Thinking about Thinking as Rational Computation

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

Abstract

Theory of Mind enables us to attribute mental states to others. But we not only make inferences about mental states (like what someone believes or wants), but about mental *processes* (like if someone is distracted or whether they remember something). Here, we present a computational formalization of these kinds of inferences. We propose that inferences about mental processes are structured around a *principle of rational mental effort*: the expectation that other people allocate mental resources rationally so as to minimize thinking costs incurred while pursuing their goals. We develop this theory into a computational model in the context of the *Rush Hour* puzzle game. In two behavioral experiments testing different inferences about mental processing, we find that our model predicts participant judgments. This work advances our understanding of the richness of the human mind's ability to think about other minds, and even about thinking itself.

Keywords:

Theory of Mind; Computational Modeling; Social Cognition; Rush Hour

Introduction

Imagine running into an old friend on the street and, when making small talk, you ask her what she's doing downtown. Intuitively, this is an easy question that she should be able to answer immediately. If instead, your friend pauses with no answer, you might start to wonder if she's hesitant to tell you, or if she's trying to figure out why you want to know. If your friend was rushing down the street when you spotted her, you might infer that she's having a hectic day and is collecting her thoughts. And if the pause gets too long, you might infer that she's lost in thought and might not have even heard your question.

This ability to reason about other people's minds, known as Theory of Mind (ToM; Gopnik et al. 1997; Wellman 2014), is a hallmark of human cognition, emerging early in infancy (Onishi & Baillargeon, 2005), and supporting a wide range of human activities, from social learning to moral reasoning (Gweon, 2021; Kiley Hamlin et al., 2013). Research over the past decade has found that inferences about other people's mental states like their beliefs and desires are structured around a *principle of rational action*—the expectation that agents act so as to maximize the rewards that they obtain while minimizing the costs that they incur (Baker et al., 2017; Gergely & Csibra, 2003; Jara-Ettinger et al., 2016; Jern et al., 2017; Lucas et al., 2014).

But as the opening example shows, people routinely go beyond inferring mental states (like what someone wants or knows) by also inferring other people's dynamic mental processes, like when we infer that someone is trying to recall something that happened, or when we determine that someone got lost in thought. Representations about mental processes are different from standard mental state attribution in two key ways. First, others' mental states are typically representations about the world (i.e., beliefs and desires represent what an agent thinks the world is currently like, and what they want it to be like, respectively). By contrast, mental processes do not represent world states; they reflect internal computations happening in other people's minds. Second, mental states are usually (but not always, e.g., Gates et al. 2021; Zhang et al. 2023) inferred based on the observable behavior that they produce (e.g., wanting a coffee is evidenced by an agent walking towards a coffee shop). By contrast, mental processes are not directly tied to observable behavior and often leave no traces beyond pauses in behavior.

In this paper, we present a computational model of inferences about mental processes, structured around a *principle of rational mental effort*—the idea that, analogous to the principle of rational action, people expect agents to allocate mental resources rationally, so as to minimize mental computational costs incurred while pursuing goals. In other words, we expect other people to flexibly decide what to think about based on what looks most promising for achieving their goals. This idea is inspired by research showing that, in first-person behavior, mental effort is costly and allocated rationally (Shenhav et al., 2017; Ongchoco et al., 2022), and that third-person observers are sensitive to the costs (Liu et al., 2019) and limitations (Alanqary et al., 2021) of mental effort, and to the amount of time thinking takes (Richardson & Keil, 2020, 2022).

Our model is centered around a posited capacity to estimate how much computation a rational agent would have to engage in to pursue their goals, and uses this representation to infer people's mental processes. We implement and test our model in the context of the *Rush Hour* puzzle game, and we test it in two behavioral experiments. In Experiment 1 we focus on inferences about when someone might be daydreaming, and in Experiment 2 we focus on inferences about whether someone is solving a problem from scratch or remembering the solution.

Computational Model

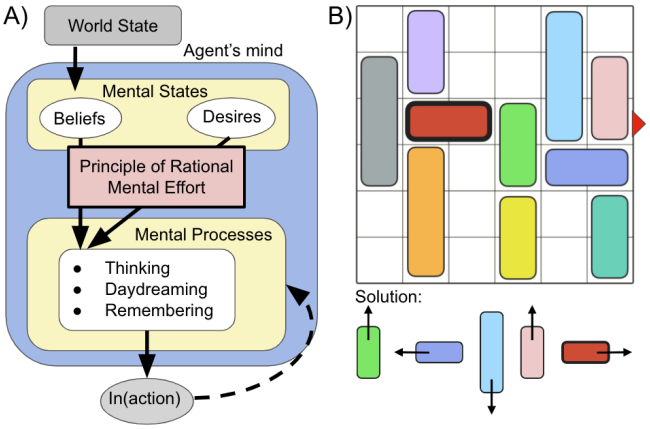

Figure 1: Conceptual Figure. A) Depicts a causal diagram, where solid arrows show causal relationships. The dotted arrow illustrates inference. The bubbles shaded in gray represent observable variables, while the white bubbles represent unobservable variables (i.e. mental states and mental processes). The principle of rational mental effort structures how an agent's beliefs (e.g., what the puzzle looks like) and desires (e.g., to solve the puzzle) relate to the mental processes the agent will engage in. Mental processes like thinking, daydreaming, and remembering influence observable actions or pauses in action. B) Depicts a Rush Hour puzzle and its solution move-by-move.

For simplicity, we present our model in the context of the Rush Hour puzzle game on which we test it. A Rush Hour puzzle (see Fig. 1B for an example) consists of a 6-by-6 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 horizontally-oriented 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 a red triangle). Then, the solution to the puzzle is a sequence of moves that clears out of the way the cars blocking the red car's path to the exit. For work on first-person thinking in Rush Hour, see Bennati et al. 2014; Bockholt & Zweig 2015; Bockholt et al. 2018; Bursley 2020; Jarušek 2013.

Our computational model aims to capture inferences that humans make simply by watching how long an agent takes to solve a puzzle, with a focus on whether the agent daydreamed (Experiment 1) or remembered the solution (Experiment 2). To achieve this, we estimate computation through a rational probabilistic solver that seeks to minimize computation. By combining this generative model of thinking in Rush Hour (i.e. the solver) with other cognitive processes (such as daydreaming or remembering), we further generate a probability distribution over expected pause lengths. This results in a full generative process of timing, which we invert using Bayesian inference. We implemented the model in Gen (Cusumano-Towner et al., 2019).

Estimating Computation

At the heart of our model is a solver that aims to find a sequence of moves that solves the puzzle. The planner begins with the target goal of moving the red car to the exit. If this move can be executed (i.e., there are no cars blocking the path), then the planner terminates successfully. If the move cannot be executed because other cars are blocking the move, the planner creates sub-goals to move each blocking car out of the way. These sub-goals, in turn, can then trigger further sub-goals. For example, in the board in Fig. 1B, the red car cannot be moved to the exit because the green, light blue, and pink cars are in the way. This leads the planner to create sub-goals to clear each of those three blockages. Clearing the light blue car's blockage in turn creates another sub-goal to move the dark blue car out of the way.

Note that it is not possible to solve sub-goals in parallel, because the moves needed to clear one blockage might affect the relevant cars for another blockage. To generate these sub-plans, the planner must therefore decide (1) which blockage to clear out first (when there are multiple blockages), and (2) where to move the blocking car (when there are multiple places where it could go). Our *principle of rational mental effort* is instatiated in these decisions, as follows. Our planner estimates how difficult it would be to clear each blockage, and probabilistically selects one via a softmax decision rule, such that $p(b) \propto \exp(C_b/\tau)$ where b is a blockage, C_b is the cost of clearing the blockage (see next section for details), and τ is the softmax temperature parameter. Similarly, when determining how to clear a selected blockage, the planner identifies all valid moves that would successfully move the car out of the way, and selects one based on the expected cost of the move, via $p(m) \propto \exp(C_m/\tau)$, where m is the target move, C_m is the expected cost of executing the move (see next section), and τ is the softmax temperature parameter.

This recursive process continues until the planner reaches a move that is immediately executable (i.e., not blocked by anything else)—which allows the planner to complete the move and return to the higher-level goal—or until it reaches a blockage that is physically impossible to unblock (e.g., a square that will always be blocked because its covered by a car so long that no matter where its moved, it will always cover the square), in which case the planner restarts.

The sequential nature of these plans can sometimes create infinite cycles, wherein unblocking one car creates a new blockage, and clearing the new blockage returns the board to its original state. Therefore, our solver also tracks state histories and restarts the planner when it identifies that it has been caught in a loop. Finally, longer plans impose higher working memory demands and increase the possibility a person might get lost and need to restart the plan. To account for this, we include a probablistic depth limit d which, when hit, restarts the planner (see Solver Parameters).

Estimating Costs The solver described so far requires estimating the approximate cost of clearing out a blockage (C_b), and the approximate cost of executing a move (C_m). These two costs are inter-connected: the cost of clearing a blockage depends on the cost of moving the blocking car out of the way (i.e. clearing a blockage becomes more costly when the move to clear it is also blocked). Likewise, the cost of moving a car is related to the cost of clearing any blockages preventing that move (i.e. a move becomes more costly as a function of how many blockages are in the way, and how hard they are to clear). We formalize this relationship through a pair of equations, where the expected cost of clearing a blockage b is given by:

$$E[C(b)] = \min_{m \in M} E[C(m)] \tag{1}$$

where M is the set of possible moves that would clear blockage b. Thus, the expected cost of clearing a blockage is the expected cost of the easiest move that clears the blocking car out of the way. Conversely, the expected cost of executing a move is given by:

$$E[C(m)] = 1 + \Big(\sum_{b \in B} E[C(b)] \Big) + \mathbb{1}(blocks(m, plan)) \tag{2}$$

The constant 1 captures the cost of executing the move once its unblocked. The second term adds the expected cost of clearing all blockages $b \in B$ that must be cleared out for the move to be executable. Finally, the third term $\mathbb{1}(blocks(m, plan))$ is an indicator for whether the destination of the car in move m blocks a future move in the current plan, penalizing the creation of a new blockage (to account for the fact that the car will need to be moved at least once more to clear this new blockage).

Note that Eq. 1 and Eq. 2 together form a recursive relation, such that $E[C(m)]$ is expressed in terms of the expected cost of clearing blockages, and $E[C(b)]$ is expressed in terms of the expected cost of future moves. The depth of recursion that the solver uses in estimating costs is regulated via a probabilistic lookahead.

Solver Parameters Rather than fixing parameters to specific choices, we instead represent each parameter through probability distributions that express uncertainty over their values. The rationality parameter τ is modeled with an exponential distribution with mean of 0.02 ($\lambda = 50$), representing a strong prior that the solver makes rational choices. A plan's depth limit d is sampled from a Binomial distribution with $p = 0.7$ and $n = 10$. This prevents unreasonably long plans but, in practice, minimally impacts the solver since the puzzles do not usually require seven layers of sub-goals to make a move executable. Finally, the lookahead use to estimate costs is controlled by a random variable drawn from a Geometric distribution on $[1, Inf)$ with parameter $p_{geom} = 0.9$, placing most of the weight on short lookaheads (for an average lookahead of 1.11 moves).

Relating Computation to Timing

The solver builds a probability distribution over the amount of computation a mind would need to solve a given puzzle, estimated in terms of the number of moves executed. We then map these computations into time durations by assuming that thinking about each move takes an amount of time t sampled from a Normal distribution with mean $\mu = 1.8$ and standard deviation $\sigma = 0.65$, obtained through an informal experiment in which five people each solved four Rush Hour Puzzles in their head. Their pause times were mapped to our model's estimate of computation (via maximum likelihood estimation). Finally, to account for the fact that people's perception of time is noisy and follows Weber's law (Halberda & Odic, 2015; Halberda, 2016), we add perceptual noise to the observed pause, using the Weber fraction $w = \frac{1}{10}$ (based on Haigh et al., 2021).

Alternative Models

To better test our theory of rational mental effort, we compare our model to two alternative models. The first alternate, the *Uniform Computation Model*, captures the possibility that, while people represent computation, they do not assume rational mental effort in others. We implement this model by replacing our solver with a breadth-first-search (BFS) strategy, which sequentially expands its search to neighboring board states and, unlike our solver, does not select moves so as to minimize expected computation. We set the thinking time parameters to $\mu = 0.13$ and $\sigma = 0.10$ seconds using the same procedure as above.

Our second alternate model, the *Action-Based Model*, captures the idea that people's inferences do not depend on any representation of computation, but merely on observable actions (like typical models of Theory of Mind). This action-based model makes inferences solely based on the number of moves in the observed solution. We tested this theory using a simple linear regression with the number of moves as the predictor. This regression was fit separately for each experiment.

Behavioral Experiments

To test our model, we conducted two experiments where participants watched an agent pause to solve a Rush Hour puzzle in their mind, and had to infer the agent's underlying mental processes. In Experiment 1, we tested people's ability to infer how long the agent spent daydreaming as opposed to thinking about solving the puzzle. In Experiment 2, we tested people's ability to infer how much the agent recalled about the puzzle as opposed to solved from scratch. These two inferences are complementary to each other: daydreaming usually extends the pause compared to focused thinking, and remembering the solution shortens the pause compared to thinking it through from scratch. All model parameters, predictions, stimuli, and analyses were pre-registered. Repositories containing the pre-registrations and data and available here: Exp1 and Exp2.

Experiment 1: Thinking vs. Daydreaming

Participants 120 U.S. participants were recruited on Prolific and randomly assigned to a subset of the trials (n=40 participants per video).

Stimuli Stimuli consisted of 36 short videos. Each video showed a static puzzle for some length of time (a pause), followed by the appearance of the words "Got it!", and an animation of the shortest solution. Although the experiment referenced that an agent named Alex 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 12 puzzles (shown in Fig. 3) with three different pause times. All puzzles were paired with a 15 second pause, a 30 second pause, and a randomly sampled puzzle-specific pause. This enabled us to both test how each puzzle's computational demands affect inferences while controlling for time (via the matched pauses), and also explore the full range of inferences that can emerge from differences in time (via the randomly selected time pauses). Each random pause was sampled from the set of pauses from 2.5 to 30 seconds by 2.5 second intervals (excluding 15 and 30, as they were already used), and from 30 to 60 seconds by 5 second intervals. We varied the intervals so as to sample more pause times between 2.5 and 30 seconds, in which range the model's inferences change the most.

Procedure Participants completed a brief tutorial teaching them the rules of Rush Hour and explaining that a character, Alex, would be solving Rush Hour puzzles. Alex would pause to solve the puzzle in their mind and then say "Got it" and produce the solution. Participants were told that Alex sometimes thinks about the puzzle for the whole pause, but sometimes, Alex daydreams, and that their job as participants was to tell what Alex was doing during the pause.

To help align the participants' priors about thinking time to the model's, participants watched two warm-up videos of Alex pausing to solve a puzzle (using a different puzzle for each video) before saying "Got it!" and producing the solution. Each warm-up video used the mean pause time predicted by the model for that puzzle, and participants were told that these videos showed Alex's average puzzle-solving speed when Alex did not daydream. Thus, this helped align participants' priors about how quickly Alex thinks.

In the testing phase, each participant watched 12 videos: one from each puzzle paired with a pause time (balanced such that 40 participants viewed each video). The order of the videos was randomized, along with which of the three pause times they saw for each puzzle. Because some of these videos have short pause times, they might not give the participant enough time to evaluate the puzzle's complexity. For that reason, 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."

Model predictions Daydreaming was modeled as a Bernoulli random variable day with probability $p_{daydream} = 0.5$. The length of the daydream t_d was sampled from a uniform distribution over $[0, 60]$ seconds. The pause time was then represented as $t * n + t_d * day$, where the first term is the time spent solving the puzzle. Given the board state and conditioning on the noisy observed pause time (with observation noise added as described previously in the *Relating Computation to Timing* section), we used the generative model to infer how long Alex daydreamed.

Results Participant judgments were averaged for each video (for a total of 36 observations) and compared directly to our model predictions. Overall, participant judgments were quantitatively aligned with our computational model ($r = 0.92$, $CI_{95\%}$: (0.86-0.96); see Fig. 2A). The middle row in Fig. 3 shows trial-by-trial results which reveal how both participants and our model showed fine-grained sensitivity to both the puzzle's complexity and to the pause time.

Interestingly, the Uniform Computation model also produced a high quantitative fit to participant judgments ($r = 0.84$, $CI_{95\%}$: (0.70-0.91)) but, critically, this fit was reliably lower than that of our main model ($\delta = 0.09$, $CI_{95\%}$: (0.02, 0.18)). That both the main and Uniform Computation models fit participant judgments well lends support to the hypothesis that people's inferences rely on an estimate of how much computation each puzzle requires (as both models estimate computation), but that the main model significantly outperformed the Uniform Computation model indicates that people's estimates of computation reflect an expectation that agents rationally decide what to think about.

Finally, the Action-Based model failed to predict participant judgments as accurately ($r = 0.67$, $CI_{95\%}$: (0.43, 0.82)), and performed reliably worse than our main model ($\delta = 0.26$, $CI_{95\%}$: (0.11, 0.45)), despite the fact that this model was directly fit to participant judgments. This suggests that people's judgments cannot be explained simply based on the number of moves in a puzzle's solution, and instead rely on some measure of how much thinking was required.

Experiment 2: Thinking vs. Remembering

Participants 120 U.S. participants were recruited on Prolific and randomly assigned to a subset of the trials (n=40 participants per video).

Stimuli Stimuli consisted of 36 short videos, like those in Exp. 1. The same 12 puzzles were again paired with three different pause times. Two of the pause times were fixed at 1 and 3 seconds. A third pause time for each puzzle was randomly sampled from 5 to 10 seconds, by 1 second intervals.

Procedure The procedure was almost identical to the procedure in Exp. 1, except that participants were told that Alex had solved all of these puzzles before. Sometimes, Alex re-

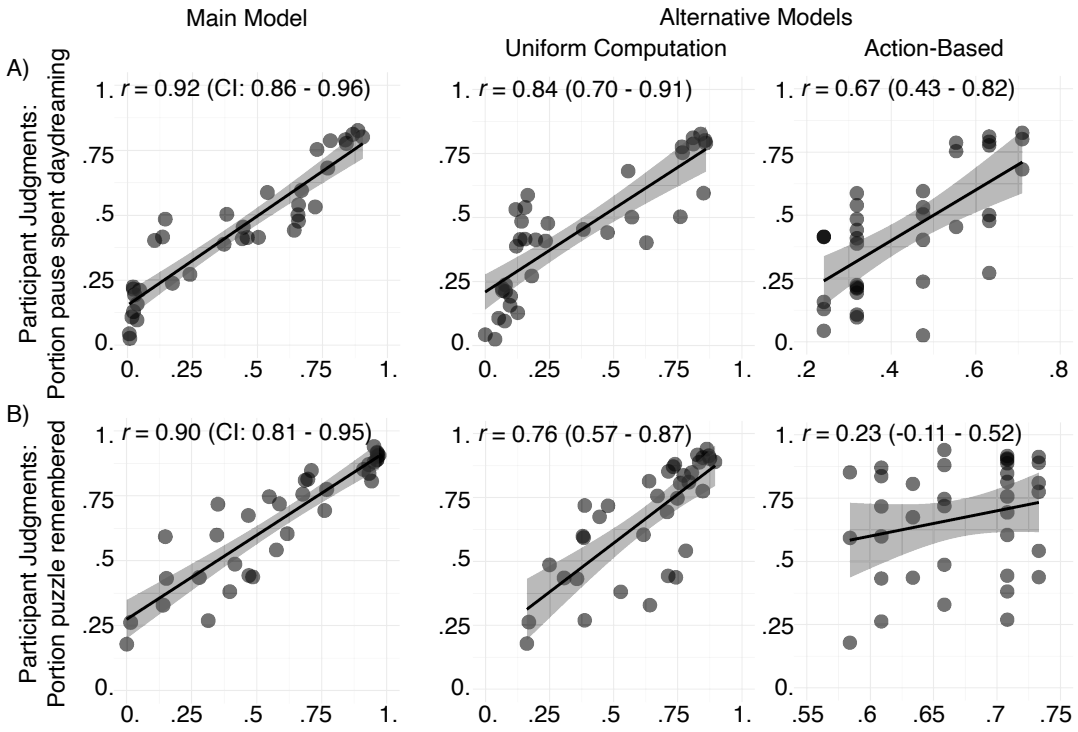

Figure 2: Results from both experiments. Each point represents the result on a video, with mean participant judgements on the y-axis and model predictions on the x-axis. The black line is the best-fit line, and shaded region gives a 95% confidence interval. Each plot includes the Pearson correlation followed by 95% confidence intervals in parenthesis. A) shows the results for Experiment 1: Thinking vs. Daydreaming, and B) shows the results for Experiment 2: Thinking vs. Remembering.

membered the entire puzzle perfectly, and other times, Alex did not remember the puzzle at all, and had to figure it out again from scratch. Still other times, Alex remembered part of the solution, but had to solve the rest of the puzzle. To align priors with the model, participants viewed the same two warm-up videos as in Exp. 1, and were told that in these videos, Alex did not remember the puzzle at all and was figuring it out from scratch. For each video, participants had to indicate how much of the puzzle Alex remembered, on a scale from "remembered 0% of the puzzle" (coded as 0) to "remembered 100% of the puzzle" (coded as 100), with the midpoint labelled "remembered about 50% of the puzzle."

Model predictions Memory was modeled as a discrete event which could happen at any of the n steps during puzzle solving, and which suddenly generated the rest of the solution, cutting the thinking short. The pause time was then $t * i$, where t is the amount of time it takes to think about each move (see *Relating Computation to Timing*) and i was sampled from a discrete uniform distribution on $[0, n]$. Given the board state and conditioning on the noisy observed pause time (see *Relating Computation to Timing* for details), we used the generative model to infer how many of the n puzzle-solving steps Alex remembered.

Results As in Exp. 1, participant judgments were averaged for each video (for a total of 36 observations). Overall, our main computational model quantitatively captured average participant judgements ($r = 0.90$, $CI_{95\%}$: (0.81-0.95); see Fig. 2B). The last row of Fig. 3 shows trial-by-trial results, revealing how both participant judgments and model predictions showed similar sensitivity to the puzzle's complexity relative to the duration of the observed pause. The fact that the same computational model was able to capture two different sets of intuitions (inferring daydreaming from time lags in Exp. 1 and inferring remembering from accelerated responses in Exp. 2) suggests that a representation of rational computation underlies people's ability to draw multiple kinds of inferences about other people's mental processes.

Like in Exp. 1, neither of the two alternative models were able to capture participant judgments as well as our main model. While the Uniform Computation model had a high quantitative fit to participant judgments ($r = 0.76$, $CI_{95\%}$: (0.57-0.87)), this fit was reliably lower than that of our main model ($\delta = 0.15$, $CI_{95\%}$: (0.03, 0.30)), confirming the importance of the expectation of rational thinking. The Action-Based model failed to reliably predict participant judgments ($r = 0.23$, $CI_{95\%}$: (-0.11-0.52)) and performed worse than our main model ($\delta = 0.67$, $CI_{95\%}$: (0.38, 1.00)), despite that it was directly fit to these judgments. This indicates that observed actions alone are not enough to explain participant judgments.

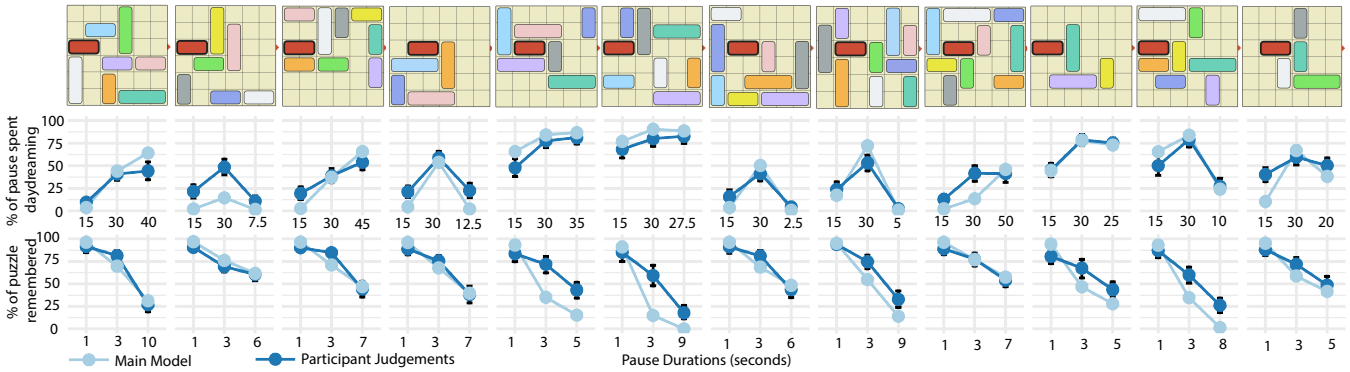

Figure 3: Trial-by-trial results. The first row shows the 12 puzzles, the middle row shows the results from Experiment 1, and the last row shows the results from Experiment 2. Main model predictions are in light blue, and mean participant judgements are in dark blue. Error bars depict bootstrapped 95% confidence intervals on mean participant judgements.

Discussion

Here, we proposed that the ability to reason about other people's mental processes is critical to human social intelligence and underlies a wide range of everyday intuitions that we use to navigate the social world (e.g., Is this person listening to me? Are they distracted or daydreaming? Is this person thinking about this for the first time? Or did they already know what I'm telling them?). We hypothesized that inferences about other people's mental processes are structured around a mental representation of computation, and furthermore, that estimation of this computation is based on a principle of rational mental effort — the expectation that other people rationally allocate their thinking resources so as to minimize the amount of thinking that they have to do to achieve their goals. Extending initial work (Berke & Jara-Ettinger, 2021) into a more general framework, we presented a computational model that describes others' thinking using a generative model of rational computation, implemented in the context of Rush Hour puzzles. Our computational model captured fine-grained human intuitions on two different kinds of everyday inferences: 1) whether and how much someone is daydreaming rather than focused on thinking and 2) how much of a problem or solution someone remembers. In both settings, participants showed sensitivity to pause time and puzzle complexity in exactly the way our model predicted, and their responses could not be explained by a model that posited thinking but lacked the notion of rational mental effort, nor by a model that attempted to draw these inferences from observable action.

While this work focused on two particular intuitions about mental processes, the approach we proposed might be able to capture a much broader set of intuitions about other people's thinking. For example, our framework could also explain inferences about which goal someone is thinking about (when each goal requires a different amount of computation), or whether someone might have a false belief (e.g. if we know that the answer to a question is surprisingly tricky, and there's a seeming obvious but incorrect solution). Our computational framework might be particularly useful applied to inferences in language, building on previous empirical work (e.g., Arnold et al. 2007; Fox Tree 2002; Heller et al. 2015; Kidd et al. 2011; Loy et al. 2017; Orena & White 2015) by modeling instances where a speaker's pauses can reveal aspects of their mental life (such as when someone pauses to remember something, or to carefully rehearse their words before broaching a sensitive topic). This suggests an exciting space of social inferences about other people's thinking that has been previously neglected by classical ToM work focused on inferences about other people's beliefs and desires.

One question that our work raises is how people acquire a generative model that enables them to estimate computation and how computation relates to behavior. Intuitively, this model is incredibly complex and context-sensitive, as our intuitions about thinking vary based on the situation. One possibility is that this generative model depends on our own first-person experience. In our debriefing questionnaires, many participants reported simulating solving the puzzles themselves. It is possible that the best way to estimate the amount of computation required to solve a problem is by solving the problem oneself. Critically, however, while first-person thinking might help estimate computation, it would be unable to explain the full range of human inferences about thinking, as we are intuitively able to reason about people who can think faster or slower than us in different tasks (rather than represent everyone else's minds as replicas of our own). Likewise, we can recognize that other people might not make the same thinking errors that we do, or they might make different mistakes (rather than assume others perform the exact same computations as we did). Understanding the origins of these intuitions about other people's thinking is a direction we hope to explore in future work.

Altogether, this work advances our understanding of the computations that underlie the human mind's ability to make rich inferences about other minds, even inferring complex processing from the simplest of behaviors, like a pause in action.

Acknowledgments

We thank Mario Belledonne, Emory Richardson, Tan Zhi Yi, and members of the Computational Social Cognition Lab for helpful conversations. This work was supported by NSF award BCS-2045778.

References

Alanqary, A., Lin, G. Z., Le, J., Zhi-Xuan, T., Mansinghka, V. K., & Tenenbaum, J. B. (2021). Modeling the mistakes of boundedly rational agents within a bayesian theory of mind. In *Proceedings of the annual meeting of the cognitive science society.*

Arnold, J. E., Kam, C. L. H., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: the on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(5), 914.

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*.

Bennati, S., Brussow, S., Ragni, M., & Konieczny, L. (2014). Gestalt effects in planning: rush-hour as an example. In *Proceedings of the annual meeting of the cognitive science society.*

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

Bockholt, M., Peters, O., Narciss, S., & Zweig, K. A. (2018). Analysis of human problem solving drafts: a methodological approach on the example of rush hour. In *Proceedings of the annual meeting of the cognitive science society.*

Bockholt, M., & Zweig, K. A. (2015). Why is this so hard? insights from the state space of a simple board game. In *Joint international conference on serious games* (pp. 147–157).

Bursley, J. K. (2020). *Evidence for recursive operations in human cognition*. Unpublished doctoral dissertation.

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).

Fox Tree, J. E. (2002). Interpreting pauses and ums at turn exchanges. *Discourse processes*, *34*(1), 37–55.

Gates, V., Callaway, F., Ho, M. K., & Griffiths, T. L. (2021). A rational model of people's inferences about others' preferences based on response times. *Cognition*, *217*, 104885.

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

Gopnik, A., Meltzoff, A. N., & Bryant, P. (1997). *Words, thoughts, and theories.* Mit Press Cambridge, MA.

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

Haigh, A., Apthorp, D., & Bizo, L. A. (2021). The role of weber's law in human time perception. *Attention, Perception, & Psychophysics*, *83*(1), 435–447.

Halberda, J. (2016). Epistemic limitations and precise estimates in analog magnitude representation. *Core knowledge and conceptual change*, 171–190.

Halberda, J., & Odic, D. (2015). The precision and internal confidence of our approximate number thoughts. In *Mathematical cognition and learning* (Vol. 1, pp. 305–333). Elsevier.

Heller, D., Arnold, J. E., Klein, N., & Tanenhaus, M. K. (2015). Inferring difficulty: Flexibility in the real-time processing of disfluency. *Language and speech*, *58*(2), 190–203.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Jarušek, P. (2013). Modeling problem solving times in tutoring systems. *Masarykova univerzita, Fakulta informatiky, Disertacnı práce*.

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

Kidd, C., White, K. S., & Aslin, R. N. (2011). Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental science*, *14*(4), 925–934.

Kiley Hamlin, J., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental Science*, *16*(2), 209–226.

Liu, S., Cushman, F., Gershman, S., Kool, W., & Spelke, E. S. (2019). Hard choices: Children's understanding of the cost of action selection. In *Proceedings of the annual meeting of the cognitive science society* (pp. 671–6677).

Loy, J. E., Rohde, H., & Corley, M. (2017). Effects of disfluency in online interpretation of deception. *Cognitive Science*, *41*, 1434–1456.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., ... Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PloS one*, *9*(3), e92160.

Ongchoco, J. D. K., Knobe, J., & Jara-Ettinger, J. (2022). Decision-making times reveal that people's thinking plans adapt to the problem they are trying to solve. https://doi.org/10.31234/osf.io/zqc9t.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258.

Orena, A. J., & White, K. S. (2015). I forget what that's called! children's online processing of disfluencies depends on speaker knowledge. *Child development*, *86*(6), 1701–1709.

Richardson, E., & Keil, F. (2020). Children use agents' response time to distinguish between memory and novel inference. In *Proceedings of the annual meeting of the cognitive science society.*

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.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual review of neuroscience*, *40*, 99–124.

Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University Press.

Zhang, C., Kemp, C., & Lipovetzky, N. (2023). Goal recognition with timing information.