No signatures of first-person biases in Theory of Mind judgments about thinking

Marlene D. Berke, Benjamin G. Sterling, Abigail L. Tenenbaum, & Julian Jara-Ettinger
Department of Psychology, Yale University, New Haven, CT, 06511
{marlene.berke, abi.tenenbaum, ben.sterling, julian.jara-ettinger}@yale.edu

Abstract

We readily get intuitions about a problem's complexity, how much thinking it will require to solve, and how long it should take, both for ourselves, and for others. These intuitions allow us to make inferences about other people's mental *processing*—like whether they are thinking hard, remembering, or merely mind-wandering. But where do these intuitions come from? Prior work suggests that people try solving problems themselves so as to draw inferences about another person's thinking. If we use our own thinking to build up expectations about other people, does this introduce biases into our judgments? We present a behavioral experiment testing for effects of first-person thinking speed on judgments about another person's thinking in the puzzle game *Rush Hour*. Although people overwhelmingly reported solving the puzzles themselves, we found no evidence for participants' thinking speeds influencing their judgments about the other person's thinking, suggesting that people can correct for first-person biases.

Keywords: Theory of Mind; simulation; social cognition; thinking

Introduction

Imagine playing chess against a friend. You advance a pawn, and then a moment later, realize that you've made a grave mistake: moving that pawn opened up your queen to an attack. You glance up at your friend's face, hoping that they haven't noticed. It seems they haven't. As you wait for their move, you study the game from their perspective: the pawn that you just advanced is now threatening their rook, and finding a safe square to move it to is proving rather tricky...

This capacity to think about other people's minds, known as *Theory of Mind*, appears to be a distinctive human capacity (Horschler* & Berke*, et al., 2023; Martin & Santos, 2016) that serves as the backbone to some of our most complex behaviors: learning and using language (Tomasello, 1992; Goodman & Frank, 2016), transferring knowledge (Gweon, 2021; Ho, Littman, MacGlashan, Cushman, & Austerweil, 2016), making socio-moral evaluations (Young, Cushman, Hauser, & Saxe, 2007; Jara-Ettinger, Tenenbaum, & Schulz, 2015), and influencing other people (Ho, Saxe, & Cushman, 2022). Theory of Mind is often conceptualized as the ability to infer other people's unobservable mental states (such as their beliefs, desires, and intentions) through an expectation that agents act rationally (Gergely & Csibra, 2003; Dennett, 1989). Consistent with this idea, computational models formalizing this process successfully capture human intuitions in a range of social tasks including how we attribute beliefs and desires (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Jern, Lucas, & Kemp, 2017), how we predict action (Jara-Ettinger, Schulz, & Tenenbaum, 2020; Baker, Saxe, & Tenenbaum, 2009), and how we engage in complex forms of social behavior (e.g., Ho et al., 2016; Bridgers, Jara-Ettinger, & Gweon, 2020; Goodman & Frank, 2016; Ullman et al., 2009).

However, we do a lot more than just infer other people's mental states like beliefs and desires. We also infer dynamic mental *processes* going on in other minds, like thinking, deciding, mind-wandering, imagining, planning, remembering, and worrying, even from an early age (Richardson & Keil, 2022). In adults, this capacity is quite nuanced, and people make surprisingly rich inferences about the processes happening in another mind just based on how long someone takes to answer a question or solve a puzzle (Berke & Jara-Ettinger, 2021; Berke, Tenenbaum, Sterling, & Jara-Ettinger, 2023; Zhang, Kemp, & Lipovetzky, 2023). To make these types of inferences, it is critical to understand the difficulty of the type of thinking someone is engaging in. For instance, the same pause will elicit different inferences depending on how complex we believe the problem is. If the problem is easy, you might think they're distracted; if it's hard you might think they're highly skilled; and if it's nearly impossible, you would suspect that they already knew the answer.

Given how central the ability to estimate problem complexity is to these inferences about other people's mental processing, this raises the question of where this ability comes from. In past work, people self-reported simulating solving problems themselves (Berke et al., 2023), suggesting a role for first-person simulation (Gallese & Goldman, 1998). And, on one hand, simulating solving the problem oneself would seem to be a good strategy—what better way to know how much thinking something requires than to think about it? But on the other hand, under some instantiations, such an approach could produce strong biases as you might inappropriately project your own patterns of thinking onto others.

Here, we investigate these questions by searching for evidence of first-person thinking in third-person judgments about what is going on in another agent's mind as they solve a puzzle. In particular, adopting the experimental paradigm and computational framework from Berke et al., 2023, we test people's inferences about how distracted someone is as they solve a puzzle. If our own first-person thinking were input directly into this inference, then faster thinkers might expect

other people to also solve puzzles quickly, and more heavily rely on distraction to explain pauses. Despite finding that participants overwhelmingly report trying to solve the puzzles themselves, we found no evidence of participants' own thinking speed influencing their judgments of distraction. This suggests that, while people may use first-person simulation to estimate the complexity of a problem, they must have some way to correct for the biases of their own thinking.

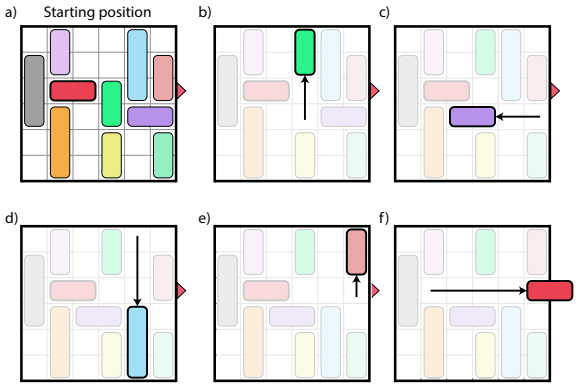

Figure 1: Rush Hour game used to study how people think about thinking. a) Example puzzle. A set of horizontal and vertical "cars" are arranged in a grid. The goal of the game is to move the red car out of the grid through the exit marked by the red triangle on the grid's right side. Cars can only move through empty squares and along their row or column. b-f) step-by-step schematic of the solution to this puzzle.

Computational Model

Domain Consider the puzzle in Fig. 1a from the puzzle game *Rush Hour*. A Rush Hour puzzle consists of a grid with non-overlapping "cars" (visualized as rectangles) of different colors and lengths, each positioned horizontally or vertically on the grid. Cars can only slide along their row or column, such that vertically-oriented cars can slide up and down, and horizontal cars can slide left and right. However, cars cannot move through other cars. The goal of the game is to move the red car to the exit on the right side of the board (indicated by the red triangle). The solution to the puzzle is a sequence of moves that clears the red car's path to the exit. For instance, the specific puzzle shown in Fig. 1a can be solved in five moves (Fig. 1b-f): green up two spaces, purple left two, light blue down three, pink up one, and red right three.

Model overview Although the computational model used in this work is the same as in Berke et al., 2023, we present a brief, high-level overview here for completeness.

The model formalizes inferences about mental processes as Bayesian inference over a generative model of mental processing. The generative model is a probabilistic solver used to estimate the amount of computation needed to solve a puzzle. This solver is built to make efficient use of mental

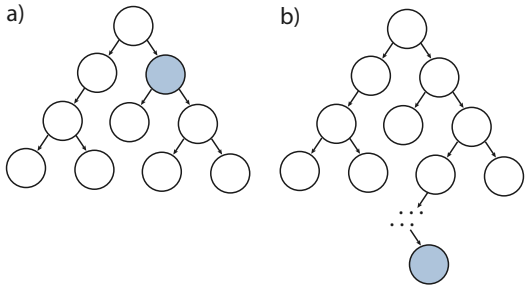

Figure 2: Conceptual illustration of the logic behind how computation is represented in our approach. Each tree represents the space of an agent's possible thinking trajectories through the puzzle. a) If the solution (shown in light blue) to a thinking problem is one layer deep, then we expect agents to answer quickly; if they don't, we get the intuition that they got distracted. b) If, on the other hand, the solution is several layers deep and the agent responds slowly, we may instead get the intuition that they were thinking deeply.

resources—i.e., minimize the total amount of computation needed to solve the task, and the memory load of plans—and has three design features supporting that end. First, to break the problem down into sub-goals, the solver uses means-ends-analysis (Newell & Simon, 1961; Newell, Simon, et al., 1972). Second, the solver implements an expectation that thinking is costly and people allocate their thinking rationally to reduce the amount of expected computation. This is implemented using a heuristic-guided search that estimates the amount of future computation and selects sub-goals and moves to reduce it. Finally, the solver interleaves planning and mental execution of moves to reduce memory load (similar to the approach in Korf, 1990). In the context of the high-level schematic in Fig. 2, this means that the solver assumes that agents are simultaneously constructing the thinking tree as they reason through the problem, deciding which branches to explore and expand by estimating which branches might require the least computation (i.e., the fewest number of additional moves) to reach a solution. For further details on the solver, please refer to Berke et al., 2023.

The solver, by virtue of being probabilistic, produces a distribution over how much computation it takes to solve a puzzle. This reflects the intrinsic uncertainty about how much another agent might have to think to find the solution, which depends on which sub-goals they prioritize thinking about (such that, on one extreme, they might be lucky and hit upon the solution immediately and, on the other extreme, they might first explore initially promising but ultimately fruitless avenues). By assuming that computation takes time and a prior over speed, this distribution over computation is transformed into a distribution over the time it should take to solve a puzzle. Combining this generative model of thinking in Rush Hour (i.e. the solver) with other cognitive processes (like daydreaming; see Berke et al., 2023) yields a full generative process of timing, which is inverted using Bayesian inference.

Behavioral Experiment

This experiment tests whether participants' thinking speed influences their judgments about the mental processes going on in someone else's mind. The first phase of the experiment tested participants' inferences about how long someone spent distracted as opposed to thinking about solving a puzzle, given an observed pause. Making this inference successfully requires an estimate of the difficulty of the puzzle and an expectation of how long it should take to solve. The second phase of the experiment tested participants' speed at solving Rush Hour puzzles. All data, stimuli, and materials are available: `https://osf.io/e3y9f/?view_only=a8f094e9a20a45c285046045b453ce78`

Participants

300 U.S. participants (Age: mean = 38 years, range = 19-77 years) were recruited on Prolific and randomly assigned to one of three conditions consisting of a subset of the trials (N = 100 participants per condition).

Stimuli

The stimuli in the phase of the experiment testing participants' ToM judgments (see Procedure) consisted of 18 short videos. Each video showed a static puzzle for some variable length of time (a pause), followed by the appearance of the words "Got it!" and an animation of the solution. Although the experiment referenced that a person was solving the puzzles (see *Procedure*), the videos only showed the puzzle and never showed any agents. The videos were generated by pairing a set of 6 puzzles (shown in the first row of Fig. 3) with three different pause times. The pause times were selected so as to test videos where people's judgment might show the largest variability,[1] as predicted by the model. The second row of Fig. 3 shows the standard deviation of the posterior distribution over the proportion of the pause spent daydreaming (the target of the model's inferences) as a function of the pause duration (in intervals of 2.5 seconds). For each puzzle, the pause duration that maximized predicted variation (the peak) was selected, as this trial should show the greatest effect. Then, intervals of pause times resulting in a predicted standard deviation of 0.2 or greater were constructed. From these intervals, a shorter and a longer pause time were randomly sampled, constrained (when possible) to be more than 2.5 seconds away from the already selected pause time (to ensure that the selected trials would be meaningfully different from each other). The vertical lines in the second row of Fig. 3 depict the selected pause times, with the pause time for the peak shown in orange, the shorter sampled pause in red, and the longer sampled pause in blue.

The stimuli in the speed-test phase of the experiment consisted of six other puzzles. The total set of twelve puzzles used in the two phases of experiment had been previously validated against the model (Berke et al., 2023).

[1] Short pause times might lead all participants to judge that the character did not daydream, while very long pause times might lead all participants to judge that the character daydreamed nearly the whole time, resulting in homogeneous judgments. This motivated using the model to identify the trials most likely to elicit heterogeneous judgments reflecting individual differences in thinking speed.

Procedure

Participants first completed a short tutorial on the puzzle game Rush Hour. They were then introduced to a character named Alex, who they would watch solving puzzles. Alex, they were told, would pause to solve the puzzle in their head, before saying "Got it!" and producing the solution. Sometimes, Alex would spend the whole pause thinking about how to solve the puzzle, but other times, Alex would daydream. Participants were instructed to try to tell what was going on in Alex's head during the pause before they said "Got it!"

In the phase testing judgments about Alex's mental processing, participants were assigned to one of three conditions, determining which set of 6 short videos they would view. The order of the videos was randomized. Because some of these videos might not give participants enough time to fully simulate solving the puzzle themselves, participants were shown a static preview of the puzzle for 10 seconds before they were allowed to proceed to the video. For each video, participants were asked to answer the question, "What was Alex doing?" by positioning a slider with endpoints "thinking for the whole pause" (coded as 0) to "daydreaming for the whole pause" (coded as 100) and the midpoint labeled "thinking for half, daydreaming for half." See Fig. 4. After completing all trials, participants answered the free-response question, "Did you try to solve the puzzles in your head?"

The last phase tested participants' puzzle-solving speed. Participants were instructed to solve a puzzle in their head, and then click the "Next" arrow as soon as they finished. Then, they were shown the puzzle again, and asked to click on the pieces whose movements were part of the solution that they found. Even if the solution that they found was wrong, they were instructed to click on the pieces that were part of it. After a practice trial, participants solved and reported their solutions for the six speed-test puzzles in randomized order.

Results

First-person reports of simulation Participants' responses to the question "Did you try to solve the puzzles in your head?" were coded based on whether the answer indicated that the participant tried to solve the puzzles themselves at least some of the time vs. never. All but 6/300 participants reported trying to solve the puzzles at least some of the time.

Effect of first-person thinking speed on ToM judgments Any effect of first-person thinking on third-person judgments may be small and difficult to detect. To tackle this challenge, we analyze the data in three different ways, trading off the sensitivity of the analysis for its assumptions.

The simplest approach to analyzing the data while making the fewest assumptions is to correlate participants' average time spent solving puzzles with their average judgment across trials. As it does not use the computational model, we call this analysis *Model-free*. However, as will be explained

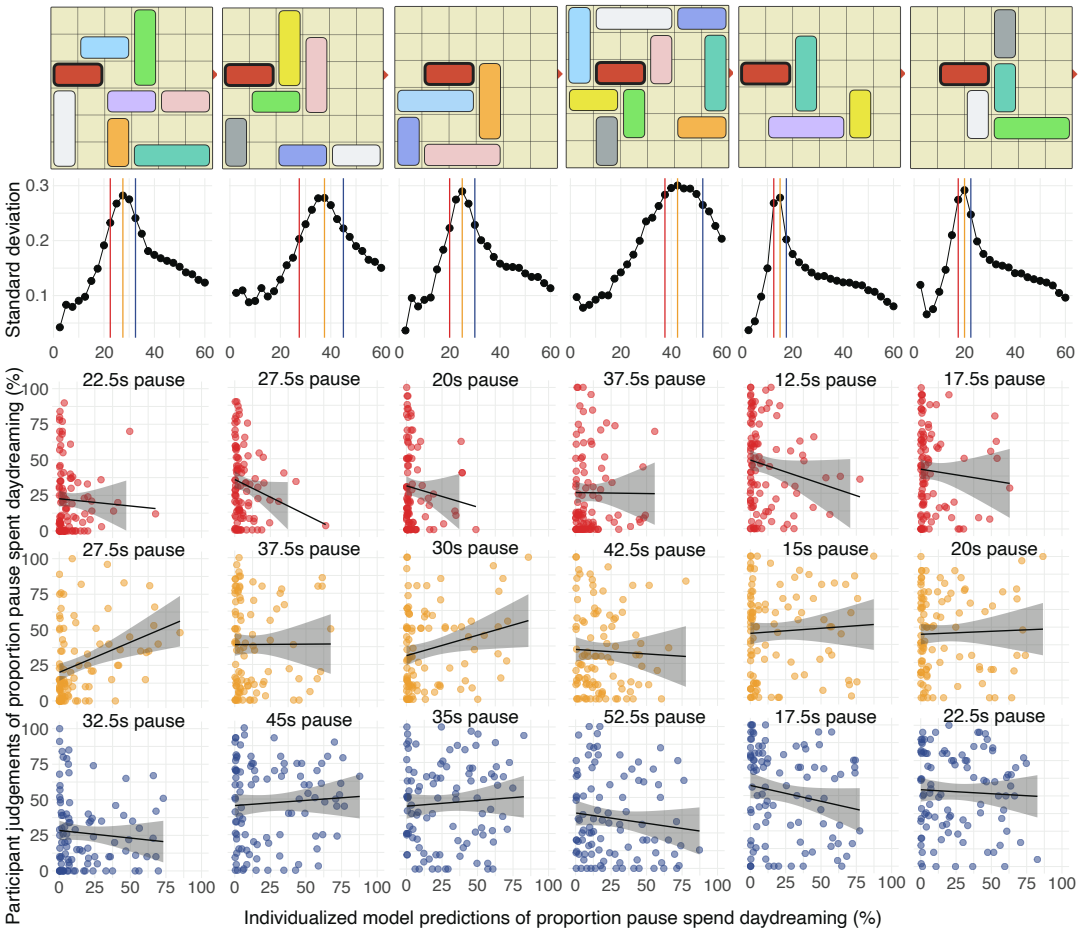

Figure 3: Trial-level stimuli design and results. Row 1 shows the six puzzles used in the thinking vs. daydreaming judgment phase. Row 2 shows, as a function of pause duration, the model's prediction about the variability of people's judgments for the puzzle in that column. Vertical lines indicate the selected pause durations (color-coded by their relative length: short:red; mid-length:orange; long:blue). Rows 3-5 show the results from the individualized analysis. Each plot displays the results for one trial, consisting of the puzzle in that column, plus a pause time. Row 3 gives the results for the shortest pause time, Row 4 for the mid-length pause time, and Row 5 for the longest pause time. These pause times were selected/sampled to elicit heterogeneous inferences (see *Stimuli*). Points show model predictions on the x-axis, and the participant judgments on the y-axis. If the model were able to predict individual differences in participant judgments at the trial level, we would expect each plot to show a positive correlation. The black line is the best-fit line, and the shaded region gives a 95% confidence interval.

shortly, this approach does not make efficient use of the data and results in a high exclusion rate.

Using the computational model to adjust for the varying difficulties of the speed-test puzzles enables us to calculate each participant's "time-per-compute"—a model-based metric for comparing across participants who solved different puzzles correctly during the speed tests. Participants' time-per-compute are then correlated with their average judgment across trials. This analysis includes more participants and is therefore more powerful, but it assumes that the model correctly estimated the amount of thinking each speed-test puzzle requires. We call this the *Model-based analysis*.

The most sensitive approach uses the model to infer the amount of thinking that each individual participant spent on each speed-test puzzle, so as to calculate a time-per-compute, which is used to produce individualized predictions of each person's judgment on each trial (*Individualized analysis*). This analysis assumes that the model not only correctly estimates the amount of thinking for each puzzle, but also captures how people infer daydreaming. In exchange, the fine-grained predictions lend greater power.

Model-free analysis If participants' first-person thinking systematically biases their third-person ToM judgments, then people who solve Rush Hour puzzles quickly might expect other people to also solve Rush Hour puzzles quickly. In the context of this experiment, this suggests that people who take less time to solve Rush Hour puzzles may judge Alex to have spent more of the pause daydreaming, and people who

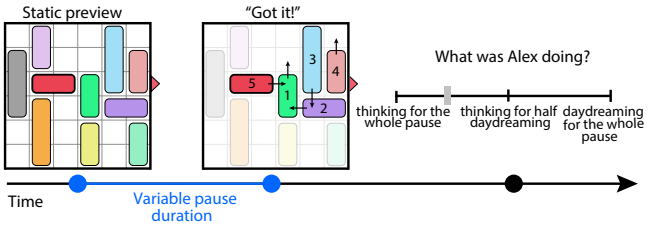

Figure 4: Schematic of a trial in the experiment phase testing judgments about thinking vs. daydreaming. Participants first see a ten-second preview of the puzzle. Next, they watch a video with a variable pause duration, followed by the message "Got it!" and a video of the solution. Participants the report the portion of the pause that the character spent daydreaming.

take longer to solve Rush Hour puzzles may judge Alex to have spent less of the pause daydreaming. Therefore, the first analysis looks for a negative correlation between each participant's average time spent on the speed tests and their average judgment of how much of the pause Alex spent daydreaming.

Unfortunately, not all participants were able to solve all six puzzles correctly in the speed testing phase. Since the six puzzles varied in difficulty and therefore how long they take to solve, simply averaging across participants' times on the puzzles that they solved correctly would result in averages that reflect which puzzles the participant solved. And averaging over the time spent on each puzzle, regardless of whether it was solved correctly, could introduce noise resulting from participants giving up or making mistakes. Thus, the cleanest way to extract average puzzle-solving times is to restrict the analysis to participants who all correctly solved a subset of the puzzles, and to only use the times for those puzzles.

Following this logic, we restricted this analysis to the 212 participants who were able to correctly solve the four easiest puzzles in speed-testing. We further excluded from this analysis participants whose mean solving time for these four puzzles was outside of 1.5 times the interquartile range (IQR).[2] This excluded 10 participants whose mean solving time was above 19 seconds for these easy puzzles.

For each participant, we now have an average time spent solving the puzzles. To compare participants' judgments who were assigned to different conditions and therefore viewed different trials, we first averaged each participant's judgments across the six trials that they viewed, and then z-scored each participant's average judgment when compared to other participants within their condition (thus accounting for any differences between conditions). No significant correlation could be found between each participant's average time spent on the speed tests and z-scored average judgment of how much of the pause Alex spent daydreaming ($r = 0.02$, $CI_{95\%}$: (-0.12, 0.16). See Fig. 5A. Furthermore, correlations between participants' average time spent solving the puzzles and average judgment within each condition similarly yielded no evidence: A: ($r = 0.20$, $CI_{95\%}$: (-0.04, 0.42); B: ($r = -0.23$, $CI_{95\%}$: (-0.45, 0.02); C: ($r = 0.09$, $CI_{95\%}$: (-0.14, 0.32).

[2]Because the distribution of solving times is skewed (rightward), IQR is a better measure of variability and better for defining outliers than standard deviation, which assumes a symmetric distribution.

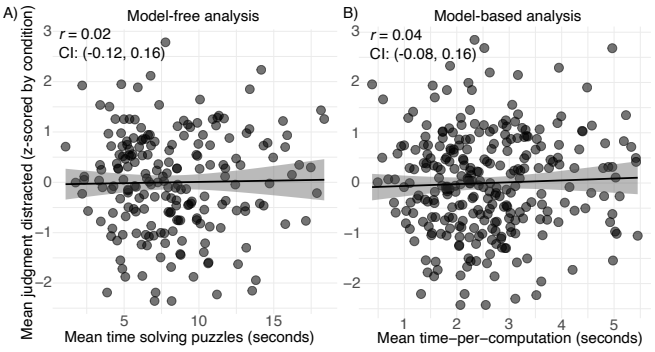

Figure 5: Results from A) the model-free analysis B) and model-based analysis. The black line is the best-fit line, and the shaded region gives a 95% confidence interval. The plot includes the Pearson correlation and its 95% confidence interval. In both plots, each point represents one participant, with their z-scored mean judgment of daydreaming on the y-axis. In A) the x-axis shows the mean time spent solving the puzzles during the speed testing phase, and in B) the x-axis shows the mean time-per-compute estimated using the model. If people base their judgments on their own experience of solving the puzzles themselves, then we would expect a negative correlation between how much time it takes them to solve puzzles and how much they judge Alex to daydream; faster people should expect Alex to solve the puzzles faster and therefore rely more on daydreaming to explain pauses.

While this analysis is direct and involves minimal processing, it is not the most powerful, for two reasons. First, it only included participants who solved at least the four easiest speed-test puzzles correctly (resulting in a 29% exclusion rate). To address this limitation, we conducted a *Model-based analysis* using the computational model to estimate the difficulty of each speed-test puzzle, allowing us to compare participants who successfully solved different puzzles during the speed tests. Second, it involved averaging judgments across trials, which may lose power. This limitation will be addressed later by the *Individualized analysis*.

Model-based analysis To measure the thinking speeds of participants who correctly solved different puzzles in the speed-test phase, we used the Rush Hour solver to estimate the difficulty of each puzzle (the amount of mental computation that each puzzle would take to solve). We estimated how much computation each puzzle required by using the MLE of the distribution over computation predicted by the model. This measure includes the expectation that agents may sometimes think about moves that look promising, but are not always correct. For each puzzle, each participant's solving time was divided by this measure to give an estimate of the participant's time-per-compute on that puzzle. Averaging across the

puzzles that they solved correctly gave an estimate of their overall time-per-compute. This approach yielded a time-per-compute for 296/300 participants (four participants failed to solve any puzzles correctly). Removing outliers (defined as having a time-per-compute outside of 1.5 times the IQR, i.e., taking more than 5.5 seconds per mental move) resulted in 282 participants (for a 6% exclusion rate).

As in the *Model-free* analysis, we averaged participants' judgments across trials. To compare participants across conditions, we z-scored their average judgments. As before, no evidence of a significant relation was found between each participant's average time-per-compute and average judgment of daydreaming ($r = 0.04$, $CI_{95\%}$: (-0.08, 0.16)). See Fig. 5B. Correlations within each condition similarly yielded no evidence: A: ($r = 0.10$, $CI_{95\%}$: (-0.10, 0.30); B: ($r = -0.10$, $CI_{95\%}$: (-0.29, 0.11); C: ($r = 0.14$, $CI_{95\%}$: (-0.07, 0.33).

This analysis leveraged our computational model to search for correlations between first-person thinking speed and third-person ToM judgments. However, this analysis still uses the judgment data inefficiently by averaging across trials to produce one overall mean judgment of distraction. If we could use first-person thinking speed not just to predict participants' overall but their trial-by-trial judgments, we may still be able to detect an effect. With this goal in mind, the *Individualized analysis* uses the model to actually infer each participant's time-per-compute and then predict each participant's judgments on each trial. This more fine-grained analysis relies heavily on the model and its assumptions, but it may have the power to detect effects that the coarser analyses missed.

Individualized analysis First, we used the model to infer each participant's time-per-compute. This involved conditioning on each participant's times for each speed-test puzzle, and taking the expected value of the resulting posterior. The advantage to this approach is that making use of the model's prior over thinking speed may help handle outliers without resorting to exclusion, but if the model's priors are misspecified, this could introduce further errors.

Time-per-compute was estimated for each of the 296 participants who solved at least one puzzle correctly. For each participant, this estimate was then used as the mean of the generative model's prior for mapping units of computation to time. Running the model forward with individualized priors over thinking speed generated individualized predictions of daydreaming judgments for each trial.

Fig. 3 shows scatterplots of each participant's judgment on each trial against the model's predictions. Combining data across trials, the correlation between model predictions and participant judgments was small but significant ($r = 0.06$, $CI_{95\%}$: (0.01, 0.11)). However, this correlation could reflect the model's ability to capture how average participant judgments vary from trial to trial, rather than how judgments vary from person to person. To test whether the individualized model is actually capturing the variation between participants, we performed a permutation test scrambling the pairings between each participant and their individualized model predictions, while retaining trial structure. We found that the observed correlation of $r = 0.06$ was expected under the null distribution ($p = 0.22$), indicating that the model's predictive power came solely from predicting trial-level differences rather than individual differences. No evidence of participants' thinking speed biasing their judgments was found.

Discussion

It's striking how invariant participants' judgments about another person's thinking were to the participants' own thinking speed. This is difficult to reconcile with the finding that 98% of participants reported trying to solve the puzzles themselves, and with the phenomenology of the task—it seems hard to look at a puzzle and *not* try to solve it!

In this experiment, there were two possible ways that first-person simulation could serve as input for inferences about another person's mental processing. First, people could have tried solving the puzzles themselves so as to judge their difficulty. But, since the solutions to the puzzles were shown, perhaps participants could rely on the solutions at the end of the video to extract the difficulty of the puzzle. Second, people could have used their own mapping from computation to time to anchor their expectations of how long the puzzle would take someone else to think about. If not from first-person, where did people get the mapping from complexity to timing? Perhaps we have such good expectations of how other people think that we can correct for differences between our own mind and others, at least in terms of speed.

While correcting for speed differences between our own mind and someone else's seems plausible and computationally simple (by scaling thinking times up or down), there are other sorts of differences that seem harder to correct. For example, if we make a silly mistake, we might not expect others to make the same mistake—but do we always know how much computation that mistake cost us, or what the counterfactual thinking path would be? Similarly, if our first attempt turned out to be right, we might not expect others to be quite so lucky—but would we know how much computation would have been used if we had taken a different thinking path? And if we are a beginner, we might have no idea of the strategies that experts use. For these kinds of differences, it is unclear whether simulation is a useful strategy. And if we do still simulate in these cases, what corrections do we apply?

This leads to an intriguing question: would someone whose mind is more "typical" or "average" possess a more accurate Theory of Mind than someone whose mind works very differently from those around them? The stereotype of the socially inept genius seems to support this idea—a very unusual mind might be a poor model of other minds, making simulation misleading for predicting or drawing inferences about others.

In future work, we hope to further explore cases where people may or may not use their own mind as a starting point for understanding other minds. Altogether, this work helps uncover (or rule out) possible inputs to human social cognition.

Acknowledgments

This work was supported by NSF award BCS-2045778. We thank our CogSci reviewers for their helpful comments.

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*(4), 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Berke, M., & Jara-Ettinger, J. (2021). Thinking about thinking through inverse reasoning. In *Proceedings of the annual meeting of the cognitive science society.*

Berke, M., Tenenbaum, A., Sterling, B., & Jara-Ettinger, J. (2023). Thinking about thinking as rational computation.

Bridgers, S., Jara-Ettinger, J., & Gweon, H. (2020). Young children consider the expected utility of others' learning to decide what to teach. *Nature human behaviour*, *4*(2), 144–152.

Cusumano-Towner, M. F., Saad, F. A., Lew, A. K., & Mansinghka, V. K. (2019). Gen: a general-purpose probabilistic programming system with programmable inference. In *Proceedings of the 40th acm sigplan conference on programming language design and implementation* (pp. 221–236).

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

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, *2*(12), 493–501.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910.

Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. *Advances in Neural Information Processing Systems*, *29*.

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with theory of mind. *Trends in Cognitive Sciences*, *26*(11), 959–971.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334.

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not so innocent: Toddlers' inferences about costs and culpability. *Psychological Science*, *26*(5), 633–640.

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64.

Korf, R. E. (1990). Real-time heuristic search. *Artificial Intelligence*, *42*(2-3), 189–211.

Martin, A., & Santos, L. R. (2016). What cognitive representations support primate theory of mind? *Trends in Cognitive Sciences*, *20*(5), 375–382.

Newell, A., & Simon, H. A. (1961). *GPS, a program that simulates human thought.* Rand Corporation Santa Monica, CA.

Newell, A., Simon, H. A., et al. (1972). *Human problem solving* (Vol. 104) (No. 9). Prentice-hall Englewood Cliffs, NJ.

Richardson, E., & Keil, F. C. (2022). Thinking takes time: Children use agents' response times to infer the source, quality, and complexity of their knowledge. *Cognition*, *224*, 105073.

Tomasello, M. (1992). The social bases of language acquisition. *Social Development*, *1*(1), 67–87.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. (2009). Help or hinder: Bayesian models of social goal inference. *Advances in Neural Information Processing Systems*, *22*.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235–8240.

Zhang, C., Kemp, C., & Lipovetzky, N. (2023, Jul.). Goal recognition with timing information. *Proceedings of the International Conference on Automated Planning and Scheduling*, *33*(1), 443-451.