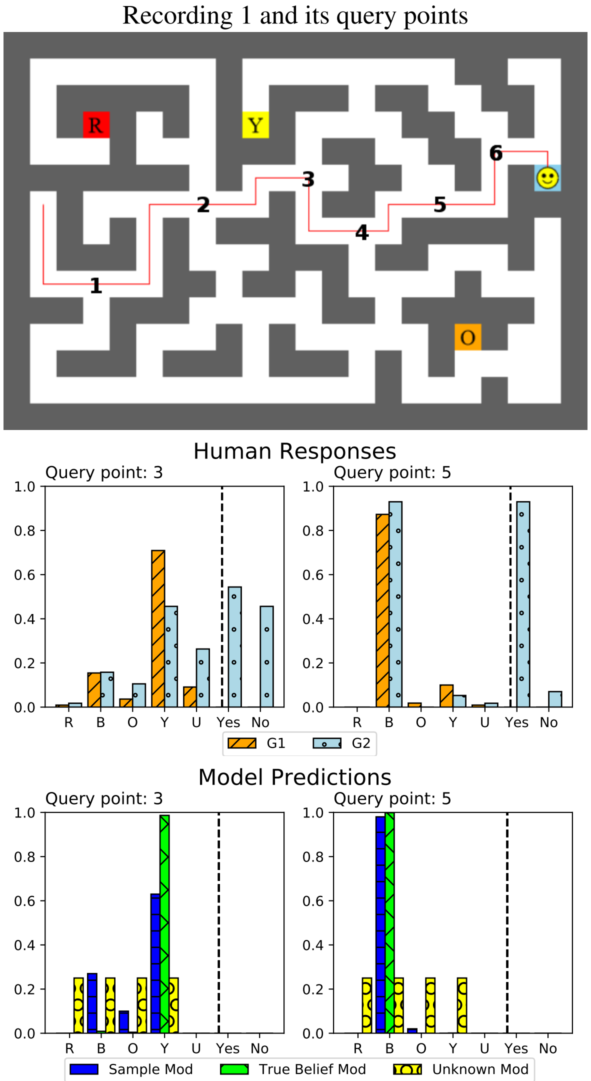

Figure 3: Relative response frequencies by the participants in both groups for Recording 1, query points 3 and 5 and the corresponding predictions made by the models using the likelihood modification

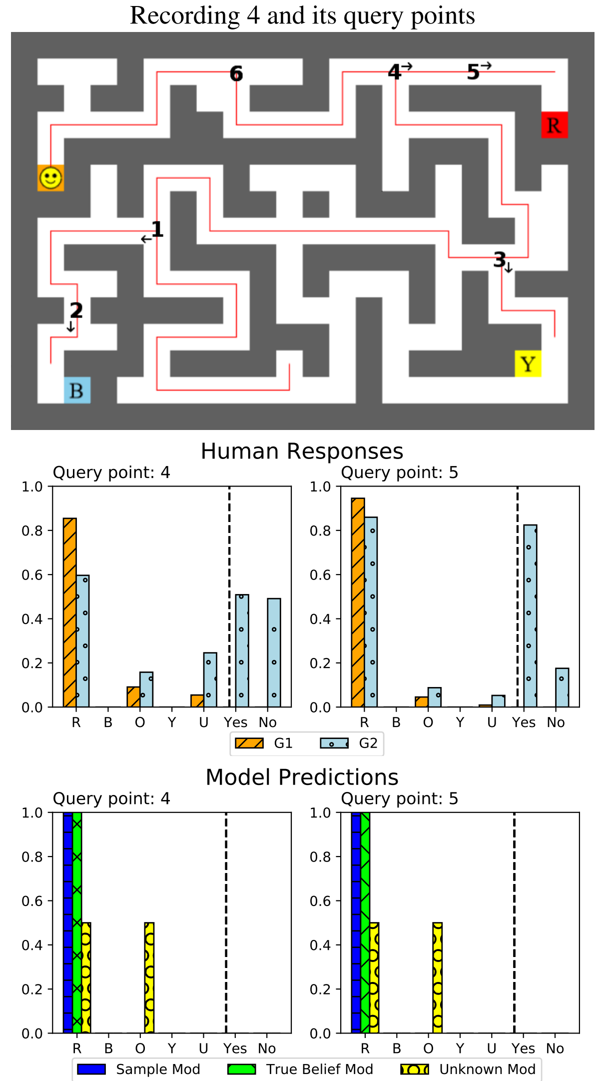

Figure 4: Relative response frequencies by the participants in both groups for Recording 4, query points 4 and 5 and the corresponding predictions made by the models using the likelihood modification.

towards assuming that the agent is seeking the destination behind the closest exit. This also holds true for points at which an agent's behavior was optimally directed to multiple exits (cf. Yellow responses for QP 3 in figures 3). We also find that participants ignore that the agent may have a knowledge state different from their own. For example, in the Red responses at QPs 4 and 5 in figure 4, they had already seen the agent turning away from two exits. Thus they should assume that the agent does not know which exit leads to which destination, even if they see the agent moving towards Red. They thus show an egocentric tendency in their reasoning.

When looking at the second group, i.e. participants that were first asked about the knowledge state of the agent before trying to identify the agent's desire, we find significantly different desire response distributions for 12 of the 22 QPs ($p \ll 0.05$ according to Fisher's exact test). These differences manifest themselves primarily in a difference of the percentage of Uncertain responses, which is significantly higher for the second group (14.9% vs 4.7%, $t = 3.86$ $p < 0.001$), also visible at QP 3 in figure 3 and QP 4 in figure 4. We further find interesting results regarding the preceding question about their belief of the agent's knowledge state: "No" responses (i.e. they believe the agent does not know) increase after an agent turned away from an exit it saw, as expected (e.g. in recording 4 "No" responses are only around 20% at QP 2, but increase to around 45% at QP 3). However, the percentage for "No" never exceeds 55% and quickly decreases again as the agent moves towards any particular exit, as seen for QP 5 in figure 4.

Computational Modeling

These results indicate that our participants in groups 1 and 2 performed their ToM reasoning differently, but neither group appears to make use of their ToM capabilities to their fullest extend. In fact, participants may even employ different strategies at different parts of the recordings. In this section we explore different models for the desire ToM task and present how they correlate with our empirical data.

BToM models The first two models we are considering, which are taken from previous work (Pöppel & Kopp, 2018), follow the general BToM framework and were designed to correspond to the mental states induced by the two conditions described above.

The *True Belief* model assumes that the agent has full knowledge regarding which exit leads to which destination:

$$P(a_{t+1}|\boldsymbol{a}_t) = \sum_{d \in D} P(a_{t+1}|d, b_d^*) P(d|\boldsymbol{a}_t) \tag{1}$$

The *Unknown* model assumes that the agent does not initially know which exit leads to which destination, which is why it needs to consider all possible combinations:

$$P(a_{t+1}|\boldsymbol{a}_t) = \sum_{\substack{d \in D \\ b_d \in B_d}} P(a_{t+1}|d, b_d) P(b_d|\boldsymbol{a}_t) P(d|\boldsymbol{a}_t) \tag{2}$$

With D being the set of desirable destinations and B_d the set of beliefs about which exit leads to which destination. b_d^* is the true belief, i.e. it maps exits to destinations correctly. $\boldsymbol{a}_t = a_1, .., a_t$ is the sequence of past actions, observed up to this point t.

The likelihood $P(a_{t+1}|d, b_d)$ is modeled following the commonly used Boltzmann noisy rationality:

$$P(a_{t+1}|d, b_d) = \frac{exp(\beta U(a_{t+1}, b_d, d))}{\sum_{a_i \in A} exp(\beta U(a_i, b_d, d))} \tag{3}$$

with β specifying the degree of rationality. Low values of β will allow more sub-optimal actions, while a larger β will result in the probability mass to be concentrated on the action with the highest utility $U(a_{t+1}, b_d, d)$ which in this simple scenario can be equated to the remaining distance to the exit leading to d according to belief b_d after executing the action a_{t+1}. The belief b_d is updated when the agent actually sees one of the exits by dismissing any beliefs which do not conform to the evidence.

Simple sampling model As a third model we introduce a model correlating to shallow processing with egocentric tendencies by implementing a very naive sampling approach: At the start of the recording, the model samples one destination from the prior $P(D)$. After observing each action, its likelihood $P(a|d, b_d^*)$ is computed using eq. 3. We keep this sample with the probability of the likelihood. Conversely, we draw a new sample with probability $1 - P(a|d, b_d^*)$ again from the prior $P(D)$, while ensuring not to pick a previously discarded destination. This way, the worse a sample can predict the observed actions the more likely it is to be replaced. Once all destinations have been discarded, we are considering all of them again, as we must have discarded the correct one along the way. The prior $P(G)$ is computed every time we need to draw a sample and depends on the remaining distance between the agent's current position p and the destination:

$$P(d) \propto exp(-\beta dist(p, d)) \tag{4}$$

For our results presented below, we fit β in the range of 0.1 to 3 at 0.1 intervals via a grid search to maximize correlations for each model separately.

Modifications As we are interested in what kind of models are required to model different ToM reasoning strategies employed by humans, we further tested the following modification to the likelihood function (eq. 3) in order to be able to better reflect the biases found in our data. While these modifications may improve the correlations in this case, we note that they may actually decrease correlation with human judgments that employ more thorough ToM reasoning.

To better reflect the bias for the closest exit found in the data, we changed the rationality constant β to a dynamic variable, which decreases with the distance to the exit, effectively dampening the likelihoods for exits that are further away and boosting optimal actions towards closer exits.

$$\beta \propto \alpha exp(-\gamma dist_m(p, d)) \tag{5}$$

where $dist_m$ is the current Manhattan distance between the agent's position p and the considered destination d. In this case α and γ are meta parameters that were fit to maximize correlation with a grid search between 2 and 4 at 0.1 intervals for α and between 0.025 and 0.75 at 0.05 intervals for γ for the results.

Model evaluation We compare our models with our participants' responses both on each recording separately, as well as over all responses. As has been done in previous BToM research (e.g. (Baker et al., 2017)) we considered the correlations between participants' average responses and the models' predicted distributions at each of the different query points. For this we stack the relative response frequencies for the four possible destinations for all QPs within a single recording, resulting in a vector of $4 \times 6 = 24$ elements (16 for recording 3 as there were only 4 QPs). Likewise we stack the destination distributions predicted by our models before computing the Pearson's r correlation. For the sampling model, we generated 100 independent responses and used the resulting normalized frequencies as the model's distribution. We then further stack the vectors for all recordings for the overall comparison, yielding a vector with 88 elements. We are deliberately evaluating in favor of our models in order to consider a best case scenario: All meta-parameters (β, α and γ)

have been fit to maximize the resulting correlation across all recordings. Furthermore, we spread all Uncertain responses across the other alternatives proportional to the model's distribution.

In order to test how well the models match the participants individually, we further had the models create actual predictions and compared these to the responses of each participant. We sampled 100 discrete responses from our models' predictions and computed how often these responses match the participants' responses at each of the different query points. We then averaged these number of matches over all query points for each recording and over all participants to get the average matching performance of our models. Again, in order to evaluate in favor of the different models, we count Uncertain responses as matches.

Tables 1 and 2 summarize the resulting correlations as well as the average number of matching responses (values in brackets) between our models and their modifications with the human responses of the first and second group respectively. Missing correlation values (–) are due zero variance in predictions of the Unknown model in those recordings. Note that Recording 3 contained only 4, instead of 6 query points.

Exemplary model outputs can also be found in figures 3 and 4, which shows the response distributions of the different (modified) models for the same QPs as the human responses.

The first thing to note is that the Unknown model, being the most rational with the least amount of biased assumptions, performs significantly worse than all others. This holds true for both the average correlation, as well as the number of matches. In Recordings 1 and 3 where we cannot compute the correlation due to zero variance, the Unknown model fails to make any predictions, always yielding a uniform distribution, which turns the Unknown model into a random model when comparing response matches. The slightly higher than chance performance of the Unknown model can mostly be attributed to the U responses. Furthermore, we find that the Sampling model correlates best with our human data, with a significant difference to the True Belief model. These results are also reflected by the average number of response matches. All models without the modification, except for the Unknown model correlated significantly more with participants in group 2 than in group 1. With regard to the modifications introduced by eq. 5: The True Belief model can improve its correlation significantly for both groups, while the Sampling model only improves significantly for the first group. Finally, it is noteworthy that the best meta-parameters for the Sampling model differ quite strongly between the two groups. (All significance claims achieved $p < 0.05$ on a t-test using the correlation coefficients after employing a Fisher transformation.)

Discussion and Conclusion

The results reported here suggest that humans can deviate quite strongly from optimal ToM reasoning. The rare use of the U(ncertain) response overall indicates that participants do not always consider the likelihood of all valid alternatives, but rather focus on single alternatives. In particular, they often fail to give U responses even after it became apparent that the agent is not aware of the location of the desired destination. In contrast, optimal reasoning would dictate the use of U responses whenever more than one destination is the most probable, or whenever multiple alternatives have a non-zero probability. Instead, participants show egocentric tendencies by ascribing their own map knowledge to the agent, and moreover a strong bias towards the closest exit as destination. This is also reflected in the decrease of "No" responses in the second group as soon as the agent moves towards any exit: even participants that briefly suppressed this tendency after having observed the agent moving away form a seen exit, tend to discard this evidence again at the next QPs. The results of the second group indicate that posing a question about the mental state of the agent before requesting the desire inference, increases the number of considered ToM alternatives slightly. Still, even participants of the second group that correctly realised that the agent's knowledge state differed from their own, often did not account for it properly when reasoning about the desire of uncertain agents. These findings support the hypothesis that humans may perform ToM reasoning differently. The task to give likelihood ratings for all alternatives (as e.g. in (Baker et al., 2017; Jern et al., 2017)) might evoke more controlled and complex ToM reasoning, suppressing cognitive biases and resulting in good correlations with optimal Bayesian models.

One might object that the observed bias towards the closest exit may stem from interpreting the instructions as "*where do you think the agent is currently going?*". However, the actual instruction was deliberately chosen to prevent this interpretation by stating "*Please specify which destination you think the agent wants to reach after leaving the maze*." While we cannot be certain about the actual interpretation by participants in the online study, we do believe that the biases are more likely to originate from inherent tendencies to use simpler, less demanding mentalizing strategies.

Looking at the correlations with different computational models, we find only comparatively weak correlations of the Unknown model with the empirical data, indicating that participants' responses are quite different from optimal Bayesian reasoning. Instead, the exhibited egocentric tendencies and biases are matched better by the True Belief and Sampling models. The better correlations of the Sampling model compared to the True Belief model can be attributed to the fact that the True Belief model compares all alternative destinations equally, while the Sampling model sticks with the first best guess, which conforms to a closeness bias, as long as it is not invalidated. When introducing likelihood modifications that shift the focus to the closer exits, the True Belief model starts to behave similarly. The lower difference between the correlations with the True Belief models and the Sampling models in Group 2, as compared to in Group 1, also indicates that priming participants with an explicit ToM-related question reduced these biases.

Table 1: Average correlations and number of response matches (in brackets) of models with ratings of Group 1.

Model	Recording 1	Recording 2	Recording 3	Recording 4	Overall
True Belief ($\beta = 0.3$)	0.85 (4.73)	0.95 (2.43)	0.68 (3.84)	0.85 (5.24)	0.85 (4.06)
True Belief Mod ($\alpha = 2.5, \gamma = 0.125$)	0.98 (4.81)	0.99 (2.82)	0.89 (4.02)	0.87 (5.23)	0.93 (4.22)
Unknown ($\beta = 1.9$)	– (4.52)	0.30 (2.12)	– (3.41)	0.62 (4.40)	0.40 (3.61)
Unkown Mod ($\alpha = 2.5, \gamma = 0.025$)	– (4.87)	0.30 (3.03)	– (4.22)	0.62 (4.95)	0.40 (4.27)
Sampling ($\beta = 1.9$)	0.94 (1.89)	0.96 (1.18)	0.82 (1.85)	0.99 (3.03)	0.94 (1.99)
Sampling Mod ($\alpha = 3.7, \gamma = 0.125$)	0.98 (1.94)	0.98 (1.18)	0.96 (1.73)	0.98 (2.95)	0.98 (1.95)

Table 2: Average correlations and number of response matches (in brackets) of models with ratings of Group 2.

Model	Recording 1	Recording 2	Recording 3	Recording 4	Overall
True Belief ($\beta = 0.3$)	0.95 (5.09)	0.96 (4.75)	0.76 (2.75)	0.93 (3.99)	0.91 (4.15)
True Belief Mod ($\alpha = 2.5, \gamma = 0.125$)	0.98 (4.95)	0.98 (4.73)	0.94 (2.84)	0.94 (4.21)	0.96 (4.18)
Unknown ($\beta = 1.7$)	– (4.57)	0.34 (4.50)	– (2.42)	0.72 (3.82)	0.45 (3.83)
Unkown Mod ($\alpha = 2.5, \gamma = 0.075$)	– (4.98)	0.34 (4.82)	– (3.14)	0.72 (4.41)	0.45 (4.34)
Sampling ($\beta = 0.7$)	0.97 (3.36)	0.97 (2.40)	0.89 (1.33)	0.98 (2.19)	0.96 (2.32)
Sampling Mod ($\alpha = 2.7, \gamma = 0.075$)	0.98 (3.23)	0.97 (2.44)	0.99 (1.28)	0.97 (2.32)	0.97 (2.32)

Overall, the actual ToM reasoning of humans appears to be more differentiated than assumed in the BToM literature. Mental reasoning is computationally expensive, especially when considering mental states of others. Unless explicitly triggered, humans appear not to perform a full-blown ToM reasoning but to resort to simpler heuristics instead. Artificial social systems can make use of these findings by adapting to different ToM models employed by their users and assisting when they might overlook important information.

References

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*, 0064.

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. doi: 10.1016/j.cognition.2009.07.005

Dennett, D. C. (1989). *The intentional stance*. MIT press.

De Weerd, H., Verbrugge, R., & Verheij, B. (2013). How much does it help to know what she knows you know? an agent-based simulation study. *Artificial Intelligence*, *199*, 67–92.

Haselton, M. G., Nettle, D., & Murray, D. R. (2015). The evolution of cognitive bias. *The handbook of evolutionary psychology*, 1–20.

Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science*, *344*(6190), 1243091.

Jara-Ettinger, J., Baker, C., & Tenenbaum, J. (2012). Learning what is where from social observations. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34).

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other peoples preferences through inverse decision-making. *Cognition*, *168*, 46 - 64. doi: https://doi.org/10.1016/j.cognition.2017.06.017

Keysar, B. (2007). *Communication and miscommunication: The role of egocentric processes.* Walter de Gruyter.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25–41.

Knill, D. C., & Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, *27*(12), 712–719.

Nakahashi, R., & Yamada, S. (2018, July). Modeling human inference of others' intentions in complex situations with plan predictability bias. In J. Z. Chuck Kalish Martina Rau & T. Rogers (Eds.), *Cogsci 2018* (pp. 2147–2152).

Nickerson, R. S. (1999). How we knowand sometimes misjudgewhat others know: Imputing one's own knowledge to others. *Psychological bulletin*, *125*(6), 737.

Pöppel, J., & Kopp, S. (2018). Satisficing models of bayesian theory of mind for explaining behavior of differently uncertain agents: Socially interactive agents track. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems* (pp. 470–478).

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*(4), 515–526.

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child development*, *75*(2), 523–541.