UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
People use mixed strategies to make efficient but structured inferences about agents in roles

Permalink
https://escholarship.org/uc/item/44r9b95k

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors
Baker, Aaron
Sharma, Khushi
Dunham, Yarrow
et al.

Publication Date
2025

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

People use mixed strategies to make efficient but structured inferences about agents in roles

Aaron Baker[1], Khushi Sharma[2], Yarrow Dunham[1], Julian Jara-Ettinger[1]
[1]Department of Psychology, Yale University
[2]Department of Computer Science, Georgia Institute of Technology

Abstract

Roles are a pervasive part of our social landscape, but little is known about the mental models people use to reason about agents who occupy roles. In this paper, we test three computational models for role-based reasoning against participant performance in a social inference task. We find evidence that people exhibit mixed approaches which broadly track the computational efficiency of simpler models, but still retaining the structure of Bayesian inference models. These findings shed light on the mechanics of this important social cognitive system and pave the way for future work in this area.

Keywords: roles, Bayesian models, social inference

Introduction

Imagine it is your first day volunteering with an organization that builds houses and you're looking for a mentor. You want to do a good job, but there are dozens of other volunteers to choose from. You notice one volunteer hammers nails some of the time and scrolls on their phone otherwise. Another volunteer paints walls sometimes and hammers nails otherwise. How do you decide who is the more dedicated volunteer to be your mentor? This example taps into the intuitive dynamics at play when reasoning about agents who occupy roles.

One of the most important skills in our social cognitive toolkit is figuring out what people like and how much they like it. Research has shown that both adults and children rationally infer people's preferences by inverting causal models of decision-making (C. L. Baker, Saxe, & Tenenbaum, 2009; Kushnir, Xu, & Wellman, 2010; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). This ability to infer the preferences of others underpins a wide variety of social interactions. It guides how and when we help others (Gweon & Schulz, 2019), how we predict their behavior (Ho, Saxe, & Cushman, 2022), and how influence them (Wu, Schulz, & Saxe, 2024). While these inferential processes are robust, they can be computationally expensive (Jara-Ettinger, 2019; Ho et al., 2022). However, all of this research has been done in contexts where personal desire is isolated as the only motivation behind people's actions. But very often people do things not because they have a direct desire to do so, but because it's their job.

Roles are a pervasive part of our social lives. Many of the people we interact with every day occupy roles, such as bus drivers, doctors, and security guards. Recent work has argued that roles constitute an critical yet relatively understudied part of our social cognition (Jara-Ettinger & Dunham, 2024; Tomasello, 2020), and that understanding human sociality requires a robust understanding of how roles operate in cognition. Some empirical work has shown that people can readily infer the roles people occupy within an institution (Davis, Dunham, & Jara-Ettinger, 2022) and use roles to make quick and generalizable inferences about the people who occupy them (A. Baker, Dunham, & Jara-Ettinger, 2024; Noyes, Dunham, Keil, & Ritchie, 2021). However, little is known about how these representations precisely interact with mental state representations. This gap is a crucial piece of the puzzle: How do we negotiate whether an agent's behavior should attributed to their mental states (i.e. desires) or the role they occupy? This is a challenging problem because it is under-determined. Any behavior from someone occupying a role can be explained by fully appealing to personal preference, fully appealing to institutional roles, or a combination of both.

In the volunteer example, we intuitively consider how a volunteer's obligations interact with their personal desires. For example, if a volunteer spends as much time scrolling on their phone as painting walls, we can assume that they like scrolling on their phone more. In this way, role-based reasoning requires people to perform joint inferences over an agent's relationship to the role and their personal desires. But this can be achieved in many ways. Following from Bayesian models of Theory of Mind, one possibility is that people implement a process of inverse decision-making to infer these joint motivations. But these algorithms are costly, and adding another causal factor (the influence of the role on decisions) can make them costlier. Therefore, it is important to also consider ways people might reason in these contexts in more efficient ways.

In this paper, we test possible algorithms people use to make these joint role-desire inferences. We use a simple paradigm where participants observe an agent in a role choosing between tasks and make judgments about their personal preferences and their affinity for the role. We evaluate their performance by comparing it to a normative model that explains judgments as a process of inverse decision-making. This model is compared with two alternative models which explain participant judgments as quick but coarse estimations. If people are making structured inferences about the agent's desires and attitude about the role, our main model should match participant responses more closely than our alternative models across a wide range of judgments.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

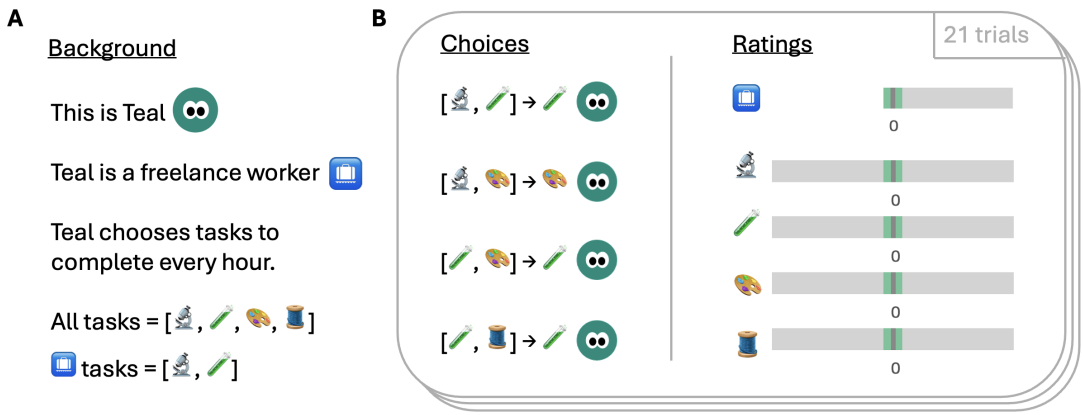

Figure 1: **A)** Task Backstory: Participants are introduced to Teal, who chooses tasks to complete based on how much they like a task and how much they care about doing tasks that pay as a freelance worker. **B)** Main Activity: Participants complete 21 trials (20 trials in Experiment 2) in random order. In each trial, participants see four choices Teal makes. They then rate how much Teal cares about the freelance worker job (from 0 to 10) and how much Teal likes or dislikes each task (from -10 to 10).

Paradigm

In our paradigm, participants are introduced to agents who choose one task to complete every hour (Figure1A). There are four tasks total in the paradigm, but every hour the agent is only offered two to choose from. Agents are described as having preferences for which tasks they like and dislike doing. Importantly, agents are also described as "freelance workers" who can complete two of the four tasks for pay. When they complete one of the two freelance worker tasks, they receive some fixed pay. When they complete one of the other two tasks they receive no pay, but they still enjoy any task according to their preferences regardless of pay. In each trial, participants and models observe four choices made by an agent and use slider scales to make judgments about how much the agent cares about doing the job and likes each individual task (a total of 5 judgments per trial).

To make this clearer, consider the stimuli in Figure 1B. The agent chooses the test tube task (a role task) every time it is offered, but also chooses the paints task (a non-role task) over the microscope task (the other role task). Intuitively, there are many ways to explain this pattern. For example, we could say that the agent likes the test tube task the most and the paints task the second most. Alternatively, we could say that the agent cares about doing the job, but really hates the microscope task. By having participants offer judgments over many trials, we can see which models best capture the strategies participants use with these role agents more generally.

Main Model

Our main model is based on Bayesian models of cognition[1], which have been shown to effectively capture human Theory of Mind (Jern, Lucas, & Kemp, 2017; C. L. Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017) and institutional reasoning (Davis et al., 2022). In these models, observers assume agents use a "forward model" of decision-making in which they choose actions that maximize some rewards. Then, by observing a set of choices, the observer uses Bayesian inference to estimate what an agent's set of rewards must be. Our model builds on this framework by assuming an agent who occupies a role gets additional reward for the actions that are required of the role.

Our model uses tasks $T = \{A, B, C, D\}$ where some tasks are explicitly associated with the role $T_{\text{role}} = \{A, B\}$. Each agent has a "personal" reward they get from each task t: $R_{\text{p,t}} \sim U(-1, 1)$ and a reward for any time they complete a role task: $R_{\text{role}} \sim U(0, 1)$. The overall reward an agent gets for choosing a task t is given by the following reward function:

$$R_{\text{t}} = R_{\text{p,t}} + \begin{cases} R_{\text{role}} & \text{t} \in T_{\text{role}} \\ 0 & \text{otherwise} \end{cases}$$

This forward model takes in four task pairs as input to generate four choices. To do this, the model samples values for $R_{\text{p,t}}$ for all $\text{t} \in T_{\text{role}}$ and R_{role}. Then, for each task pair, the model probabilistically chooses one of the two tasks in proportion to it's overall reward. These probabilities are given by a softmax function over the two reward values, which uses a rationality parameter $\tau = 0.4$. This value for τ was chosen based on a pilot experiment which ran a different set of trials and showed little qualitative difference in model fit using τ between 0.1 to 0.7, and so we selected the midpoint of 0.4.

This forward model is used with a series of observed choices (C) to jointly infer the most likely values for an agent's personal task rewards ($R_{\text{p,t}\in T}$) and role reward (R_{role}). The probability of the agent having a set of reward values given a set of observed choices is given by Bayes' theorem:

$$P(R_{\text{role}}, R_{\text{p,t}\in T} | C) \propto P(C | R_{\text{role}}, R_{\text{p,t}\in T}) P(R_{\text{role}}, R_{\text{p,t}\in T})$$

[1]Code for all models is available at https://osf.io/b7tjf/

Where $P(C|R_{\text{role}}, R_{\text{p,t}\in T})$ is the likelihood of making choices C given some reward values, and the prior probabilities of selecting a set of reward values $P(R_{\text{role}}, R_{\text{p,t}\in T})$ is constant because all rewards are sampled from uniform distributions. The final reward estimates are the expected values calculated using importance sampling over 10,000 samples.

Alternative Models

While Bayesian inference algorithms are robust tools for social reasoning, they can be computationally demanding. Because of this, we wanted to explore two additional questions in this paper. First, is it possible that people are using a simpler strategy to make these inferences which reduces the computational burden? Second, if so, is it possible that these are approximations of inverse decision-making? To answer these questions, we designed two alternative models that implement simpler strategies for estimating an agent's motivations in a role.

Alternative model 1: Ratio Tracking

In this model, each task estimate is calculated by summing the number of times that task was chosen divided by the number of times it was offered to the character (or 0 if it was never offered). Similarly, the role estimate is calculated by counting the number of times either role task was chosen divided by the number of times at least one role task was offered.

Alternative Model 2: Frequency Tracking

In this model, each task estimate is calculated by summing the number of times that task was chosen overall. Similarly, the role estimate is calculated by counting the number of times any role task was chosen overall.

Experiment 1

In this experiment, participants observed agents in a role making decisions about which tasks to complete. From these observations, participants made inferences about the agent's personal desires and their affinity for the role, which can be directly compared to model predictions. All aspects of the study were preregistered unless explicitly noted[2].

Participants

We recruited 50 adult participants from the US via Prolific to complete the experiment online (M_{age}=32.92, SD_{age}=12.50; 52% Female, 46% Male, 2% Non-binary). An additional 16 participants were excluded due to comprehension check failures. In a small deviation from our preregistration language, we resampled participants to compensate for exclusions until reaching our target sample of N=50.

Materials

The study was made up of 21 trials, where one trial consisted of seeing a new character choose one of two tasks over 4 hours (Figure 1B). Based on these four choices, participants were asked to rate 5 things about the character: How much they care about doing the freelance job (1 scale from 0 to 10) and how much they like doing each individual task (4 scales from -10 to 10). Participants had to click each of the sliders at least once before moving on to the next trial.

Trials were sampled from a compiled set of all trial permutations that can be made with the 4 tasks[3]. This process led to an overall set of 2,016 possible trials to sample from.

The sampling process used estimates from our main model run over all trials. The trials were grouped based on key estimates (judgments about role, judgments about role tasks, judgments about non-role tasks). One trial was sampled from each of these groups to ensure a substantial spread of trials along the ranges of these key model estimates. This process resulted in a final set of 21 trials used in the experiment.

Procedure

After providing consent, participants were introduced to our character Teal and the background story of the task (see Paradigm and Figure 1A). During the introduction, participants answered three comprehension checks about key details of the task (e.g. "Select all of the tasks a freelance worker can complete for pay"). Participants had to answer these comprehension checks correctly before moving forward. Participant who answered any one of the comprehension checks incorrectly more than twice were excluded. After the introduction, participants were familiarized with trial formatting and slider mechanics before beginning the main activity. In the main activity, participants completed 21 trials in random order (Figure 1B). Participants then answered optional demographic questions before finishing and collecting compensation.

Results

In Figure 2 we see participant responses and model estimates for individual trials. A positive value on the y-axis indicates a higher rating than average for that slider (Teal likes that task more than average or cares about the job more than average). A negative rating indicates a lower ratings than average. Before being compared to model predictions, participant responses were z-scored at the subject level and slider level (role vs task) then averaged. This resulted in a total of 105 data points to be compared with model estimates. Each model's estimates were also grouped by slider type and z-scored before comparison.

We first tested how participant judgments correlated with our main model (Figure 3). The main model showed a high quantitative fit with participant responses ($r = 0.87, CI_{95\%}$: $(0.82, 0.91)$), suggesting that our main model explains participant responses well across a wide range of possible choice patterns. Next, we correlated our alternative models with participant responses to test whether they can also be explained by by computationally cheaper strategies. Both the Ratio

[2]Materials and preregistrations for both studies are available at https://osf.io/b7tjf/

[3]Trial generation and selection code is available at https://osf.io/b7tjf/

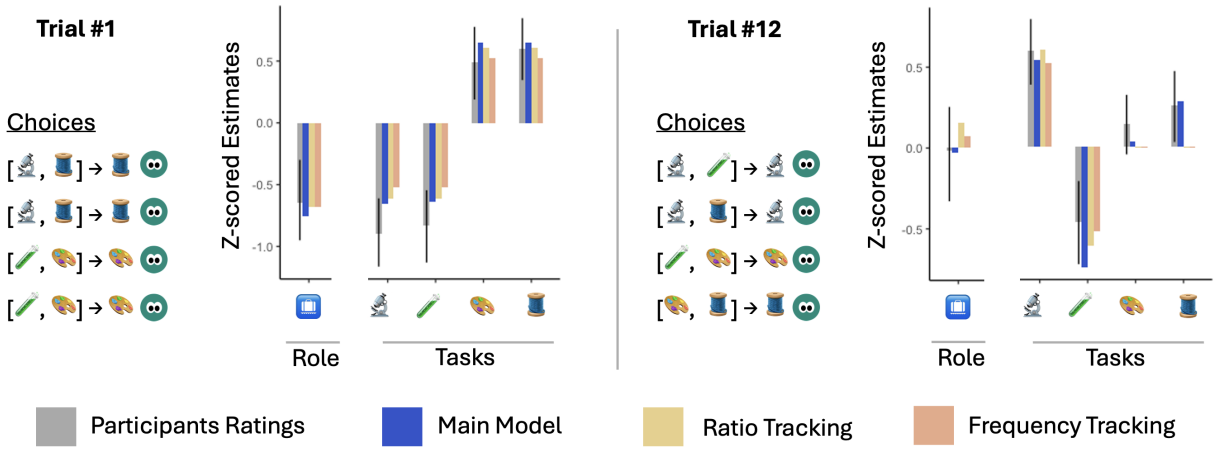

Figure 2: Example Trials. For each Trial, the character's choices can be seen on the left. On the right are average participant responses (with 95% confidence intervals) and model estimates for each measure. To better visualize the relationship between participants and the models, model estimates have been linearly transformed to fit the same scale as participant responses.

Tracking Model ($r = 0.89, CI_{95\%} : (0.84, 0.92)$) and the Frequency Tracking Model ($r = 0.92, CI_{95\%} : (0.89, 0.95)$) also showed high quantitative fits with participant responses (Figure 3). We used bootstrapped differences to test whether either of the alternative models show a significantly higher correlation than the main model. Neither the Ratio Tracking Model ($CI_{95\%} : (-0.05, 0.02)$) nor the Frequency Tracking Model ($CI_{95\%} : (-0.10, 0.01)$) showed a significantly higher correlation. These results indicate that all models explain participant judgments at a high level. Because the results from these models are so tightly matched, we investigated whether the alternative models could be approximating the main model by directly comparing the models to one another.

To test the relationships between our models, we ran exploratory analyses correlating the main model to the two alternative models. For the trials in Experiment 1, we found that the main model has a strong correlation with both the Ratio Tracking Model ($r = 0.92, CI_{95\%} : (0.89, 0.95)$) and the Frequency Tracking Model ($r = 0.81, CI_{95\%} : (0.74, 0.87)$). However, these trials are a subset of all the possible trials this paradigm allows. Therefore, we also tested how these models correlate across all 2,016 trials we generated in the trial selection process (see Materials). The correlation between the models remain relatively strong across all trials (Main Model and Ratio Tracking: $r = 0.87, CI_{95\%} : (0.87, 0.88)$); Main Model and Frequency Tracking: $r = 0.68, CI_{95\%} : (0.67, 0.67)$). These results support the notion that the alternative models effectively estimate the main model.

Although the alternative models correlate highly with participant responses, both suffer from coarse predictions. Each alternative model only give a handful of estimate values, so the data for each of them are clustered into groups (see Figure 3). These patterns imply that the alternative models treat the data points in these clusters as indistinguishable, and so participant responses for them should not meaningfully vary. The main model, on the other hand, produces a continuous range of estimates, which affords us an opportunity to test whether the main model captures something the alternative models miss. In a preregistered analysis, we partitioned our data into subsets based on these alternative model clusters. Then, for each subset, we correlated participant responses with the main model to see if it made predictions that the alternative model was missing. In our preregistration, we did not specify how many data points were sufficient for a cluster to be included in the analysis, so for this paper we excluded clusters with less than 10 data points. This yielded six total clusters, three from the Ratio Tracking Model and three from the Frequency Tracking Model. For five out of six of these clusters, the main model showed significant positive correlations with participant responses (Ratio Tracking clusters: $r = [0.57^{**}, -0.02, 0.71^{***}]$; Frequency Tracking clusters: $r = [0.63^{***}, 0.67^{***}, 0.53^{*}]$). These results suggest that, although all models follow the general trend of participant responses, the alternative models miss important nuance captured by the main model.

Discussion

In Experiment 1, we found that all three of our models broadly captured how participants reasoned about an agent occupying a role. However, the main model showed more nuance than the alternative models, fitting participant data in places where the alternative models make no distinctions.

We also found that these models correlate with one another, making it hard to distinguish between two general possibilities. The first is that participants are using the Bayesian model, and the two alternative models happen to approximate that pattern well. The second is that participants are using one of the more efficient alternative models, but they also take care in making small distinctions that the Bayesian model captures well and the alternative models miss.

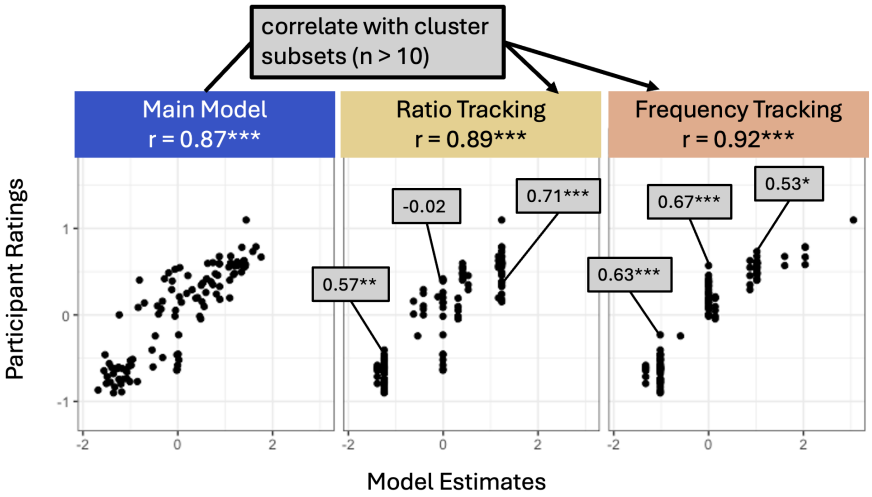

Figure 3: Experiment 1 Results. Each subplot contains 105 data points corresponding to one of the five measures (role and four tasks) for all 21 trials. All models show high overall fit with participants. However, both alternative models are grouped into vertical clusters which fail to capture meaningful meaningful variation in participant responses. For 5/6 of the larger clusters the main model positively correlates with participant responses, indicating a more nuanced fit with participant judgments.

In order to test these hypothesis further, we ran a second experiment, which is identical except that it deliberately uses trials that minimize the correlations between model predictions. By using trials where the models disagree, we can more clearly see which models align with participant intuitions.

Experiment 2

In Experiment 2, we tested a new set of participants in the same paradigm as Experiment 1, but with trials systematically selected to minimize similarities between model predictions. All aspects of the study were pre-registered unless explicitly noted[4].

Participants

We recruited 50 adult participants from the US via Prolific to complete the experiment online (M_{age}=37.44, SD_{age}=14.01; 52% Female, 46% Male, 2% Non-binary). An additional 4 participants were excluded due to comprehension check failures. We resampled participants to compensate for exclusions until reaching our target sample of N=50.

Materials

The study was made up of 20 trials and followed the structure and formatting as Experiment 1 (Figure 1B).

The key difference between Experiment 1 and Experiment 2 is how the trials were sampled. In Experiment 2, trials were selected to intentionally minimize the correlation between the main model and each of the alternative models[5].

Like Experiment 1, we first compiled a list of all permutations of 4 task pairs (the two tasks that are offered) and 4 task choices, leading to a total 2,016 trials. However, we identified that this list contained trials made up of the same pairs and choices but in different orders. Because our predictions do not consider the order of the pairs and choices, we removed all duplicates. This left us with a total of 1,365 trials to consider.

We then ran our main model and alternative models over these 1,365 trials. Using these model estimates, we selected trials in two steps: first we selected 10 trials that yielded the lowest overall correlation between the main model and the Ratio Tracking Model, then we removed these trials from the overall pool and selected 10 trials that yielded the lowest overall correlation between the main model and the Frequency Tracking Model. We combined these two groups of 10 trials, giving us a total of 20 trials for Experiment 2 that had correlations close to 0 for both model comparisons (see Results).

Procedure

Experiment 2 follows the exact same procedure as Experiment 1. Participants were introduced to the task (see Paradigm and Figure 1A), answered comprehension check questions, and were familiarized with trial mechanics. In the main activity, participants completed 20 trials in random order (Figure 1B). Participants then answered optional demographic questions before claiming their compensation.

Results

As in Experiment 1, participant responses were z-scored at the subject and slider (role vs tasks) level and averaged. This yielded one value for each measure across all 20 tasks, a total

[4]Materials and pre-registrations for both studies are available at https://osf.io/b7tjf/

[5]Trial generation and selection code is available at https://tinyurl.com/rolemodelsfiles

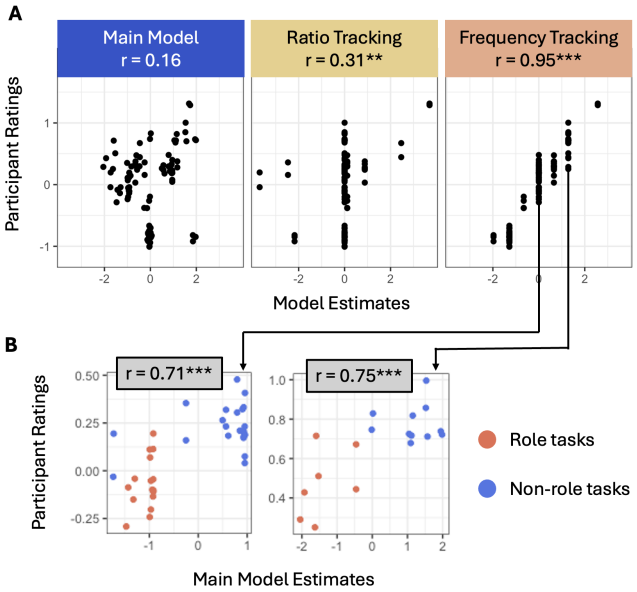

Figure 4: Experiment 2 results. **A)** Model correlations with participant responses, where Frequency tracking shows the highest overall model fit. **B)** Main model correlations with participant responses for clusters of data in the Frequency Tracking model. The main model demonstrates strong predictive power in these subsets, which appears driven by the main model's ability to infer that agents tend to like role tasks less than non-role tasks.

of 100 data points. Model estimates were also z-scored at the slider level before being compared to participant ratings.

Because trials were selected to intentionally minimize the correlation between the main model and alternative models, correlations between models for the trials used in this task were weak (Main Model and Ratio Tracking: $r = 0.12, CI_{95\%} = (-0.08, 0.31)$; Main Model and Frequency Tracking: $r = 0.00, CI_{95\%} = (-0.20, 0.19)$). However, it's worth noting that the models still correlate strongly with one another over all possible trials, even after removing order duplicates (see Materials; Main Model and Ratio Tracking: $r = 0.89, CI_{95\%} = (0.89, 0.90)$; Main Model and Frequency Tracking: $r = 0.74, CI_{95\%} = (0.73, 0.75)$).

For our primary analysis, we tested the overall correlations between participants and models (Figure 4A). The main model showed no significant overall correlation with participant responses ($r = 0.16, CI_{95\%} : (-0.03, 0.35)$), the Ratio Tracking model showed a weak positive correlation ($r = 0.31, CI_{95\%} : (0.12, 0.48)$), and the Frequency Tracking model showed a strong positive correlation ($r = 0.95, CI_{95\%} : (0.92, 0.96)$). Bootstrapped differences revealed that the Frequency Tracking Model showed a significantly higher correlation than the main model ($CI_{95\%} : (-0.97, -0.58)$) but the Ratio Tracking Model did not ($CI_{95\%} : (-0.41, 0.12)$). These results show that the Frequency Tracking model very closely captures participants overall pattern of responses.

To investigate these results deeper, we replicated the analysis from Experiment 1 in which we correlate the main model with participant responses for clusters of data in the Frequency Tracking model (Figure 4B). In this exploratory analysis, we found an interesting pattern: in each of large clusters, where the Frequency Tracking model predicts no variance in participant responses ($r = 0$), the main model positively correlates with with participant responses. Two of these correlations are small and not significant, but two of them are substantial ($r = 0.71, CI_{95\%} : (0.51, 0.84)$ and $r = 0.75, CI_{95\%} : (0.44, 0.90)$). Looking closer, we found that the data in these clusters organized neatly into role tasks and non-role tasks. This pattern shows that participants and the main model make inferences about role tasks and non-role tasks as categories, a distinction that the Frequency Tracking model completely misses.

General Discussion

Across two experiments, this paper investigates how people make intuitive judgments about agents occupying roles. We found that judgments are broadly consistent with the more efficient method of Frequency Tracking, and that Frequency Tracking serves as an effective approximation of the Bayesian model. However, participants also reliably made nuanced distinctions captured by the main model which Frequency Tracking alone cannot account for.

These results suggest that participants are using mixed strategies that are not fully accounted for in any one of the models tested. Here we entertain two possibilities for what this strategy could be.

One possibility is that people combine these two models in a way that plays to their strengths. Frequency Tracking allows them to arrive at a reasonable estimate in a more computationally efficient way. Then, Bayesian inference allows them to fine-tune these estimates and capture more nuance. This approach could resemble amortized inference (Gershman & Goodman, 2014), which has been proposed as a way that cognition can save computational resources in inference contexts such as these. Another possibility is that people are using a more sophisticated variant of frequency tracking which captures more nuance than the implementation used here and bypasses Bayesian inference entirely.

This project provides initial insights into the cognitive mechanisms of role-based reasoning. And yet, there is still much more to explore. First, we will test how people make predictions about what an agent will do when they exit a role. If people treat roles as causally distinct from personal desire, then predictions about what an agent will do next should depend on whether they are still in the role or not. Second, these experiments probe role motivation in a broad way, by asking how much the agent "cares" about the role. In reality, people can have many motivations for executing a role, such as money or status. Future work should disentangle these intuitions and shed light on specific motivations related to roles and their occupants.

Acknowledgments

This work was supported by NSF award BCS-2438827

Thank you to the members of the Computational Social Cognition Lab and the Social Cognitive Development Lab for their input throughout the project, as well as the family and friends for their role in supporting science (especially as gracious pilot participants).

References

Baker, A., Dunham, Y., & Jara-Ettinger, J. (2024). Roles guide rapid inferences about agent knowledge and behavior. *Proceedings of the 46th Annual Meeting of the Cognitive Science Society*. doi: https://doi.org/10.31234/osf.io/ad9u5$_v$1

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), 1–10. doi: 10.1038/s41562-017-0064

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

Davis, I., Dunham, Y., & Jara-Ettinger, J. (2022). Inferring the internal structure of social collectives. *Proceedings of the Annual Meeting of the Cognitive Science Society*. doi: 10.31234/osf.io/t5hpb

Gershman, S., & Goodman, N. (2014). Amortized inference in probabilistic reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Gweon, H., & Schulz, L. (2019). From Exploration to Instruction: Children Learn From Exploration and Tailor Their Demonstrations to Observers' Goals and Competence. *Child Development*, *90*(1), e148–e164. doi: 10.1111/cdev.13059

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with Theory of Mind. *Trends in Cognitive Sciences*, *26*(11), 959–971. doi: 10.1016/j.tics.2022.08.003

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110. doi: 10.1016/j.cobeha.2019.04.010

Jara-Ettinger, J., & Dunham, Y. (2024, April). The Institutional Stance.
doi: 10.31234/osf.io/pefsx

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. doi: 10.1016/j.tics.2016.05.011

Jern, A., Lucas, C., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*. doi: 10.1016/j.cognition.2017.06.017

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young Children Use Statistical Sampling to Infer the Preferences of Other People. *Psychological Science*, *21*(8), 1134–1140. doi: 10.1177/0956797610376652

Noyes, A., Dunham, Y., Keil, F. C., & Ritchie, K. (2021). Evidence for multiple sources of inductive potential: Occupations and their relations to social institutions. *Cognitive Psychology*, *130*, 101422. doi: 10.1016/j.cogpsych.2021.101422

Tomasello, M. (2020). The role of roles in uniquely human cognition and sociality. *Journal for the Theory of Social Behaviour*, *50*(1), 2–19. doi: 10.1111/jtsb.12223

Wu, S., Schulz, L., & Saxe, R. (2024). How to Change a Mind: Adults and Children Use the Causal Structure of Theory of Mind to Intervene on Others' Behaviors. *Proceedings of the Annual Meeting of the Cognitive Science Society*.