Detecting social biases using mental state inference

Mika Asaba (mika.asaba@yale.edu)[1], **Isaac Davis (isaac.davis@yale.edu)**[1],
Julia Leonard (julia.leonard@yale.edu), **Julian Jara-Ettinger (julian.jara-ettinger@yale.edu)**
Department of Psychology, Yale University, New Haven, CT 06511

Abstract

Social biases can negatively impact our sense of belonging, achievement, and social relationships. However, it is unclear what inferential processes underlie how people detect biases. We propose that people infer social biases by positing prior beliefs to account for potential gaps between what someone observed (e.g., seeing you succeed on a challenging task) and how they responded (e.g., recommending you try something easier). We present a computational model formalizing this process, and validate it with two experiments. We find a strong quantitative fit between model predictions and participant judgments across a range of inferences, namely which prior belief the coach held (i.e., which team the coach thought the player was on, or which bias the coach has). This work bridges computational methods with social psychological research on social biases, by showing how mental state inferences contribute to our ability to rapidly detect biases.

Keywords: Social Cognition, Theory of Mind, Computational Models; Social Biases

Introduction

Imagine trying out for a basketball team as a relatively short player. While the coach is watching, you proudly score four three-pointers in a row. However, after the try-outs, the coach sends you to the lowest-ranked group. Why did the coach assign you to that team? Given what the coach saw, they should think that you are quite good at shooting. You might suspect that the coach is biased against you because of your height—perhaps the coach has a strong belief that short people are generally bad at basketball and dismissed your shots as luck.

Situations like this are prevalent in every day life. A wealth of research in social psychology has found that people can easily detect social biases beyond blatant sexism and racism, such as when a seemingly innocuous comment or suggestion reveals that others are interpreting our behavior based on superficial aspects of our appearance (Spencer, Logel, & Davies, 2016; Sue et al., 2007; Dardenne, Dumont, & Bollier, 2007). It has also been well-documented that social biases negatively impact people's affective states (e.g., feeling hurt or offended, Sue, 2010), and a range of life outcomes, from academic achievement to social relationships to self-perceptions (often for the worst; Dovidio, Gaertner, Kawakami, & Hodson, 2002). However, we know relatively little about what inferential processes underlie our ability to detect others' biases in the first place.

Here we propose that these inferences stem from our ability to think about other people's minds—our *Theory of Mind* (Gopnik, Meltzoff, & Bryant, 1997; Wellman, 2014). That is, detecting social biases may reflect inferences about what mental states explain how other people treat us. For instance, in the above example, imagine that the coach watched you miss an easy layup and was distracted when you made your three-pointers. When the coach places you in the lowest-ranked group, you may not think that the coach holds a bias, because their action is justified given what they saw. This proposal is consistent with recent research showing that both children and adults can use other people's observations and behavior to infer their beliefs about subjective traits or qualities (e.g., what the coach thinks of my competence; Asaba & Gweon, 2022; Bass, Mahaffey, & Bonawitz, 2021; Kleiman-Weiner, Shaw, & Tenenbaum, 2017).

At the same time, explaining how we infer social biases through Theory of Mind poses a challenge. Past work on how people infer others' beliefs about them shows that these inferences are, broadly speaking, rational. That is, our inferences about what others think about us are structured around an assumption that people's social representations respond rationally to the evidence that our behavior provides. However, social biases, by definition, lack a rational justification. To accommodate this tension, we hypothesized that inferences about social biases might consist of inferring what prior beliefs an agent holds about us. Returning to the basketball example, we might expect the coach to rationally update their beliefs given our performance and think positively of our skill. If your performance was strong, but the coach treated you as a beginner, their recommendation can only be explained under an assumption that the coach had a sufficiently strong prior that you were a poor player, which could not be overridden by their observation.

Our proposal makes the commitment that inferences about others' biases reflect a form of mental state inference. However, it is possible that evaluations of others' biases do not rely on belief representations at all. Rather, people may rely on heuristics based on surface features of these scenarios. First, people may simply represent some behaviors as intrinsically biased, without considering whether their observations justify these behaviors (e.g., the coach placing an agent in the lowest-ranked group always suggests a negative bias). If this is the case, we would expect people to infer a bias based only

[1] These authors contributed equally to this work and are listed alphabetically.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

on the coach's actions (their feedback), and ignore what they saw. Second, people may mistakenly associate others' observations with their biases (e.g., the coach's observations of a player's poor performance means that the coach has a negative bias). If this is the case, we would expect people to infer a bias based solely on what the coach saw, before the coach provides any actual feedback.

Here we provide an initial test of our proposal and these alternative accounts. We present a computational model of social bias detection through prior belief inference, along with lesioned models that formalize the two simpler heuristics that might underlie bias detection. Intuitively, our model performs a joint inference over our own true competence, an observer's beliefs about our competence, and the priors that the observer must have to justify this belief. We evaluate our model in two behavioral experiments. Experiment 1 tests inferences about which prior someone holds (i.e., which basketball team a coach thinks a player is on). Experiment 2 explicitly tests inferences about biases (i.e., which bias a coach holds about a player). Data, analyses, and model code and predictions can be found here (https://tinyurl.com/krsut42u) and preregistration here (https://osf.io/cej83).

Computational Framework

For simplicity, we explain our computational model in the context of our experimental task, which consisted of two stages. In the first stage, an agent (the player) makes initial attempts at a skill-based task—attempting to throw a ball through one of three different basketball hoops—resulting in a set of outcomes $O_S = \{o_1, \ldots, o_n\}$, each specifying the basket that was attempted as well as the outcome ("hit" or "miss"). Critically, a second agent (the coach) observes some, but not all, of these outcomes $O_C \subset O_S$. After this first stage, the model updates both the player's belief and the coach's belief about the player's ability, and these beliefs might conflict with each other due to differences in the observed evidence (e.g., the player succeeded on the easy and hard shots, but the coach only saw the easy shots). This belief formation is affected by three considerations. First, we assume that agents can sometimes succeed or fail simply due to luck, but the agents cannot directly observe which shots were lucky or unlucky. Second, we assume that coaches rationally update their beliefs about the player based on the observed performance O_C (which we explain in detail below). Finally, we assume that the coach has some initial belief about the player's competence, which might include some biases (e.g., the coach might have been told, possibly incorrectly, that the player is on the advanced team, and therefore have higher prior expectations about the player's skill level), and that this initial belief influences the coach's updated belief.

In the second stage of our task, the coach suggests that the player make one final attempt on one of the baskets. The coach's recommendation is based on their belief about the player's skill, with the goal of getting the player to succeed at the hardest shot they are good at (e.g., recommending the easiest shot implies that coach doesn't think the player would succeed at either of the harder shots).

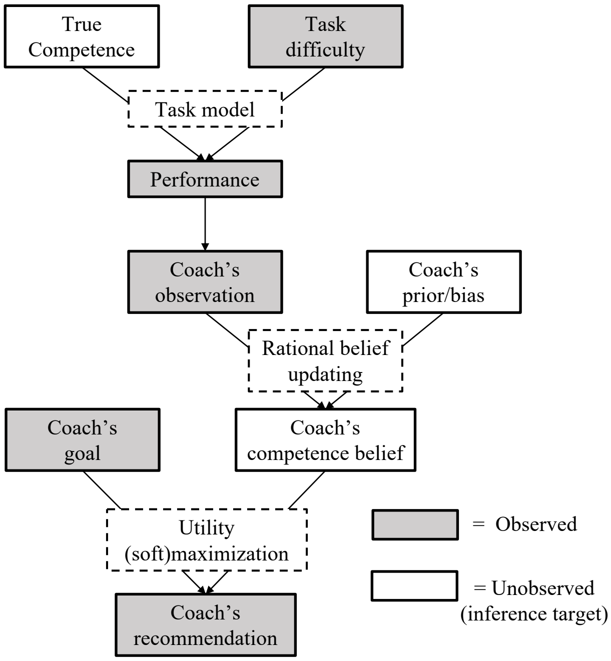

Figure 1: Schematic of our computational model. Dashed lines denote processes, and solid lines denote variable. Gray boxes indicate observable variables, and white boxes denote unobservable variables.

We modeled this full procedure as a causal mental model (Fig. 1) of the entire process, which we assume people can invert through a form of Bayesian inference. Previous work has leveraged similar assumptions to explain a range of inferences about mental states and behavior (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Jara-Ettinger, Schulz, & Tenenbaum, 2020; Lucas et al., 2014; Jern, Lucas, & Kemp, 2017).

Generative model

Figure 1 shows a schematic of our generative model, which consists of three components: a task model, a model of belief updating, and a model of how the coach chooses a recommendation.

Part 1: Task Model The task model assumes that the player's performance is determined by their true competence and the task's known difficulty. Formally, the player has some skill level $s \in (0,1)$, and each basket has some difficulty level $d \in (0,1)$. Given these parameters, the agent's probability of making a shot is given by $P(\text{hit}|s,d) = 1/(1+exp(-\beta * (s-d)))$, i.e.: a logistic function with growth rate β. We further assume that each shot has a small probability ε of automatically hitting or missing, independent of skill or difficulty.

Part 2: Belief Updates The task model produces a set of outcomes O_S, each specifying a basket and an outcome (hit or miss), of which the coach observes a subset O_C. Both agents then update their beliefs about the player's skill level via Bayesian inference:

$$P(s|O) \propto P(O|s)P(s) \tag{1}$$

where O is either O_S or O_C (depending on which agent is updating), $P(O|s)$ is the likelihood of observing the outcomes in O, given by the task model, and $P(s)$ is the agent's prior beliefs about the player's skill.

To constrain the space of possible prior beliefs, our task used a structure where players could be on one of three different teams—Beginner, Intermediate, or Advanced—each associated with a known competence distribution ($P_{Beg}(s)$, $P_{Int}(s)$, or $P_{Adv}(s)$). We further assume that the player does not know their own true skill level, but does know which team they are on, and the distribution of skill levels on that team. Thus, the player uses $P_{trueTeam}$ as the skill prior for updating, and the coach uses $P_{coachTeam}$ as the skill prior, where *coachTeam* is the team that the coach believes. This prior belief is how we represent the coach's potential bias (e.g., a prior belief that a player must be from the advanced team without evidence to support that belief).

Part 3: Coach's Recommendation The final part of the model generates a recommendation by combining the coach's beliefs about the player's competence with their goal—maximizing the expected number of points that the player will receive from the shot. The expected value of a shot is given by $P(\text{hit}|s, d_i) * V(i)$, where $V(i)$ is the value of succeeding at shot i. Because the coach does not know the player's exact skill level s, the expected value of recommendation i is therefore integrated over the coach's belief about skill:

$$EV(i) = \int_{s=0}^{s=1} [P(\text{hit}|s, d_i) * V(i)] * P(s|O_c) ds \tag{2}$$

where $P(s|O_c)$ is the coach's belief in each player's level of skill (Obtained from Eq.1). Given an expected value associated with each possible recommendation, the coach uses a softMax decision policy to make a recommendation.

Inference

Our experimental task asks participants to infer a) the player's belief about their own skill, b) the coach's belief about the player's skill, and c) the coach's prior bias about the player (before observing the player's performance). To generate predictions for these three variables, we inverted the generative model defined by equations 1 and 2 via Bayesian inference, conditioning on the values specified by the trial (i.e.: the player's performance, the coach's observation, and the coach's recommendation). This yielded, for each trial, three posterior distributions, one for each participant response variable. We then took the expected value of each posterior distribution to generate our model predictions.

Behavioral Experiments

Our behavioral experiments have two goals. First, to validate our computational framework and test if it can explain how people infer another agent's prior beliefs (Experiment 1; preregistered). Second, having validated that our model captures how people infer others' prior beliefs, we test whether this model explains how people react to evidence that an agent might have a prior belief reflecting a social bias (Experiment 2; not preregistered).

Experiment 1

Methods

Participants 150 adults (M_{Age}(SD) = 33.8(12.5), range: 18-71; 71 women, 68 men, 4 non-binary, 7 no response) were recruited from Prolific. An additional 8 subjects were recruited and excluded for failing one or more comprehension check questions (preregistered criteria).

Stimuli Experiment 1 used 30 different stimuli (see Fig. 2 for examples). Each stimulus depicted: a player who attempted four shots (and the player's performance on each shot), each of which could be Short, Medium, or Long shot; a coach who observed one of those shots (coach observation); and the coach's recommendation for which of the 3 shots the player should attempt next (coach recommendation). The stimuli space was constructed by crossing 5 patterns of player performance with 2 different coach observations and the 3 possible recommendations, to create 30 total trials. The 5 patterns of player performance varied which shots the player made versus missed, with the purpose of creating an even distribution of player competence from low to high. Across trials, the player's beliefs about their competence and the coach's beliefs about the player's competence ranged from low to high, and each trial had closely matched versions (i.e., that only differed in the coach's observation or recommendation). Each participant saw 10 trials, which always contained the 10 combinations of player performance and coach observation.

Procedure Participants first underwent a brief tutorial. They were introduced to three basketball teams, Beginner, Intermediate, and Advanced, and shown each team's average success rate for throwing a ball into each type of hoop (Short, Medium, and Long distance). Next, participants learned that they would watch a player practice in the presence of a coach (who would occasionally leave and therefore not watch the entire practice). Critically, participants were told that the coach had a pre-existing belief about the player's team, but this belief could be incorrect (as the player told the coach which team they are on, but the coach may have misheard). At the end of the practice, the coach recommends which shot the player should do to maximize the expected number of points the player receives (1 point for Short; 2 points for Medium; 3 points for Long). Participants were required to answer five comprehension check questions correctly, to ensure that they understood the story; if they answered incorrectly, they were prompted to try again. Finally, participants did a training with the pie chart scale used for the team inferences.

After the tutorial, participants underwent 10 trials (see Fig.

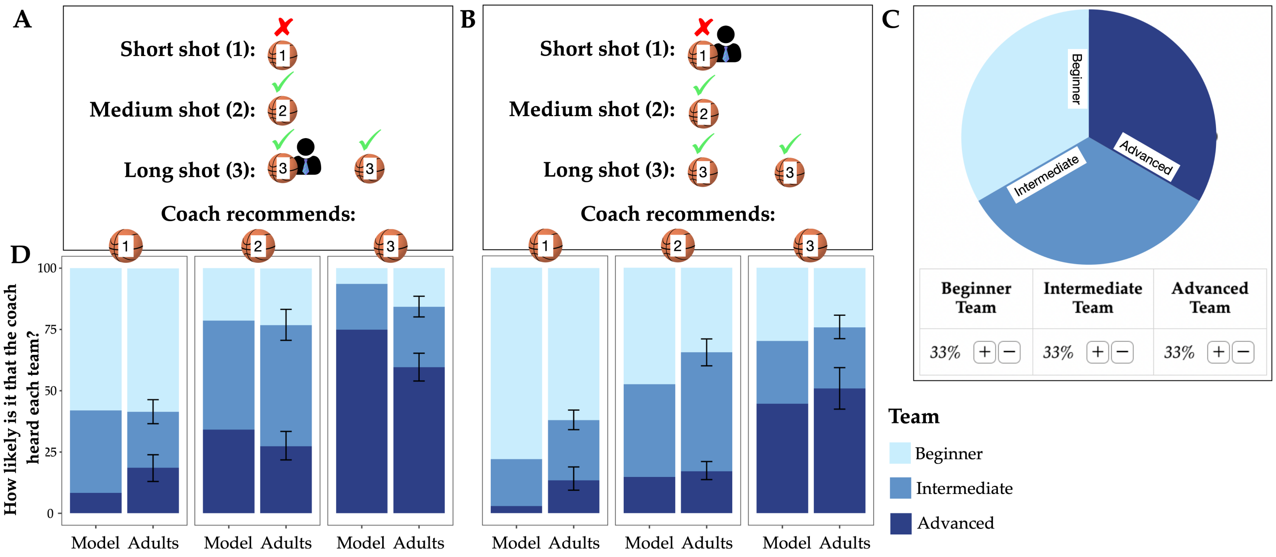

Figure 2: Example experiment stimuli (A-C) and results (D). (A) and (B) depict trials in which the player fails at the short shot, succeeds at the medium shot, and succeeds twice at the long shot. In (A), the coach only observes a success on the long shot. In (B), the coach only observes a failure on the short shot. After the player's performance and coach's observation, the coach recommended a shot for the player to do next: the short (1), medium (2), or long (3) shot. Finally, participants responded to a question about the coach's team prior: "Which team did the coach hear?" (C) shows the pie chart rating scale for the team inference; participants could drag the pie chart sections or click the '+' or '-' buttons. (D) shows Full model predictions and participant judgments (means and 95% bootstrapped CIs) for the team questions.

2A and B for two examples). In the first stage of each trial, participants saw the player's performance outcomes and which shot the coach observed. The player always attempted four shots and the coach always observed only one of the shots. The player's performance was depicted as a red "X" (miss) or a green check mark (hit). In the second stage, the coach recommended a basket for the player to do for points.

On each trial, participants responded to a check question about the coach's observation ("Which shot did the coach see?"), which they were forced to get correct. Then, they responded to three test questions. The first two test questions concerned the player's competence, the player's own beliefs about their competence ("What does Player [name] think of themselves?"; 100-pt sliding scale from "Extremely bad" to "Extremely good") and the coach's beliefs about the player's competence ("What does Coach [name] think of Player [name]?"; same scale). The third question concerned the coach's beliefs about which team the player is on ("How likely is it that the coach heard each team?"). Participants responded using a pie chart scale with three sections, one for each team. Participants could drag each section of the chart or click buttons to indicate their response (i.e., the probability that the coach thinks the player is on each team; see Figure 2C). Each participant did a set of 10 trials (randomized order).

Results

Each of the 30 trials produced 5 data points: 2 competence inferences (one for the player, one for the coach), and 3 team likelihood inferences (nonindependent; one per team). Our full model showed an overall strong quantitative fit with participant judgments (*r*=.90, 95% CI: [.86, .92]; preregistered), and model fit was similar for each inference type (competence: *r*=.95, 95% CI [.92, .97]; team: *r*=.87, 95% CI [.82, .92]; both preregistered).

Our Full model assumes that people consider *both* the coach's observations and their recommendation to infer the coach's prior belief (which team the coach heard). However, it is possible that participants simply relied on one of these dimensions (either the coach's observation or their recommendation). We tested these accounts with two models that lesioned off one variable: the Observation only (which considered the coach's observation and ignored their recommendation) and Recommendation only (which considered the recommendation and ignored the observation). The Observation only model had an overall correlation of *r*=.60 (95% CI: [.49,.69]), which was reliably lower than the Full model (δ = .30, 95% CI: [.20, .42]; preregistered).

The Recommendation only model had an overall correlation of *r*=.90 (95% CI: [.86,.92]), which was not different from the Full model (δ = .00, 95% CI: [-.03, .02]; preregistered). Despite the high numerical correlation, this model showed less sensitivity to human judgments. This is visualized in Fig. 3, where the vertical lines show collections of judgments where the model produced the same inference (hence not varying on the x axis) but participants had different intuitions (hence varying on the y axis). To test whether

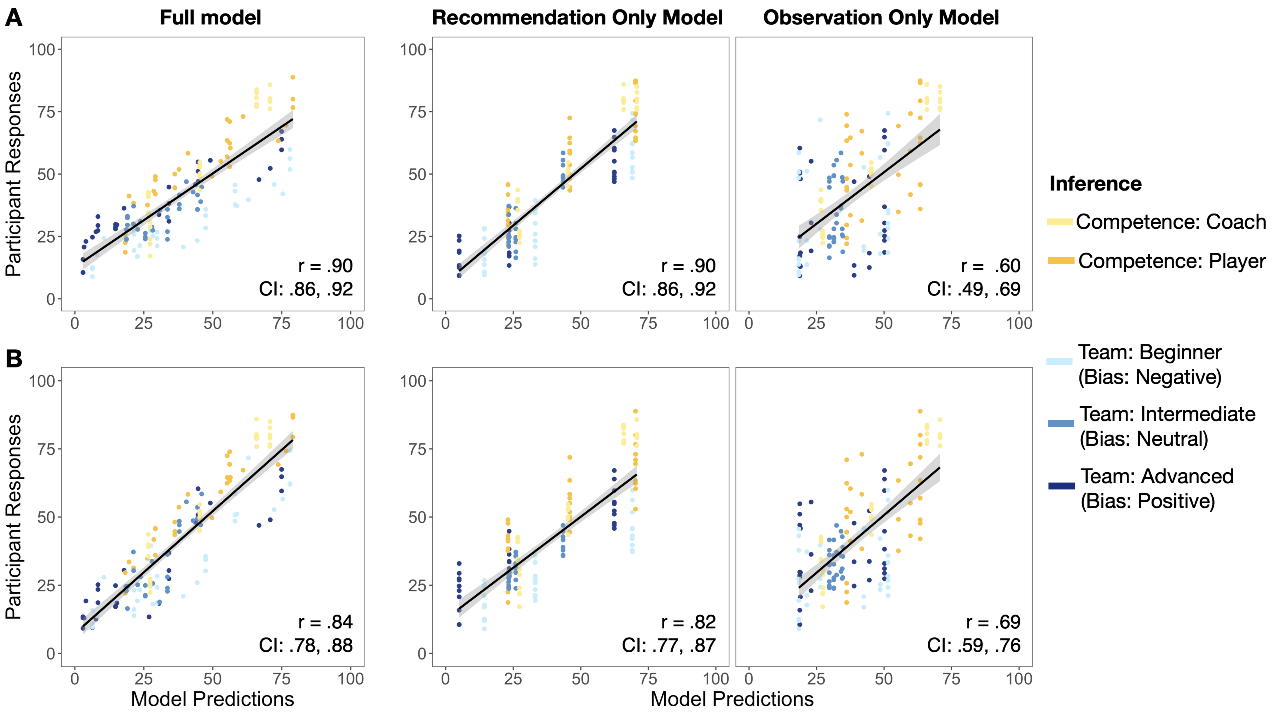

Figure 3: Experiment 1 (A) and Experiment 2 (B) model correlations with participant judgments: Full Model (left), Recommendation Only Model (middle), and Observation Only Model (right). Each point represents a distinct trial and color shows inference type. Black lines show the best linear fit lines between model and participants with 95% CI bands in gray.

the Full model captures this variability, we ran correlations between the Full model and average participants' responses within the sets of trials for which the Recommendation only model makes identical predictions (preregistered; see Figure 3 middle panel). We found a positive correlation for all sets of trials, which is significantly greater than what would be expected by chance (.0002%). These results suggest that the Recommendation Only model captures the broad, aggregate pattern of participant judgments, but failed to capture more nuanced patterns that the Full Model was able to capture. This suggests that people consider both an agent's observations and their subsequent action to infer their prior belief.

Experiment 2

Experiment 1 (preregistered) provided an initial test of our model by examining how people infer others' prior beliefs (i.e., which team the player is on) to make sense of their apparent beliefs. Experiment 2 (not preregistered) investigates whether participants rely on the same inferential process when making judgments about others' *biases*. This experiment was highly similar to Experiment 1, except participants were tasked with inferring which bias the coach holds, rather than which team the coach thinks the player is on.

Methods

Participants 86 adults (M_{Age}(SD) = 38.06(12.6), range: 20-68; 36 women, 37 men, 2 non-binary, 11 no response) were recruited from Prolific. An additional 8 subjects were excluded for failing one or more of the check questions.

Stimuli Same stimuli as in Experiment 1.

Procedure The procedure was similar to Experiment 1, with the following changes to the cover story and measures. First, participants learned that the coaches form biases (negative, neutral, or positive) about each player based on the their physical appearance, instead of learning that the coaches form beliefs about which team the player is on. Second, the average success rates (same %s as Experiment 1) indicated the coaches' expectations for players' performance on the shots given each bias, rather than indicating each team's performance. The success rates for the Beginner team were replaced with the negative bias expectations, Intermediate team with the neutral bias expectations, and Advanced team with the positive bias expectations. Third, we asked participants to judge the coaches' biases ("How likely was it that Coach [name] holds each bias?"), rather than which team the coach thinks the player is on. Participants responded to this question using the same pie chart scale as in Experiment 1, with three sections that each represented one of the biases. All other aspects of the procedure were identical.

Results

As in Experiment 1, each trial produced 5 data points: 2 competence inferences and 3 bias likelihood inferences. We used the same model predictions from Experiment 1, and found that the Full model again showed a strong quantitative fit with participant judgments (*r*=.84, 95% CI: [.78, 88]). Model fit was high for each inference type (competence: *r*=.92, 95% CI: [.88, 95]); bias: *r*=.82, 95% CI: [.74, 88]).

Next, we tested to what extent the alternative models (same as Experiment 1) account for participants' responses. The Observation only model had an overall correlation of *r*=.69, 95% CI: [.59, .76], which was significantly lower than the Full model (δ=.15, 95% CI: [.08,.25]. The Recommendation only model had an overall correlation of *r*=.82, 95% CI: [.77, .87], which was not reliably lower than the Full model (δ=.01, 95% CI: [-.02, .04]). As in Experiment 1, we ran correlations between the Full model and average participants' responses within sets of trials for which the Recommendation only model makes identical predictions (see Figure 3 middle panel). We found a positive correlation for 11 of the 12 sets of trials, which is significantly greater than what would be expected by chance (.0002%).

Collectively, these results suggest that our model indeed captures people's inferences about *biases*. Specifically, our model comparison results suggest that people consider both the coach's observations and their shot recommendation to infer the coach's bias, rather only one or the other.

General Discussion

Here we sought to explain people's ability to detect when someone might be socially biased. We proposed that social bias detection is (at least partially) the process of positing prior beliefs to explain the gaps between the social information someone observed (e.g., watching a player perform well on hard basketball shots) and how they reacted (e.g., recommending that they try something easier). Importantly, this hypothesis entails that judgments about social bias should not depend on the agent's full performance, only those parts of the performance seen by the observer, and hence depend on tracking shared knowledge. Across two experiments, we found a high quantitative fit between model predictions and participants' inferences about others' prior beliefs (Experiment 1) and biases (Experiment 2). Together, these experiments demonstrated that participants attended to the coach's observation, rather than the player's full performance, and their subsequent action when evaluating the coach's bias.

Our experimental paradigm manipulated a coach's observation and their subsequent recommendation (which shot the player should do next). Comparing our full model to alternative, lesioned models revealed that both the coach's observation and recommendation, rather than only one or the other, supported participants' inferences about the coach's belief and which prior bias best explains this belief. Though the Recommendation only model produced a similarly high numerical correlation as the Full model in both experiments, further investigation revealed that the Full model was able to capture more nuanced patterns in participant responses. Furthermore, we found that the Recommendation Only model produced a stronger quantitative fit than the Observation Only model (where the model only considered the coach's observation of the player, not their recommendation), suggesting that the coach's recommendation was a stronger signal for their biases than the coach's observation in our experimental context. Of course, there may be cases where the coach's observation would be more informative than in our scenarios. For example, consider a scenario in which the coach *chose* to observe a specific shot: if they decided to watch the player practice the short shot (but could have chosen the medium or long shot), one might infer that the coach was uncertain that the player could succeed on the easy shot.

Our work opens new questions for future research. First, our experiments and model focused on situations where the space of possible prior beliefs was categorical and known to the observer (i.e., the coach believes that the player is on one of three teams). In the real world, however, we rarely have access to such structured information about the potential prior beliefs of others. Rather, we often make general inferences about the magnitude and direction of social biases (e.g., "the coach seems strongly biased against me," or "the coach seems a little biased in my favor"). In future work, we hope to develop a variation of our model that can make general inferences about the magnitude and direction of social biases, without the need for specific information about possible prior beliefs. Evaluating these more general bias inferences is an important next step for understanding how people detect social biases in more naturalistic situations.

Second, future work can explore how inferences about others' biases inform one's subsequent actions and even self-representations. For example, when the coach has a strong negative prior against the player, one possibility is that the player would then try really hard to counteract this prior by showing lots of positive evidence (e.g., getting many hard shots in). This prediction is related to stereotype threat (Spencer et al., 2016), for which one proposed mechanism is putting in more effort than is necessary, which can ironically lead to underperformance. It is also possible, however, that the player would give up and decide that it's not worth trying to correct the coach's belief. In some cases, people might even wonder if the coach's bias is justified and infer that they must not be as good as they originally thought (i.e., modify how they think about their abilities). Future research should explore how people decide when to take the effort to correct others' beliefs and when to update their own self-representations.

This study is a first step towards understanding how and when people detect others' social biases. We find that people consider whether others' actions are rational given their observations, and what priors or biases are necessary to justify their actions. This work contributes to research showing the negative consequences of social biases, by investigating how we infer them in the first place.

References

Asaba, M., & Gweon, H. (2022). Young children infer and manage what others think about them. *Proceedings of the National Academy of Sciences*, *119*(32), e2105642119.

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.

Bass, I., Mahaffey, E., & Bonawitz, E. (2021). Do you know what i know? children use informants' beliefs about their abilities to calibrate choices during pedagogy. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Dardenne, B., Dumont, M., & Bollier, T. (2007). Insidious dangers of benevolent sexism: consequences for women's performance. *Journal of personality and social psychology*, *93*(5), 764.

Dovidio, J. F., Gaertner, S. E., Kawakami, K., & Hodson, G. (2002). Why can't we just get along? interpersonal biases and interracial distrust. *Cultural diversity and ethnic minority psychology*, *8*(2), 88.

Gopnik, A., Meltzoff, A. N., & Bryant, P. (1997). *Words, thoughts, and theories* (Vol. 1). Mit Press Cambridge, MA.

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.

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

Kleiman-Weiner, M., Shaw, A., & Tenenbaum, J. (2017). Constructing social preferences from anticipated judgments: When impartial inequity is fair and why? In *Cogsci.*

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.

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual review of psychology*, *67*(1), 415–437.

Sue, D. W. (2010). Microaggressions, marginality, and oppression: An introduction.

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A., Nadal, K. L., & Esquilin, M. (2007). Racial microaggressions in everyday life: implications for clinical practice. *American psychologist*, *62*(4), 271.

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