Less Egocentric Biases in Unsolicited Theory of Mind When Observing Agents in Unbalanced Decision Problems

Jan Pöppel
Social Cognitive Systems Group
Bielefeld University, Germany

Stacy Marsella
Centre for Social, Cognitive and Affective Neuroscience
Glasgow University, Scotland, UK

Stefan Kopp
Social Cognitive Systems Group
Bielefeld University, Germany

Abstract

Theory of Mind (ToM) or mentalizing is the ability to infer mental states of oneself and other agents. Theory of mind plays a key role in social interactions as it allows one to predict other agents' likely future actions by inferring what they may intend or know. However, there is a wide range of ToM skills of increasing complexity. While most people are generally capable of performing complex ToM reasoning such as recursive belief inference when explicitly prompted, there is much evidence that humans do not always use ToM to their full capabilities. Instead, people often fall back to heuristics and biases, such as an egocentric bias that projects one's beliefs and perspective onto the observed agent. We explore which (internal or external) factors may influence the mentalizing processes that humans employ *unsolicitedly*, i.e., employ without being primed or explicitly triggered. In this paper we present an online study investigating unbalanced decision problems where one choice is significantly better than the other. Our results demonstrate that participant's are significantly less likely to exhibit an egocentric bias in such situations.

Keywords: Theory of Mind; Egocentric bias; Behavior prediction;

Introduction

Mindreading, often also called Theory of Mind (ToM), is the ability to infer the mental states of oneself and others (Premack & Woodruff, 1978). This ability allows us to infer the intentions and beliefs that determine other agents' actions, thus making their actions more understandable. ToM is crucial for social interactions, both cooperative as well as competitive, as it allows the prediction of other agents' future actions, allowing one to react and adapt appropriately.

Over the past decades, a lot of research has gone into understanding the development and limits of human's ToM capabilities. While it is not yet clear if any or how many of these abilities are innate or developed during childhood, different levels of sophistication in people's ToM capabilities have been discovered: Children learn to first infer another agent's desires before their (potentially false) beliefs and finally their emotions (Wellman & Liu, 2004). Recently, Bayesian Theory of Mind (BToM), which is based on inverse planning in causal probabilistic models, has been very successful in matching people's mentalizing responses for a range of different scenarios and mental states, including intentions, beliefs, preferences but also emotions (C. Baker, Saxe, & Tenenbaum, 2011; C. L. Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Ullman, Tomer, Baker, Macindoe, Goodman, & Tenenbaum, 2009; Jern, Lucas, & Kemp, 2017; Velez-Ginorio, Siegel, Tenenbaum, & Jara-Ettinger, 2017). These studies usually explicitly query participants to make judgments regarding an observed agent's mental state in a particular situation and compare these judgments to the model's predictions. Mental inferences can even be recursive up to several levels, albeit with increasingly more errors (Dunbar, 1998).

However, at the same time, there is substantial research documenting cases where people do not employ their mentalizing capabilities to their fullest extent and/or fail to accurately infer another agent's mental perspective. These findings emerge primarily in scenarios where participant's were asked to predict behavior or act themselves without being explicitly queried about the mental state of the other agent (e.g. (Keysar, 2007; Pöppel & Kopp, 2019)). We will refer to such mental processes that people employ without being primed or explicitly triggered as *unsolicited*[1]. It is unclear to what extend participants in these *unsolicited* mentalizing scenarios actually perform explicit reasoning. The mentioned findings may also be the result of a lack of explicit reasoning. However, we conjecture that adults can "choose", potentially automatically, to employ explicit or deliberate mentalizing processes *unsolicitedly*.

Studies with such *unsolicited* mentalizing have shown that participants appear to be likely to employ heuristics resulting in biases, such as an egocentric bias where participants project their own mental state onto the observed agent, even if they have sufficient evidence to infer that the observed agent's mental state should be different. For example, the general BToM framework, which models an ideal observer, has been adapted by Nakahashi and Yamada to account for situations in which participants were less primed to consider the actual mental states of agents (Nakahashi & Yamada, 2018). Pöppel and Kopp, similarly, had to modify a BToM model to make strong use of visual heuristics to best fit participants' predictions regarding an observed agent's future actions (Pöppel & Kopp, 2019). Interestingly, they found that asking a question related to the observed agent's knowledge, before asking for a prediction, would significantly influence people's predictions to be closer to predictions made by an unmodified BToM model.

Other evidence for heuristics or biases in mentalizing can

[1] This should not be confused with implicit social cognition which is argued to be present from a very young age (Frith & Frith, 2008).

be found in the work by Keysar: He observed that participants in an object naming task did not take the perspective of their interlocutor into account correctly (Keysar, 2007). Participants would often make decisions based on their own perspective instead of the known different perspective of the other agent. While criticism regarding these findings has been raised (Hawkins & Goodman, 2016), other studies such as (Pöppel & Kopp, 2019; Nakahashi & Yamada, 2018) provide similar evidence in different scenarios. The study presented here will also provide additional evidence for egocentric biases in *unsolicited* mentalizing.

These diverging findings raise the question about the cause for these different "modes" of mentalizing. One explanation could be that there are differences in the complexity between different ToM processes which may incur considerable computational demands. One can assume that mentalizing, much like general decision-making, is subject to bounded rationality (Simon, 1955), i.e. it needs to take into account not only the external factors but also internal ones such as mental resources. Following the "resource rational analysis" paradigm by Lieder et al. (Lieder & Griffiths, 2019), it is only rational to employ heuristics to approximate mental inferences in face of limited resources depending on the context and the involved costs. Under this assumption, it is likely that egocentric biases arise from the use of heuristics to simplify the mentalizing task, e.g. by not explicitly inferring the other's belief state but instead assuming it is identical to one's own.

We generally assume that *unsolicited* mentalizing is subject to a (often automatic) choice that takes the costs and benefits of different ToM processes into account. This process needs to balance the computational costs of more complex inferences against the higher accuracy of the inferred mental states. In this way, we would expect a poker player in the finals of a tournament to employ much more elaborate ToM processes to infer the mental state of their opponent than a person just taking a walk would employ to infer the mental state of other pedestrians she needs to avoid. In the former case, the inferred state can directly determine their tournament placement, while failing to (correctly) infer the other's mental state would in the worst case result in a slightly awkward situation in the latter case. While it may be obvious that there are differences in the mentalizing employed in these very different situations, it is unclear what factors (or "costs"), both externally and internally, influence a person's *unsolicited* mentalizing processes in general. Likewise, it is unclear how significant these factors need to be before a different, more complex mentalizing process is employed that, in turn, may change the inferred mental state.

In this paper we present first results towards uncovering the role of different factors for different "modes" of mentalizing. In particular, we conducted a study on the effects of the observed agent's decision situation on the observer's mentalizing. More specifically, we look at how the relation between the (subjectively) expected utilities of two alternative actions an observed agent may choose from, influences the predictions that human observers make for it. In this situation, participants will need to employ a ToM to assess utilities from the agent's perspective. We assume that participants are likely to exhibit egocentric biases when performing *unsolicited* mentalizing, as projecting their own perspective is easier then actually inferring another agent's mental state. Yet, we also hypothesize that participants will perform more elaborate mentalizing, less susceptible to an egocentric bias, in situations where the observed agent is forced to decide between two *unbalanced* options, i.e. in a scenario where the expected utilities for the two options differ greatly.

Method

In order to be able to draw conclusions with regard to possible factors influencing *unsolicitated* mentalizing, the design of the presented study had to fulfill three conditions:

1. Judge mental state inferences without priming participants with respect to the mentalizing they employ explicitly.
2. Control alternative factors that may influence participant's mentalizing.
3. Account for strong subjective differences between people's mentalizing capabilities and tendencies.

The stimuli, procedure and used conditions outlined below address these three requirements. In particular, previous research with the BToM framework has shown that directly querying participant's inferred mental state is likely to prime them to employ more sophisticated mentalizing. Indirect measurements for participant's inferred beliefs, here in the form of predictions for the observed agent's next action, need to be used instead. The chosen stimuli was designed to allow these kinds of predictions while keeping the second requirement in mind.

Stimuli

Behavior in complex domains can be explained by multiple alternative mental states. In order to be able to study the effect of a particular factor on participant's mentalizing (requirement #2), we needed a very simple domain (cf. (C. Baker et al., 2011)). The domain needs to limit or at the very least control any other factors that may influence participant's mentalizing. After much experimentation, we settled on a simple 2D home scenario (see Figure 1). For this study, the agent living in the home can have a belief regarding the position of a book. The book can be in any of three bookshelves (represented as green squares in Figure 1).

Participant's would be shown a trajectory of the agent looking for the book. The agent would take a rational path depending on its mental state, e.g. if the agent knew that the book is located in the top left bookshelf it would go there directly. For the study reported here, the agent did not know the books location. Furthermore, the book is always located in the bookshelf that is furthest away from the agent's starting position, resulting in a search behavior that checks all three

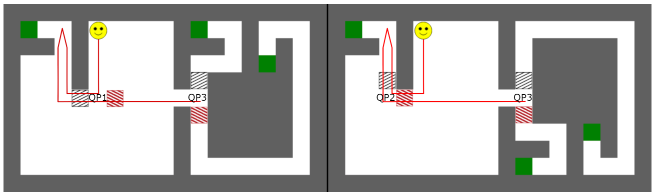

Figure 1: Overview of the used environment in two orientations (third bookshelf *DOWN* on the left and *UP* on the right). The red line indicates the agent's trajectory. At three query points (QP), participants have to predict the agent's next move (QP1 and QP3 shown on the left; QP2 and QP3 on the right). The possible predictions (Black or Red tile) were only presented when the agent reached that particular QP.

bookshelves in a rational order. The agent's behavior in combination with the structure of the home environment were set up such that we can make judgments about the inferred mental state based on participants' predictions (requirement #1): From the point of view of an observer, the agent's behavior should only depend on the agent's belief regarding the book's location once its desire and competence are established: The agent could either know the book's true location and would thus directly head towards the correct bookshelf. Alternatively, the agent would believe the book to be at a wrong location (false belief), in which case it would still head directly towards the wrong bookshelf before having to reconsider. Finally, the agent may not know where the book is located and a rational search behavior would be the most expected behavior. The 2nd case would also likely turn into the third after confirming that the initially targeted bookshelf was incorrect.

Procedure

Participants would be made aware of the agent's desire in order to control for possible other desires participants may attribute to the agent (requirement #2) In this case participants were told that the rational agent they are observing is trying to get a blue book as quickly as possible.

In order to allow participants to "choose" their mentalizing *unsolicitedly* instead of priming specific processes, we did not ask for predictions regarding an agent's mental state but rather regarding that agent's next action. Participants would be shown the agent move in the 2D environment and be asked to predict the agent's next action at different query points (QP) along the trajectory.

For this study, participants were given only two options for the agent's next actions which were highlighted on the environment next to the agent at each QP (Black and Red tiles in Figure 1). Participants would then choose one of the two options with 3 different certainty ratings each: Certainly, Very Likely, Likely. Alternatively, participants could state "I do not know" if they did not want to commit to either option, resulting in a 7-point Likert scale.

In this study we used the three QPs along the trajectory shown in Figure 1 which were chosen deliberately: The first one (QP1) is an initial check right at the start before participants could realistically infer the agent's mental state. This serves to measure any prior assumptions participants may make about the agent. QP2 followed closely after and serves primarily as an attention check as well as to ensure participants realize that the agent first goes towards the left.

The third and final QP (QP3) was the important one that we argue allows the correlation between participants' prediction and inferred mental state: As can be seen in Figure 1, the third QP forces a decision at the intersection between the second and third bookshelf, after the agent had already checked the first one. A rational agent should usually check the closer bookshelf first, if it did not know the book's location. A prediction for the longer path assumes that the agent knows that the book is located there. If a participant predicts the agent to take the longer path, they either did not infer the mental state of the agent correctly, or did not take the agent's mental state into account when making the prediction.

Conditions

In order to determine if the agent's situation, in particular the expected utilities for its possible actions, can influence an observer's ToM process, we varied the distance towards the third bookshelf that always contained the book. A longer distance would result in a higher cost if the agent were to go there in vain thus reducing the expected utility for that option. In the *SHORT* condition, the final bookshelf is only marginally further away from the intersection than the 2nd one. In the *LONG* condition, the difference is very significant as can be seen in Figure 2.

In order to account for subjective differences in participant's ToM reasoning (requirement #3), we constructed a "baseline" condition by varying the book's visibility for the observing participants, as can be seen in Figure 2. In the *HIDDEN* conditions, participants would not see the book's location until the agent actually found the book, i.e. they effectively had the same mental state regarding the book's location as the agent by design. Conversely, participants in the *VISIBLE* condition could directly see the book's location from the start of the trajectory. Since participant's knowledge is identical to the agent's in the *HIDDEN* condition, we would assume that there cannot be a directed egocentric bias of projecting their belief regarding the book's location onto the agent. In the *VISIBLE* condition, however, the participants' knowledge differs from the agent's and if they, in this case incorrectly, project their mental state onto the agent, they would be more likely to predict the agent to go directly towards the bookshelf containing the book. By considering the difference between these two visibility conditions, we can measure the influence our *distance* variable has on participant's egocentric bias while accounting for individual differences. With these conditions, the following hypotheses can be formulated:

- Participants should be more likely to predict the agent to go directly towards the book's location in both the *SHORT* and *VISIBLE* conditions as compared to their counterparts.

In the case of the *VISIBLE* conditions this would show an egocentric bias.

- The differences in participants' responses between the two visibility conditions should be larger in the *SHORT* conditions than in the *LONG* conditions.

Additionally, we introduced two variants of the *SHORT* and *LONG* conditions where we flipped the right hand side of the environment as a counter-balancing variable. You can see the flipped version of the *LONG* conditions in Figure 1. This was intended to counter-balance any effect the visual location of the bookshelves may have.

Overall, these variables constitute a 2x2x2 between-subject design. While within-subject studies allow us to better filter out individual differences, we realized in earlier experiments that participants were often influenced by what they saw in previous conditions when encountering subsequent ones. This influence was even present when introducing "new" agents and environments.

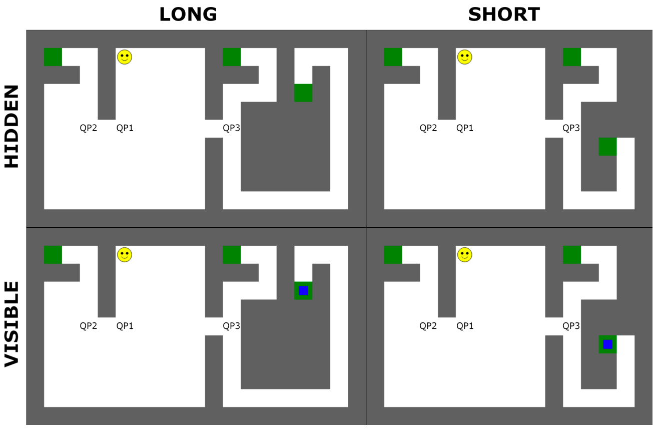

Figure 2: Overview of the layouts for the different conditions in the *DOWN* orientation from the participant's point of view. Each query point (QP) only became visible once the agent reached that tile.

Instructions

At the start of the study, participants were told that they were going to see "a rational agent, shown as a smiley, navigating in his home, a place it knows well." On the instruction screen participants would see the home environment corresponding to their distance condition, as in the top two images in Figure 2, but without an agent and the QPs. Participants were further told, that a friend brought a blue book and placed it in one of the three bookshelves in the agent's house and that the agent was now "in urgent need of the book", stating that the agent wants to get the book "as quickly as possible". We emphasized the agent's rationality as well as their urgency to limit the possible explanations participants could come up with regarding the agent's behavior. The colored squares were introduced as the bookshelves in question and an image of the book was introduced. Note that only participants in the *VISIBLE* conditions did actually see the book. We then informed the participants that they will be asked to predict the agent's next action while they observed the agent.

We deliberately did not state whether or not the agent knew in which bookshelf the book was placed as this belief was to be inferred by the participants.

Participants

After eliminating participants that did not give responses at all QPs, i.e. that aborted the study, we ended up with 242 participants from Amazon's Mechanical Turk. These participants were randomly assigned to the eight different conditions resulting in 31 participants in the *VISIBLE, LONG, DOWN*, 30 in the *HIDDEN, LONG, DOWN*, 36 in the *VISIBLE, LONG, UP*, 30 in the *HIDDEN, LONG, UP*, 29 in the *VISIBLE, SHORT, DOWN*, 26 in the *HIDDEN, SHORT, DOWN*, 30 in the *VISIBLE, SHORT, UP* and 30 in the *HIDDEN, SHORT, UP* condition. We only ordered participants that had a rate of completed HITs greater than 80%. Participants were financially compensated.

Results

The rating options were always presented to participants in the same order ("Certainly Black", "Very Likely Black", "Likely Black", "I do not know", "Likely Red", "Very Likely Red" and "Certainly Red") which we coded from 1 to 7 for numerical analysis. The orientation condition only affects the 3rd QP. We converted the responses for *UP* orientation so that they represent the same meaning with respect to the actual book location as in the *DOWN* orientation for the results presented here. This way, values from 1 to 3 (or involving "Black") always correspond to predictions towards the next closest bookshelf (away from the actual location of the book), while values of 5 to 7 ("Red") correspond to predictions towards the book's actual location for all query points.

Since neither the non-parametric Kruskal-Wallis test (H=2.24, p>0.1) nor a chi-square test on the Likert values ($\tilde{\chi}^2(5, N = 242) = 3.317, p = 0.673$) revealed a significant difference between participants' predictions in the two orientations (*UP* and *DOWN*), we collapsed the data and only report on the effects for the *visibility* and *distance* variables.

Figure 3 presents the average Likert scores for the different QPs by their conditions. The figure reveals strong differences between the *visibility* conditions for the QP1, most notably in the *LONG* condition: A Scheirer-Ray-Hare-Test (SRH) (an extension of the Kruskal-Wallis test to also take interactions into account) revealed significant main effects for both *visibility* (H=24.72, p<0.0001) and *distance* (H=7.21, p<0.01) as well as a significant effect for the interaction (H=18.50, p<0.0001). In order to better understand the interaction, we tested the effect both factors individually on the two sub-groups of the other factor, e.g. the effect of *visibilty* in the *SHORT* conditions. A Kruskal-Wallis test on these subgroups revealed significant differences for *visibility* in both *distance* conditions (H=22.88, p<0.0001 for *LONG*; H=5.91,

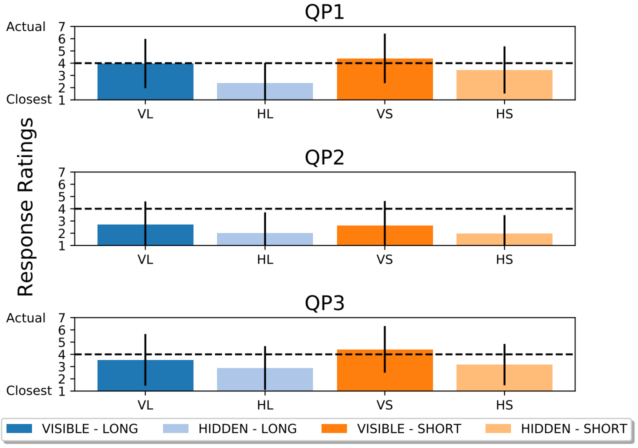

Figure 3: Average response values and their standard deviations for the different query points and conditions. The dashed line represents the “I do not know” response. Values below 4 correspond to predictions towards the next closest bookshelf, while those above towards the book’s actual location.

$p<0.02$ for *SHORT*)). However, *distance* only has a significant effect in the *HIDDEN* conditions ($H=10.92$, $p<0.001$).

For QP2 the SRH test reveals a main effect for *visibility* ($H=7.65$, $p<0.006$) and the interaction ($H=18.48$, $p<0.0001$) but not for *distance*. The Kruskal-Wallis test on the subgroups only reveals a significant difference for *visibility* in the *LONG* conditions ($H=5.721734$, $p<0.02$).

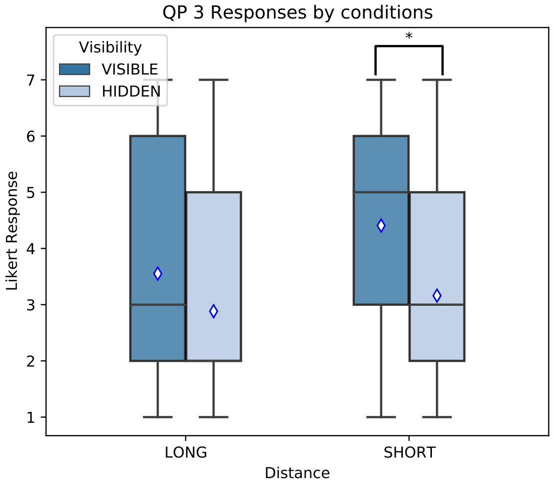

Figure 4: Box-and-whisker plot including the means for the responses for QP3 by conditions. The * symbolizes the significant difference for *visibility* in the *SHORT* conditions (Kruskal-Wallis $H=11.95$, $p<0.0006$).

The critical QP3 does show some noteworthy differences again between the *visibility* and *distance* conditions in Figure 3.

The SRH test revealed significant main effects for both *visibility* ($H=13.35$, $p<0.0004$) and *distance* ($H=5.66$, $p<0.02$) as well as for the interaction ($H=20.04$,$p<0.0001$). Regarding the sub-groups tests for this third QP, we find significant differences for *visibility* in the *SHORT* conditions (Kruskal-Wallis $H=11.95$, $p<0.0006$). Similarly, we find a significant difference for *distance* in the *VISIBLE* conditions (Kruskal-Wallis $H=4.727685$, $p<0.03$). This means, that the differences between participants’ responses in the *VISIBLE* and *HIDDEN* conditions were only significant in the *SHORT* conditions, not in the *LONG* ones. This interaction can be seen in the boxplot for the third QP shown in Figure 4.

So far we have looked at significant differences in participants responses on the Likert scale. The differences found so far could, however, have been differences in certainty or confidence for the same bookshelf instead of resulting in actual differences in the predicted bookshelf. In order to test this possibility, we combined all predictions towards a particular direction (towards the *closest* bookshelf or book’s *actual* location). Figure 5 shows the resulting relative frequencies of participant’s responses. Performing a statistical analysis on this collapsed data for the critical QP3, we find the same significant effects as before: Main effects for both *visibility* ($H=9.50$,$p<0.003$) and *distance* ($H=4.24$,$p<0.04$) and the interaction ($H=20.54$,$p<0.0001$). Likewise the pairwise comparison also reveals the same effects: We find a significant effect for *visibility* only in the *SHORT* conditions ($H=10.911$, $p<0.001$) and for *distance* only in the *VISIBLE* conditions ($H=6.90$,$p<0.001$).

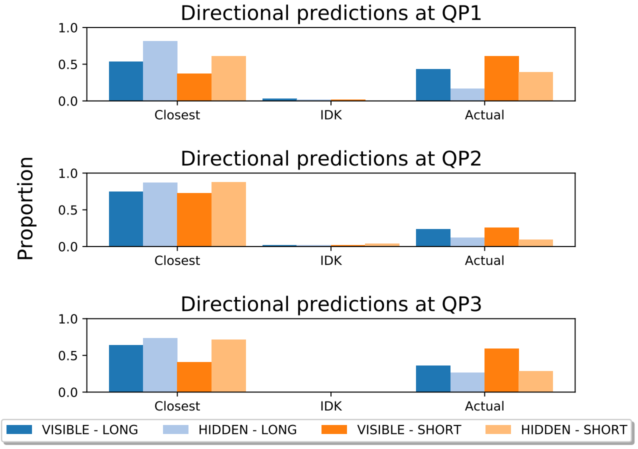

Figure 5: Proportion of participant’s predictions when grouped by their predicted direction (towards the *closest* bookshelf or towards the book’s *actual* location). No participant chose the “I do not know” option at QP3.

Discussion

The results for QP1 give additional evidence to earlier findings of an inherent egocentric bias in *unsolicited* mentalizing. A significant portion of participants predicted that the agent would go towards the right despite the left bookshelf being closer when they knew the book's location in the *VISIBLE* conditions. This difference was not influenced by the distance condition. Participants in the *HIDDEN* condition were, however, more likely to predict the closest bookshelf in the *LONG* condition compared to those in the *SHORT* condition. However, an initial egocentric bias may actually be reasonable in this situation: Participants did not know whether or not the agent would know the book's location and the agent's behavior up to this point did not allow for any clear inference regarding that question. An alternative explanation could be that there are two bookshelves on the right and only one on the left. While this did not appear to make much of a difference for the *HIDDEN* visibility conditions, it may influence some participants to lean towards that direction.

At QP2 the majority of participants predicted the agent to go towards the closest bookshelf with certainty. This is not surprising, as the agent clearly headed towards that location. This condition primarily serves as an attention check to see if participants inferred the intent of the agent at all. The significant difference for *visibility* does indicate, that the few participants that predicted the agent would go back towards the right side of the home, primarily belonged to the *VISIBLE* condition.

The critical QP3 reveals interesting differences between the conditions in line with our hypotheses. Both variables *distance* and *visibility* have significant main effects on participants' responses by themselves. The main effect for *distance* confirms our first hypothesis: Participants were more likely to predict the agent to go to the closer bookshelf, if the distance to the other one is larger. These results are in line with what we would expect from rational decision-making, as the costs for having to turn around is larger if one takes the long route without finding the book there. At the same time, such a response would be expected if participants employ visual heuristics for choosing the likely next target, even without employing mentalizing about the other agent.

Conversely, participants were more likely to predict the agent to directly take the longer path when they knew the book is located there. This is also in line with rational decision-making, assuming the agent knew where the book was located. Yet, in all conditions, the first part of the trajectory up to QP3 should have made it clear that the agent did not know where the book was located. We suspect that participants in the *VISIBLE* conditions were more likely to ignore this information and instead projected their own knowledge of the book's location onto the agent, thus exhibiting an egocentric bias. Crucially, when considering the two distances separately, there is no significant difference for visibility in the *LONG* conditions but only in the *SHORT* ones. Since the significant difference between our *HIDDEN* "baseline" and target *VISIBLE* conditions vanishes in the *LONG* conditions, we can accept our second hypothesis: Higher costs for the observed agent, here in the form of a longer distance towards a potentially wrong target, influences participant's mentalizing in such a form that they are less likely to exhibit an egocentric bias.

The second significant interaction for *distance* only in the *VISIBLE* conditions also serves as a quality check for our experiment. Participants that did not know the book's location, i.e. that were in the *HIDDEN* conditions, were not significantly influenced by the different distances towards the two bookshelves. Instead, the majority predicted the agent to go the closer book first.

Our hypothesis does not make any claims about the reason for this effect. There is evidence for automatic ToM components, usually referred to as implicit mentalizing (Frith & Frith, 2008). Different studies have discovered automatic components involved in empathy. According to these studies, at least some components involved in perspective-taking operate automatically; e.g. (Li & Han, 2010; Zaki, 2014).

One explanation for this effect could be that participants take the observed situation, including the utilities or costs the observed agent faces into account for their own mental decision-making process. Similar to how people share other's feelings, the absolute costs the observed agent faces could be added to their own internal and external costs. The higher total costs based on the situation may then offset the higher costs for the more thorough mental inference process that actually infers the agent's likely mental state. If the (perceived) external costs of the observed agents are smaller using a simpler heuristic resulting in egocentric bias may be more resource-rational.

Another explanation could judge the complexity of the situation the agent faces: In the *SHORT* conditions, both options have similar expected utility if the book's location is unknown. If a participant realizes that the agent's decision does not have a large effect on the outcome, it may be more beneficial to save the mental resources by employing heuristics. Conversely, if the observed decision problem is more complex it may be worse the additional mental effort to consider the situation more carefully.

A distinction between these two explanations could potentially be made in future work by performing a similar study, that also includes a condition where both options lead to long paths but where there is only a small, if any, difference between the distances. The second explanation would assume that similar distances would be a fairly simple situation for the agent regardless of the distance, as its decision does not matter all too much. It would, therefore, not predict to find a reduction of the egocentric bias in that condition.

Conclusion

In this paper we have presented a study designed to investigate whether an observed agent's potential costs for wrong actions can influence an observer's *unsolicited* mentalizing by reducing the likelihood for making predictions with an egocentric bias. We believe that the presented study design can be used to study other potential factors that may influence the *unsolicited* mentalizing processes. The effect of an egocentric bias is measured by the difference between the baseline and target condition, thus reducing the effect of individual differences. Our results confirm our hypotheses by revealing a significant increase in egocentric bias only in the scenario where the observed agent's costs are close for both possible actions. We did not find such an effect when there is a large difference in expected utility between the actions for the observed agent. We hope that uncovering and understanding these factors will help to understand people's *unsolicited* mentalizing capabilities and potentially replicating them in intelligent systems.

References

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proc. CogSci* (Vol. 33).

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.

Dunbar, R. I. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, *6*(5), 178–190.

Frith, C. D., & Frith, U. (2008). Implicit and explicit processes in social cognition. *Neuron*, *60*(3), 503–510.

Hawkins, R. X., & Goodman, N. D. (2016). Conversational expectations account for apparent limits on theory of mind use. In *Proc. CogSci* (Vol. 38).

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's 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.

Li, W., & Han, S. (2010). Perspective taking modulates event-related potentials to perceived pain. *Neuroscience letters*, *469*(3), 328–332.

Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 1–85. doi: 10.1017/S0140525X1900061X

Nakahashi, R., & Yamada, S. (2018, July). Modeling human inference of others' intentions in complex situations with plan predictability bias. In *Proc. CogSci* (Vol. 40, pp. 2147–2152).

Pöppel, J., & Kopp, S. (2019). Egocentric tendencies in theory of mind reasoning: An empirical and computational analysis. In *Proc. CogSci* (Vol. 41, pp. 2585–2591).

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

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, *69*(1), 99–118.

Ullman, Tomer, D., Baker, C. L., Macindoe, O., Goodman, N. D., & Tenenbaum, J. B. (2009). Help or Hinder: Bayesian Models of Social Goal Inference. *Nips*, 1–9.

Velez-Ginorio, J., Siegel, M. H., Tenenbaum, J. B., & Jara-Ettinger, J. (2017). Interpreting actions by attributing compositional desires. In *Proc. CogSci* (Vol. 39).

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

Zaki, J. (2014). Empathy: a motivated account. *Psychological bulletin*, *140*(6), 1608.