Detecting social biases using mental state inference

Mika Asaba*, Isaac Davis*, Julia Leonard, Julian Jara-Ettinger

Department of Psychology, Yale University

Keywords: social biases, theory of mind, computational modeling

* These authors contributed equally to this work and are listed alphabetically.

Address for correspondence: mika.asaba@yale.edu, isaac.davis@yale.edu

OSF view-only link: https://osf.io/kprz3/?view_only=25d06dd71efb4b8699ab151ab40e4bc5

Acknowledgments: This work was supported by NSF award BCS-2045778, and NSF SBE Postdoctoral Research Fellowship award 2204106.

Abstract

Social biases are prevalent in everyday social interactions, but they are often expressed in subtle ways that can make them difficult to detect. Yet, intuitively, people can often recognize when they are the subject of a bias, even in the absence of any overt behavior. How do we do this? While much research has focused on the negative consequences of being the subject of a bias, less is known about the cognitive mechanisms that allow people to explicitly detect biases in the first place. In this paper, we propose an account of bias detection which is grounded on mental state representations. We propose that people infer biases by detecting a gap between expected unbiased behavior and observed real-world behavior, which in turns reveals the hidden biases influencing other people's beliefs. We present a formal computational model of this account and, across three preregistered studies (n=720 total), we show that this model captures participants' inferences about an observer's prior beliefs (Experiment 1), general social biases (Experiment 2), and specific real-world biases (Experiments 3a–3c). Moreover, our model captures key patterns of variance in participant responses which simpler alternative models fail to capture. These findings highlight the role of Theory of Mind in social bias detection, and broaden our understanding of the human capacity to detect and reason about implicit prejudices.

Statement of Limitations

Here we detail the limitations of our research, including constraints on the generalizability of our findings. First, our sample was entirely from the United States, predominantly white, and we did not have access to participants' socioeconomic information. Thus, we are limited in the extent to which we can generalize our findings to other populations. Second, our stimuli were designed to vary only three variables (i.e., an agent's performance outcomes, an observer's presence during a subset of the outcomes, and the observer's subsequent action towards the agent), and therefore do not capture the full complexity of social interactions that reveal bias in the real world. Third, we measured inferences about a relatively limited set of biases, specifically the valence of the bias (positive bias, negative bias), rather than specific contents of biases (e.g., stereotypes).

1 Introduction

Jeremy Lin became the first Taiwanese American to play in the NBA (National Basketball Association) and is widely known for his contributions to the sport. Yet, at the beginning of his career, Lin struggled to receive recognition for his abilities. For example, as a high school basketball player, he was named the Northern California Player of the Year, but he received zero college scholarships. As a college player, he broke several Ivy League records in points, rebounds, assists, and steals, yet he received zero offers in the NBA draft. Thus, at least in his early career, Lin's accolades were at odds with recruiting decisions—how can we make sense of this?

Commentators have interpreted that Lin's early career struggles were due to racial biases and discrimination, evidenced by the lack of recognition and anti-Asian slurs he received (*38 at the Garden*, 2022). Biases like this are prevalent in everyday life. Indeed, much prior work in social psychology has demonstrated how, given the exact same evidence (e.g., identical job applications, the same feedback from others), people make stereotyped judgments about others (Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012) and even themselves (Coffman, Collis, & Kulkarni, 2019). Furthermore, people often rely on information about an individual's social group (e.g., explicitly provided base rate information) to evaluate the individual, even when such information is irrelevant to the task (Cao, Kleiman-Weiner, & Banaji, 2017, 2019). People may not be aware that they are holding these biases——explicit biases (e.g., overt sexism, racism) have decreased over the past few decades, yet subtle biases are still held by highly egalitarian individuals and exhibited by both advantaged and disadvantaged groups (e.g., men and women in the US can be sexist, Nosek, Banaji, & Greenwald, 2002).

Social biases can impact all aspects of people's lives, including their self-perceptions (Coffman et al., 2019), mental health (Monahan, Macdonald, Lytle, Apriceno, & Levy, 2020), social life (Aboud, 2003), and even career outcomes (Moss-Racusin et al., 2012). Investigating the consequences of stereotypes has been a rich and important topic of study in social psychology. For example, the study of stereotype threat has been devoted to studying how even the concern about being stereotyped or confirming stereotypes deeply impacts people's actions and performance on important tasks (Spencer, Logel, & Davies, 2016; Steele & Aronson, 1995). However, despite much focus on the effects of biases, much less is known about how people detect biases in the first place.

In this paper, we aim to fill this gap by investigating the cognitive processes that enable us to detect and reason about other peoples' biases. That is, rather than test for the existence of social biases and their consequences on people's behavior, the goal of the current paper is to characterize the experience of thinking that someone is biased. We take a computational approach to this problem. As recent work has shown (e.g., Awad et al., 2022; Davis, Carlson, Jara-Ettinger, & Dunham, 2023; Gershman, Pouncy, & Gweon, 2017; Siegel, Mathys, Rutledge, & Crockett, 2018; for a review, see Cushman, 2023), research into social psychology can benefit from computational modeling in several ways. First, formalizing verbal theories as computational models enables us to ensure that the proposal is internally consistent (i.e., that it does indeed explain the target behavior and does not contain any hidden logical contradictions). This further enables us to modify assumptions in the model to explore how closely related alternative accounts make similar and different predictions. Second, implementing a theory as a computational model enables us to generate exact quantitative predictions about how people should behave according to the theory. The ability to go beyond qualitative effects to make precise, quantitative predictions is particularly useful in situations where people might have graded intuitions. In social judgments (including the bias inferences we study here), people might go beyond simply inferring that someone has or lacks a certain bias, and they might infer relative degrees to which how much someone is biased. Our computational approach enables us to predict the exact likelihoods of biases that people should infer in different situations, which helps us test our theory more rigorously than would be otherwise possible.

To help motivate our account, consider the following scenario: suppose you are trying out for a basketball team as a new player. After watching your tryouts, a coach will assign you to one of three teams, which are separated by skill level. Ideally, the coach will base their judgment solely on an accurate assessment of your skill and assign you to the appropriate team. However, it is also possible that the coach's decision will be influenced by some social bias—for example, the coach might hold the prior belief that people from certain racial groups are "naturally" more or less skilled at basketball than others, as was the case with Lin. How could you determine whether the coach has such a bias, and what that bias is? One potentially important cue to someone's biases about you is the way they treat you relative to the way they treat others. Indeed, it is empirically well established that students are sensitive to such selective treatment and will explicitly point out that a teacher has a bias against their social group when they perceive unequal treatment from that teacher (Boysen, Vogel, Cope, & Hubbard, 2009; Marcus, Gross, & Seefeldt, 1991; Wayman, 2002). Thus, in the context of our example, you might infer that the coach has a negative bias about you if they assign most players to the intermediate or advanced team but assign you to a beginner's team.

However, the coach's differential treatment of players might not always reflect a social bias. Suppose, for example, that the coach watched each of the previous players make several consecutive three-point baskets. When it is your turn to try out, however, the coach watches you miss two easy lay-ups in a row. In this case, it does not seem unreasonable for the coach to infer that you are considerably less skilled than the previous players. Of course, this evidence may not accurately reflect your true skill level—even skilled players occasionally fumble and make mistakes. But, if the coach must make a recommendation based on the available evidence, a "beginner" recommendation does not seem to imply a negative bias. This demonstrates a second key factor that may be relevant for inferring an observer's bias: the evidence about you that the observer has access to.

This provides an intuitive motivation for our account of bias detection: We propose that detecting an observer's biases involves considering what evidence about you the observer has access to, what beliefs an unbiased observer ought to form based on that evidence, and how the observer's actual beliefs (as reflected by how they treat you) compare against the hypothetical unbiased beliefs. Thus, 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.

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).

At the same time, explaining how we infer social biases through Theory of Mind poses a challenge. Past work suggests that, when we infer other people's beliefs, we assume that other people reason and behave rationally (e.g., Baker, Jara-Ettinger, Saxe, and Tenenbaum 2017; Gergely and Csibra 2003; Jara-Ettinger, Gweon, Schulz, and Tenenbaum 2016; Lucas et al. 2014). However, social biases, by definition, lack a rational justification. To solve this tension, we propose that bias inference involves two subtly different types of belief reasoning. First, inferring an agent's *current* belief, based on an assumption that their behavior is a rational expression of what they think (e.g.: a coach's belief about a player's competence, based on their recommendation to the player). And, second, inferring the agent's *prior* belief, based on the discrepancy between the evidence the agent received, and the contents of their current belief (e.g.: what assumptions the coach must have started with to explain their current belief in light of what the coach observed). Returning to the basketball context, if the coach observed you successfully make a few 3-pointers, we might expect the coach to rationally update their beliefs given your performance, and think positively about your 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 your performance.

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. A first possibility is that people 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, see Boysen et al., 2009; Marcus et al., 1991; Wayman, 2002). If this is the case, we would expect people to infer a bias based only on people's actions (e.g., a coach's feedback), and ignore what they saw.

We therefore consider an alternative model that makes an inference solely based on the evaluator's feedback.

A second possibility is that people might simply think of bias as either a negative or positive belief about someone, rather than a *pre-existing* belief. For example, if the coach observes a player miss several easy shots, it might seem intuitive to predict that the coach will *now* have a "negative bias" against the player based on this evidence. To clarify whether participants understand this distinction between a prior bias and a posterior belief, our second alternative model ignores the observer's feedback, and simply returns the posterior belief (updated based on the observed performance) as a prediction of the observer's bias. For example, if the coach observes a poor performance, and thus forms a negatively-skewed belief about the player's competence, the alternative model will infer a negative bias.

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 accounts that might underlie bias detection. Our model performs a joint inference over an actor's (e.g., a basketball player's) true competence, an observer's (e.g., a basketball coach's) beliefs about the actor's competence, and the priors that the observer must have had to justify this belief. We evaluate our model in three behavioral experiments. Experiment 1 provides the initial validation of the model by testing inferences about which prior belief an observer 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). Finally, Experiment 3 investigates the extent to which these inferences extend to inferences about real-world biases, i.e., about race (3a), gender (3b), and age (3c). Throughout, we relied on simple, hypothetical scenarios that grant us full control over the the exact information that participants are given and allow us to parametrically vary variables of interest (e.g., the player's performance, or which shot the coach observed). This approach enables us to properly test our hypothesis against alternatives and thus, evaluate the principles by which people infer social biases. Furthermore, given our goal to understand how people detect that someone is biased, we relied on explicit judgments about biases (i.e., whether the coach held a positive or negative bias, or no bias), rather than implicit measures.

2 Computational Framework

For simplicity, we explain our computational model in the context of our experimental tasks, all of which have the same two-stage structure (instantiated in three different contexts). In the first stage, an *actor* agent makes some initial attempts at a skill-based task (e.g., a basketball player shoots some baskets) while an *observer* agent watches one of the attempts. We assume that the observer has some initial belief about the actor's competence. This initial belief may reflect some biases. For example, the observer might have been told some possibly incorrect information about the actor (e.g., a coach might have been told that a player is on the advanced team, and therefore has higher prior expectations about the player's skill level). The observer may also hold biases that relate some surface feature of the actor's identity to the actor's competence (e.g., a teacher might hold a bias that female students are generally worse at math).

Tasks can have three different difficulties—easy, medium, or hard—and the actor's attempts result in a set of outcomes $O = \{o_1, \ldots, o_n\}$, each specifying the difficulty of the task as well as the outcome (a binary "success" or "failure"). Critically, the *observer* agent observes some, but not all, of these outcomes $O_{obs} \subset O$ (e.g., a basketball coach observes one of the player's shots). After this first stage, the model infers both the actor's belief and the observer's probable beliefs about the actor's ability (which will depend on the unknown prior bias), and these beliefs might conflict with each other due to differences in the observed evidence (e.g., the actor succeeded on the easy and hard tasks, but the observer only saw the easy tasks). This belief formation is performed through the assumption that both the player and the observer rationally update their beliefs about the actor based on the observed performance O_{obs} (following the evidence that people expect each other to update beliefs rationally in light of evidence; Baker et al. 2017).

In the second stage, the observer suggests that the actor make one final attempt on one of the tasks (easy, medium, or hard). The observer's recommendation is based on their belief about the

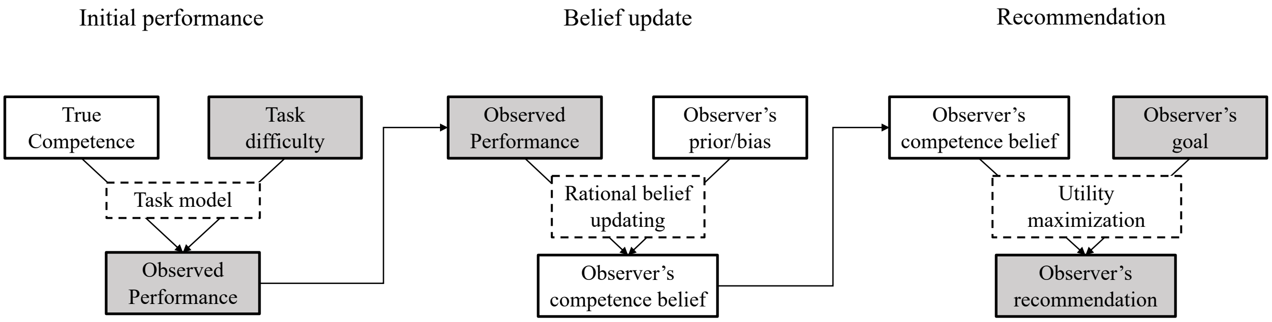


Figure 1: Schematic of the three components of our computational model. Dashed lines denote processes, and solid lines denote variables. Gray boxes indicate observable variables, and white boxes denote unobservable variables.

actor's skill, with the goal of getting the actor to succeed at the hardest task they can reliably do (e.g., recommending the easiest task implies that observer doesn't think the actor would succeed at either of the harder tasks). We modeled this full procedure as a causal mental model (i.e., a *generative 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 et al., 2017; Jara-Ettinger et al., 2016; Jern, Lucas, & Kemp, 2017; Lucas et al., 2014).

2.1 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 observer chooses a recommendation.

2.1.1 Part 1: Task Model

The task model assumes that the actor's performance is determined by their true competence and the task's known difficulty. Formally, the actor has some skill level $s \in (0,1)$, and each task has some difficulty level $d \in (0,1)$. Given these parameters, the actor's probability of succeeding at a task is given by $P(hit|s,d) = 1/(1+exp(-\beta*(s-d)))$, i.e.: a logistic function with growth rate β. We further assume that each attempt has a small probability ε of succeeding or failing, independent of skill or difficulty. This allows the observer to selectively discount subsets of outcomes when they believe they are best explained by luck rather than skill. Our initial analysis revealed that the exact value of this luck parameter had little effect on the model's pattern of inferences, so we fixed ε at a small value (.05). This has a minimal effect on the observer's overall inferences, but provides enough flexibility to discount certain observations as reflecting good or bad luck.

2.1.2 Part 2: Belief Updates

The task model produces a set of outcomes O, each specifying a difficulty level and a result (success or failure), of which the observer sees a subset O_{obs}. Both agents then update their beliefs about the actor's skill level via Bayesian inference:

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

where $P(O|s)$ 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 actor's skill.

To constrain the space of possible prior beliefs, our task used a structure where observers could have one of three pre-existing biases, ("negative", "neutral", or "positive"), each associated with a known prior distribution over skill level ($P_{neg}(s)$, $P_{neut}(s)$, or $P_{pos}(s)$). Rather than attempting to estimate participants' expectations about these three priors, we modeled these beliefs as Beta distributions that were shown to participants as part of the instructions. P_{neg} is skewed towards lower competence values (modeled as a $Beta(1,2)$ distribution), P_{pos} is skewed towards higher competence values (modeled as a $Beta(2,1)$ distribution), and P_{neut} is symmetric and neither skewed towards higher nor lower values (modeled as a $Beta(1,1)$ distribution).

We assume that the actor is more familiar with their own skill level than the observer, and therefore always uses the correct distribution P_{true} as the skill prior for updating (with each of the three possible priors being equally likely for any given player). The observer uses P_{obs} as the skill prior, where *obs* is the initial belief that the observer holds. This prior belief is how we represent the observer's potential bias (e.g., a prior belief that a player must be from the advanced team without evidence to support that belief).

2.1.3 Part 3: Observer's Recommendation

The final part of the model generates a recommendation by combining the observer's beliefs about the actor's competence with their goal—maximizing the expected number of points that the player will receive from the task, where an easy task is worth 1 point, a medium task is 2 points, and a hard task is 3 points. The expected value of an attempt is given by $P(success|s,d_i) * V(i)$, where $V(i)$ is the value of succeeding at shot i, and d_i is the known difficulty of the task. Because the observer does not know the actor's exact skill level s, the expected value of recommendation i is integrated over the observer's belief about skill:

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

where $P(s|O_{obs})$ is the observer's belief in the actor's skill level (obtained from Eq.1). Given an expected value associated with each possible recommendation, the observer uses a softMax decision policy to make a recommendation. We tuned this temperature parameter to $\beta = 2$ via maximum likelihood estimation applied to data from a small initial pilot study.

2.2 Inference and alternate models

Our experimental task asks participants to infer a) the actor's belief about their own skill, b) the observer's belief about the actor's skill, and c) the likelihood that the observer held each of the prior biases about the actor (before observing the actor'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 actor's performance, the observer's observation, and the observer'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.

In addition to our main model predictions, we generate predictions from two alternate models that each lesion one core component of the main model. The *Recommendation only model* assumes that, instead of performing the full mental state inference, people determine bias based on the recommendation alone. Under this model, a low recommendation implies a negative bias, a high recommendation implies a positive bias, and an intermediate implies no bias, regardless of the observed performance. Conversely, the *Observation only model* ignores the observer's recommendation, and infers a bias based solely on the observed performance (e.g.: observing a weak performance yields a negative bias against the actor). Under this model, an observation of negative performance (a miss) implies a negative bias, and positive performance (a hit) implies a positive bias. This model allows us to ensure that participants recognize that the observer's bias must precede the observation, and cannot be directly influenced by the observation.

While all of the model parameters and predictions were pre-registered, we later identified a minor bug in the model implementation. Our paper therefore presents the same preregistered analytical plan, but with the model predictions of the corrected model. All of the key results and conclusions are identical under the pre-registered model predictions (see SI for full details).

3 Behavioral Experiments

Our behavioral experiments (all preregistered) have three goals. First, we validate the structure of our computational model and test if it can explain how people infer another agent's prior beliefs (Experiment 1). Second, we test whether this model explains how people infer others' social biases (Experiment 2). Third, we examine whether our model captures inferences about others' biases across a series of contexts that involve real-world biases (Experiments 3a, 3b, and 3c). All analyses reported were preregistered, unless reported otherwise.

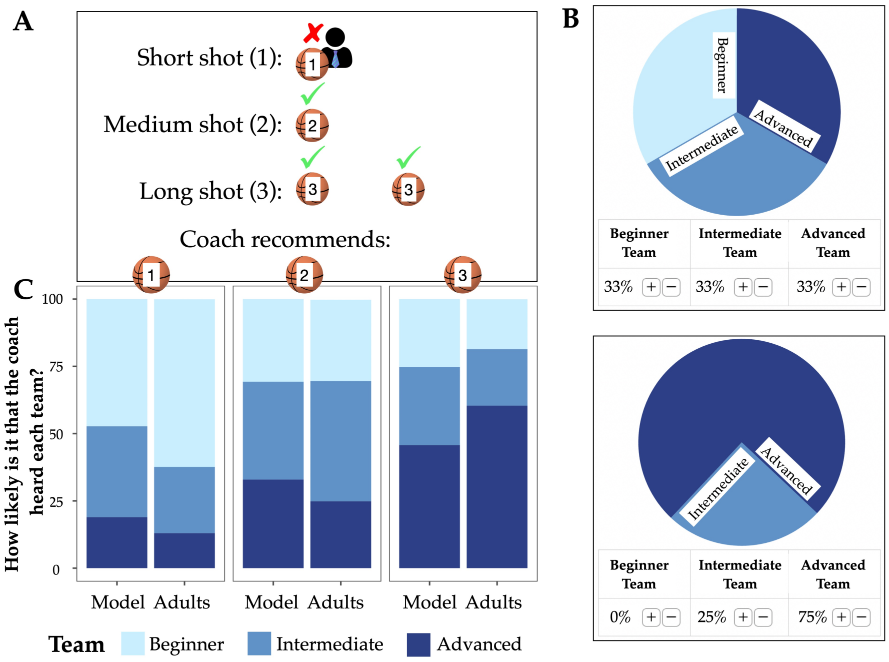


Figure 2: Example Experiment 1 stimuli (A-B) and results (C). (A) depicts a trial in which the player fails at the short shot, succeeds at the medium shot, and succeeds twice at the long shot; the coach only observes failure on the short shot (right). After the player's performance and coach's observation, the coach recommended a shot for the player to do next: the Short, Medium, or Long shot. Finally, participants were tasked with inferring the coach's team prior. (B) shows the pie chart rating scale that participants used for inferring the coach's prior belief; the pie chart was always initialized with equal proportion for each team (top), and participants can click the '+' or '-' buttons to respond or drag the pie chart sections in any way, including removing a section entirely (bottom). (C) shows *Full model* predictions and participant judgments (means and 95% bootstrapped CIs) for the team questions for the trials shown.

4 Experiment 1

According to our proposal, people attribute biases to others by inferring their prior beliefs. To test this possibility, it is first necessary to test whether our model accurately captures how people infer prior beliefs in a general sense. We test this in Experiment 1. Participants were presented with a cover story that matched the structure of our model, in a context that simply required identifying the prior beliefs of different coaches, with no mention of bias. Because we are interested in the explicit experience of detecting others' priors, we asked participants to explicitly report the prior beliefs. We measured participants' inferences about i) the actor's belief about their skill, ii) the observer's belief about the actor's skill, and iii) the observer's prior belief about which skill level group the actor is in.

4.1 Methods

4.2 Transparency and Openness

Studies 1, 2, and 3a-3c were preregistered, including study design, hypotheses, sample size, data collection, exclusion criteria, and analysis plan. All data, analysis code, research materials, and preregistrations of study and analysis plans are available at https://osf.io/kprz3/?view_only=25d06dd71efb4b8699a Data for all studies were processed, analyzed, and visualized using R 4.1.3 (R Core Team, 2022). We report all data exclusions (if any), all manipulations, and all measures in the study. Our reporting adheres to the APA Style Journal Article Reporting Standards.

4.2.1 Participants

150 adults (M_{Age}(SD) = 33.8(12.5) years, range: 18-71 years) from the United States were recruited from Prolific. Participants were 47.3% women, 45.3% men, 2.7% non-binary, and 4.7% no response, and 68% White, 10% Hispanic/Latine, 9.3% Asian, 6% Black, 3.3% American Indian or Alaska Native, .7% Middle Eastern, and 2.7% no response (gender and race/ethnicity self-reported). An additional 8 subjects were recruited and excluded for failing one or more comprehension check questions (preregistered criteria).

4.2.2 Stimuli

Experiment 1 used 30 different stimuli (see Figure 2A for examples). Each stimulus depicted: a player's performance on four shots, each of which could be the Short, Medium, or Long shot; a coach who observed only one of those shots; and the coach's recommendation for which of the 3 shots the player should attempt next. 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 four shots the player attempted, made, and missed, with the purpose of creating an even distribution of player competence from low to high. The coach observations (of one of the player's shots) were designed to create an even distribution of the coach's beliefs about the player competence from low to high. For each of the 10 combinations of player performance and coach observation, we created three distinct trials for each of the coach's recommendations (Short, Medium, or Long), for 30 trials total. Each participant saw 10 trials, which always contained the 10 combinations of player performance and coach observation, and coach recommendation for each trial varied across participants. See SI for full description of the stimuli space.

4.2.3 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). For the Short distance shot, the success rates were 60% (Beginner), 80% (Intermediate), and 90% (Advanced). For the Medium distance shot, the success rates were 30% (Beginner), 60% (Intermediate), and 80% (Advanced). Finally, for the Long distance shot, they were 20% (Beginner), 30% (Intermediate), and 60% (Advanced). Each team's success rates were shown as the proportion of agents, out of 10, that were in the team's color. For example, for the Beginner's team success rate for the Short distance shot, participants saw six agents in the Beginner team's color (yellow) and four blank agents. Participants were required to correctly type in the correct success rate percentage before continuing the tutorial, and had unlimited chances to do so.

Then, participants learned that the players are meeting new coaches today—the players told the coaches which team they are on, but the coaches may have misheard. This means that each coach holds a pre-existing belief about a player's competence based on their team, but this belief could be incorrect. Participants were told that they would watch a player practice in the presence of a coach, who would occasionally leave and therefore not watch the entire practice. 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 four comprehension check questions correctly, to ensure that they understood the story. If they answered incorrectly, they were prompted to try again and they were given unlimited chances to complete the survey (but always being required to re-read the instructions whenever they got at least one question wrong; see SI for comprehension check questions). Finally, participants did a training with the pie chart scale used for the team inferences.

After the tutorial, participants underwent 10 trials (randomized order, see Fig. 2A for an example). 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. See Figure 2A.

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 and had unlimited chances to do so. Then, they responded to three test questions. The first two test questions concerned the player's 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 2B).

After completing the 10 trials, participants responded to the same comprehension check questions as in the tutorial at the beginning of the task. Participants who responded incorrectly to at least one question were excluded from analyses (preregistered criteria).

4.3 Results

Each of the 30 trials produced 5 data points: 2 competence inferences (one for the player, one for the coach), and 3 inferences about the coaches' prior beliefs (nonindependent; one per team). Our *Full model* showed an overall strong quantitative fit with participant judgments (r=.92, 95% CI: [.89, .94]), and model fit was similar for each inference type (competence: r=.96, 95% CI [.93, .97]; team: r=.83, 95% CI [.82, .92]). See Figure 2C for results from a specific trial and Figure 3 for overall results.

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 factors (either the coach's observation or their recommendation). We tested these accounts with two models that lesioned off one variable: the *Observation only model*(which considered the coach's observation and ignored their recommendation) and *Recommendation only model* (which considered the recommendation and ignored the observation). The *Observation only model* had an overall correlation of r=.55 (95% CI: [.43,.65]), which was reliably lower than the *Full model* (Δ = .37, 95% CI: [.26, .50]).

The *Recommendation only model* had an overall correlation of r=.91 (95% CI: [.87,.93]), which was not reliably different from the *Full model* (Δ = .01, 95% CI: [-.01, .03]). Finally, as an exploratory analysis, we examined which model produced the highest fit for each participant. To do this, we ran correlations between each participant's responses and predictions from the *Full*, *Recommendation only*, and *Observation only model*s. We found that 47% participants were best fit by the *Full model*, 32% by the *Recommendation only model*, and 21% by the *Observation only model.*

Collectively, this experiment revealed two findings. First, we found that people can accurately infer other agent's prior beliefs based on what the agent sees and suggests. This validates that people have this general capacity for belief inference, which enables us to test if the same capacity is at work when inferring social biases (in Experiments 2 and 3a—3c). Our second finding is that the *Recommendation only model* produces very similar predictions to the *Full model* (and

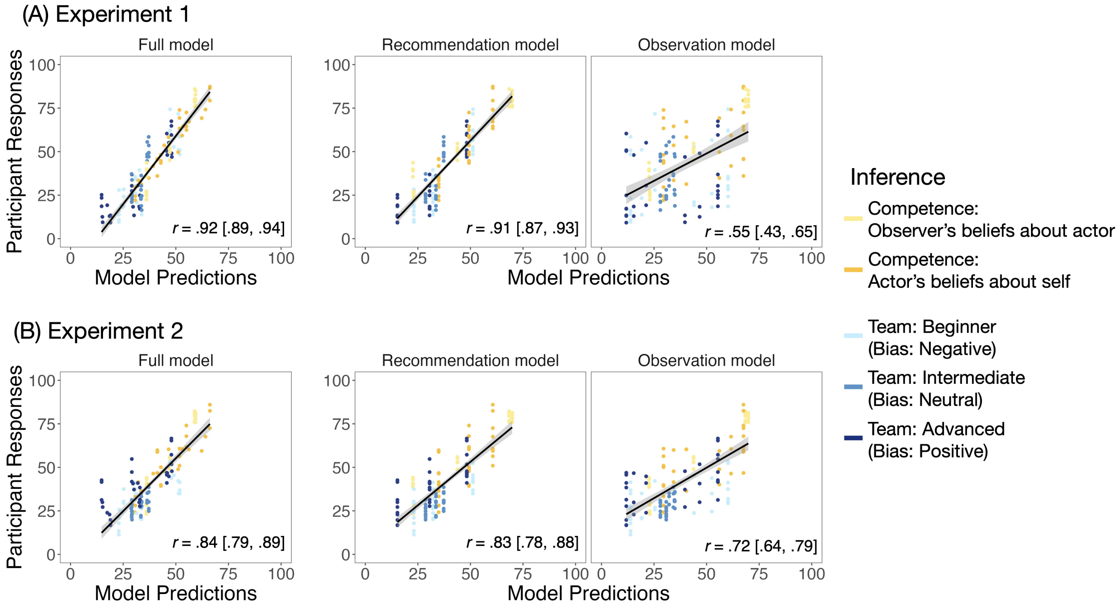


Figure 3: Experiment 1 (A) and Experiment 2 (B) model correlations with participant judgments: *Full model* (left), *Recommendation only model* (middle), and *Observation only model* (right). The measures between the experiments were identical, except that participants provided probability judgments for which team the coach thinks the player is on (Beginner, Intermediate, Advanced), whereas in Experiment 2, participants provided probability judgments for which bias the coach holds (Negative, Neutral, or Positive). 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.

they were hence indistinguishable), and it is therefore possible to approximate complex belief inferences by attending only to the recommendation, although our subject-level analysis suggest that most participants show a closer alignment with the *Full model*. This is a possibility that we further explore in the next two experiments.

5 Experiment 2

Experiment 1 established that people can infer an agent's prior prior beliefs to make sense of their behavior. Experiment 2 investigates whether participants rely on the same inferential process when making judgments about others' *biases*. This experiment therefore tests our hypothesis that bias detection is equivalent to inferring an agent's prior beliefs. 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.

5.1 Methods

5.1.1 Participants

152 adults (M_{Age}(SD) = 39.28(11.8), range: 18-69) from the United States were recruited from Prolific. Participants were 50% women, 46.1% men, and 3.9% other, and 73.7% White, 11.8% Black, 9.2% Hispanic/Latine, 3.3% Asian, 1.3% Alaska Native, and .6% (n=1) no response (gender and race/ethnicity self-reported). An additional 18 subjects were excluded for failing one or more of the check questions.

5.1.2 Stimuli

Same trials as in Experiment 1.

5.1.3 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 their physical appearance (but the dimension of physical appearance was intentionally left open to avoid external beliefs to influence the task in this first test), 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' subjective expectations for players' performance on the shots given each bias, rather than indicating each team's objective performance. Specifically, 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, including participants responding to check questions at the end of the tutorial (unlimited chances) and again at the end of the task (exclusion criteria).

5.2 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: [.79, 89]). Model fit was high for each inference type (competence: *r*=.92, 95% CI: [.86, .95]); bias: *r*=.61, 95% CI: [.46, 73]).

Next, we tested to what extent the lesioned, alternative models (same as Experiment 1) account for participants' responses. The *Observation only model* had an overall correlation of *r*=.72, 95% CI: [.64, .79], which was reliably lower than the *Full model* (Δ=.12, 95% CI: [.05,.21]. The *Recommendation only model* had an overall correlation of *r*=.83, 95% CI: [.78, .88], which was not reliably lower than the *Full model* (Δ=.01, 95% CI: [-.02, .04]).

One important feature of the *recommendation only* model is that there are collections of trials for which the model produces identical inferences, but people do not. This can be visualized in Figure 3 where the *recommendation only* model results show columns of data: Clusters of trials where the model makes the same prediction (henceforth *clusters of indifference*), resulting in the same value on the x axis, but people make different judgments, resulting in a range of values on the y axis. If the *Recommendation only model* is correct, then this variability should only reflect nothing more than noise in participant responses. Alternatively, it is possible that this variability reflects more nuanced participant reasoning that the *recommendation only* model does not capture, but that our *Full model* does. To test this possibility, we ran correlations between the *Full model* and average participants' responses within sets of trials for which the *Recommendation only model* makes identical predictions (12 sets of trials). We found a positive correlation for only 6 of the 12 sets of trials, which is not significantly greater than what would be expected by chance (p=1, Binomial Test). Finally, we calculated model fits for each participant (exploratory). As in Experiment 1, the *Full model* fit best for the highest proportion of participants (42%). The *Recommendation only model* fit best for only 21% participants, and the *Observation only model* for 37% participants.

Collectively, these results suggest that our *Full model* indeed captures people's inferences about *biases*. However, as in Experiment 1, the *Full model* and the *Recommendation only model* produced similar fits to participant data. When exploring model fits at the individual-level, we once again found that the *Full model* fit the highest proportion of participants the best, compared to the two lesioned models. However, the evidence is inconclusive so far about whether participants inferred biases by attending only to the agent's recommendation, or whether they attempted to infer the agent's prior beliefs.

6 Experiments 3a, 3b, and 3c

In the last set of experiments, we tested to what extent our model captures participants' inferences about biases, in cases where the observer is specifically making a judgment about the actor based on their social group membership (e.g., based on their race, gender, or age), rather than on their physical appearance in general (as in Experiment 2). For each experiment, we tested a specific context and group membership, for which past research has shown that people experienced discrimination and/or stereotype threat (Experiment 3a: race and athletic skills, Stone, Lynch, Sjomeling, & Darley, 1999; Experiment 3b: gender and math skills, Muzzatti & Agnoli, 2007; Experiment 3c: age and technology skills, Fritzsche, DeRouin, & Salas, 2009; Lamont, Swift, & Abrams, 2015). Each experiment was run separately but given that they shared the exact same trial structure and model predictions, we present them as a group here.

6.1 Methods

6.1.1 Participants

We report participant demographic information for each experiment separately. For all experiments, we recruited 175 participants with the aim of having 150 participants in the final analyses after exclusions (preregistered).

Experiment 3a: 151 adults (M_{Age}(SD) = 37.1(13.7), range: 18-83) from the United States were recruited from Prolific. Participants were 50.9% woman, 44.4% man, 2% non-binary, .6% agender, and 2% prefer not to respond, and 68.2% White, 13.2% Asian, 9.9% Black, 6% Hispanic/Latine, 1.3% American Indian or Alaska Native, and 1.3% no response (gender and race/ethnicity self-reported). An additional 24 participants were recruited and excluded for failing one or more comprehension check questions (preregistered criteria).

Experiment 3b: 140 adults (M_{Age}(SD) = 37.8(12.3), range: 19-74) from the United States were recruited from Prolific. Participants were 47.9% men, 45% women, 2.9% non-binary, .7% genderqueer, and 1.4% prefer not to respond, and 62.9% White, 12.1% Black, 11.4% Hispanic/Latine, 10% Asian, 2.1% American Indian or Alaska Native, .7% Middle Eastern, and .7% no response (gender and race/ethnicity self-reported). An additional 35 participants were recruited and excluded for failing one or more comprehension check questions (preregistered criteria).

Experiment 3c: 127 adults (M_{Age}(SD) = 39.5(12.7), range: 20-67) were recruited from Prolific. Participants were 48% women, 48% men, .8% non-binary, .8% genderqueer, and 2.4% prefer not to respond, and 80.3% White, 8.7% Black, 6.3% Asian, 1.6% American Indian or Alaska Native, 1.6% Hispanic/Latine, .8% Middle Eastern, and .8% no response (gender and race/ethnicity self-reported). An additional 48 participants were recruited and excluded for failing one or more comprehension check questions (preregistered criteria).

6.1.2 Stimuli

Same trials as in Experiments 1 and 2 were used, with minor wording or image changes based on the cover story (see Figure 4). Experiment 3a concerned scenarios about basketball coaches making judgments about players' basketball skill based on their race; Experiment 3b, about math teachers making judgments about students' math abilities based on their gender; and Experiment 3c, about tech recruiters making judgments about candidates' programming skills based on their age. Experiment 3a kept the basketball icon to depict the basketball shots, Experiment 3b used a calculator to depict the math problems, and Experiment 3c used a computer to depict the programming problems. Across all experiments, we did not provide names of the player/student/candidate in order for their social groups to be unknown to participants; instead, we named them each by a letter (e.g., "Player A", "Player B", etc.)

6.1.3 Procedure

The procedure for each experiment was similar to Experiment 2, with the following changes to the cover story and measures. Below, we briefly describe the cover stories for each experiment (see Figure 4):

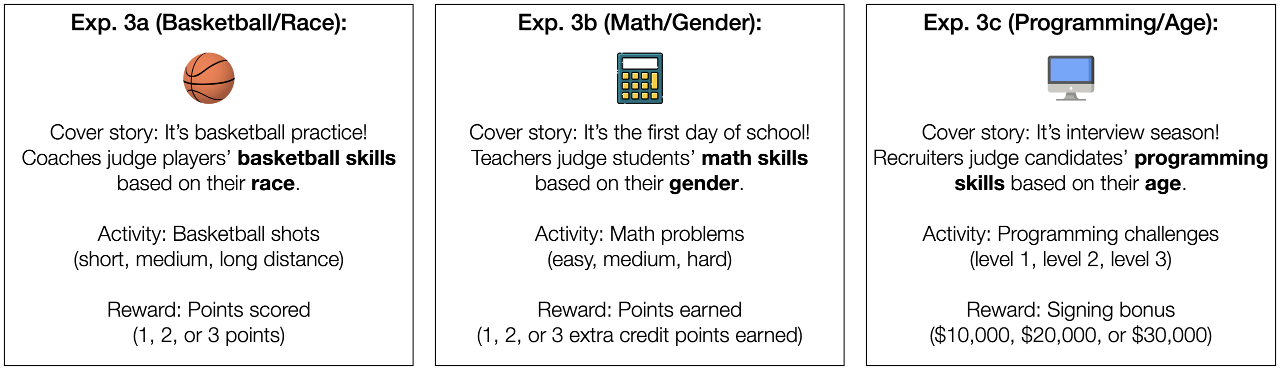
Experiment 3a: Participants read stories about basketball coaches and players at basketball practice. They learned that the basketball coaches are meeting the players for the first time, and that the coaches initially judge the players based on their race (White or Black).

Experiment 3b: Participants read stories about teachers and students in a math class. They learned that the teachers are meeting the students for the first time, and the teachers initially judge the students based on their gender (boys or girls).

Experiment 3c: Participants read stories about recruiters and software engineering candidates at a tech company. They learned that the recruiters are meeting the candidates for the first time, and the recruiters initially judge the candidates based on their age (20 year-olds or 60 year-olds).

For all experiments, the actor (the player, student, or candidate) always made four attempts on tasks (basketball shots, math problems, programming challenges) with three difficulty levels. However, the observer (the coach, teacher, or recruiter) only saw one of the actor's attempts. Furthermore, the observer always recommends one final problem and the actor's reward depends on the difficulty of the problem (points, extra credit, or bonus money, see Figure 4).

On each trial, participants responded to the same test questions as in Experiment 2: (i) the actor's competence, (ii) the observer's beliefs about the player's competence, and (iii) the observer's biases (pie chart scale). Finally, we included two exploratory questions. First, we included one question about the student's identity on each trial. We asked participants to predict the race/gender/age (depending on the experiment) of the player/student/candidate (7-point scale from Most likely Black/Boy/20s to Most Likely White/Girl/60s). We asked this question to check whether participants were genuinely thinking about the actor's race/gender/age and the observer's specific biases, rather than ignoring these details in the trials. Second, at the end of the task, we asked participants how much they endorse the stereotype that the cover story was about. Participants read a stereotype, "Boys are more skilled at math than girls", "Black people are more skilled at basketball than White people", or "Young people are more skilled at programming than old people", and then were asked to rate the accuracy of the stereotype (on a scale of 1-7, from not at all accurate to extremely accurate). The purpose of this question was to explore whether model fits



Exp. 3a (Basketball/Race):

Cover story: It's basketball practice! Coaches judge players' **basketball skills** based on their **race**.

Activity: Basketball shots (short, medium, long distance)

Reward: Points scored (1, 2, or 3 points)

Exp. 3b (Math/Gender):

Cover story: It's the first day of school! Teachers judge students' **math skills** based on their **gender**.

Activity: Math problems (easy, medium, hard)

Reward: Points earned (1, 2, or 3 extra credit points earned)

Exp. 3c (Programming/Age):

Cover story: It's interview season! Recruiters judge candidates' **programming skills** based on their **age**.

Activity: Programming challenges (level 1, level 2, level 3)

Reward: Signing bonus ($10,000, $20,000, or $30,000)

Figure 4: Cover stories for Experiments 3a, 3b, and 3c. Each experiment concerned a different activity (basketball, math, or programming), and observers made judgments about the actor based on different aspects of their identity (race, gender, age). The structure of the scenarios were similar, with 3 levels of difficulty for each activity and 3 evenly-spaced levels of reward.

depended on participants' own stereotypes (e.g., whether participants who more strongly endorsed a particular stereotype had worse model fits; see SI for exact question wording and results).

6.2 Results

Figure 5 shows the results from the experiment. The *Full model* showed a strong quantitative fit with participant judgments for Experiment 3a (r=.89, 95% CI: [.85, .92]), Experiment 3b (r=.92, 95% CI: [.89, .94]), and Experiment 3c (r=.89, 95% CI: [.85, .92]). Across experiments, model fits were high for each inference type ($r >$.94 for competence inferences and $r >$.73 for bias inferences for all experiments, see SI for exact values).

Next, we examined the performance of the *Full model* in capturing participant's responses compared to each of the alternative models. For all three experiments, the *Full model* produced reliably higher fits compared to the *Observation only model* ($\Delta >$.20 for all experiments, see SI for exact values). The *Full model*, however, did not consistently produce higher fits than the *Recommendation only model* ($|\Delta| <$ .03; all $\mathrm{CI}_{95\%}$ between $-$.03 and .07, see SI for exact values broken down by experiment).

To test whether the *Full model* captures variability that the *Recommendation only model* does not, we ran correlations between the *Full model* and average participants' responses within clusters

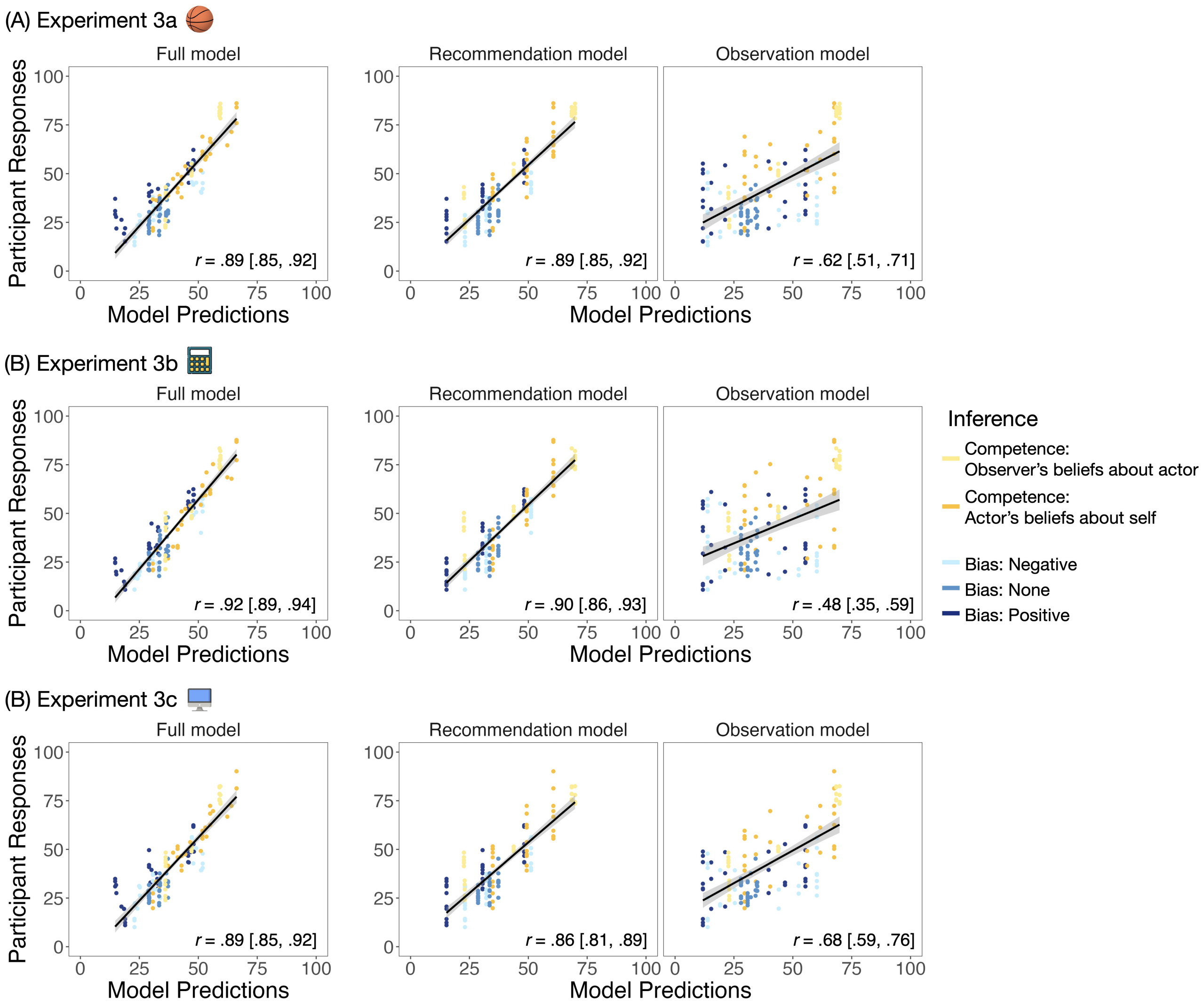


Figure 5: Experiment 3a, 3b, and 3c model correlations with participant judgments: *Full model* (left), *Recommendation only model* (middle), and *Observation only model* (right). Each point represents a distinct trial and color shows inference type. Black lines show the best linear fit lines between model and participants with 95% CI bands in gray. Experiment 3a concerned scenarios about basketball coaches making judgments about players' basketball skill based on their race; Experiment 3b, about math teachers making judgments about students' math abilities based on their gender; and Experiment 3c, about about tech recruiters making judgments about candidates' programming skills based on their age.

of trials for which the *Recommendation only* model makes identical predictions (12 clusters of trials x 3 experiments). Of the 36 clusters, 28 showed a positive correlation with the *Full model* (77.78% of trials; $p = .001$, Binomial Test). 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. That is, participants considered both the coach's observations and their subsequent action to infer their prior belief.

As in the previous experiments, we also calculated model fits for each participant (exploratory). We found that the *Full model* again fit the highest proportion of participants the best (41%), followed by the *Observation only model* (33%) and the *Recommendation only model* (25%).

Then, we asked whether participants' predictions about the identity of the actor were related to their inferences about the observer's biases. That is, we asked whether people's inferences about the direction of a bias were related to their inferences about the social category of the agent. A correlation between these two dimensions would suggest that participants were indeed reasoning about the observer's biases as they relate to real-world stereotypes when responding to the trials. For each trial, we first calculated a difference score for each trial that corresponded to the average difference in participants' positive and negative bias ratings (therefore ranging from -1 to 1). Thus, a value of 1 indicates that every participant inferred a positive bias with full confidence on that trial, a value of -1 indicates that every participant always inferred a negative bias with full confidence on that trial, and values in between reveal relative orientations of bias across participants for the same trial. We then correlated these difference scores with participants' average identity predictions (3a: race of the player, 3b: gender of the student, 3c: age of the candidate; 7-point rating scale for all). We found strong correlations for each experiment (Experiment 3a: r=.80, 95% CI: [.62, .90]; Experiment 3b: r=.74, 95% CI: [.52, .87]; Experiment 3c: r=.84, 95% CI: [.69, .92]). That is, when participants inferred that the observer was more likely to have a negative bias than a positive bias, they were more likely to predict that the basketball player was Black than White (Experiment 3a), the math student was a girl than a boy (Experiment 3b), and the programming candidate was 60 years-old than 20 years-old (Experiment 3c). These results suggest that participants were indeed reasoning about real-world stereotypes when inferring the observer's biases. Finally, we ran correlations between participants' own stereotype endorsements and their individual fits with the *Full model* and did not find evidence of a relationship between the two (see SI for more details and exact values).

7 General Discussion

Social biases are prevalent and negatively impact people's daily lives. Here, we hypothesized 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). Critically, this hypothesis entails that the ability to infer social biases depends on a capacity to track shared knowledge and to infer mental states from behavior. We tested this proposal by implementing a computational model of bias detection through Theory of Mind. We found a high quantitative fit between model predictions and participants' inferences about others' prior beliefs (Experiment 1) and biases (Experiment 2), including several real-world biases (Experiments 3a, b, c). Together, our findings provide a mentalistic account of social bias inference, showing how people consider both what an observer saw and how they reacted to infer their bias.

Note that we found some variability in model fits across experiments. At first glance, one might think that participants had an easier time inferring prior beliefs when they were about the player's team, rather than social biases (*r*=.92 in Experiment 1, compared to *r*=.84 in Experiment 2). However, the model fits in Experiment 3 were of comparative magnitude to the ones in Experiment 1 (*r*=.89, .92, and .89, for Experiments 3a-c, respectively), suggesting that these differences in correlation values may be due to experimental variability.

We also considered two alternative possibilities throughout: (i) whether people assume that some behaviors are intrinsically biased, such that they only consider others' actions to infer bias (*recommendation only* model), and (ii) whether people simply think about biases as positive versus negative beliefs, such that they use others' observations to infer bias (*observation only* model). Our *Full model* (which included both what the observer saw and their subsequent actions) had the best fit at the population level and the subject level in all five studies. At the same time, the *recommendation only* model had surprisingly similar fits to participants, and the *Full model*'s correlation was never significantly higher. Nonetheless, a sensitivity analysis in Experiment 3 showed that the *Full model* did predict unique variance across the three versions that the *Recommendation only model* did not, suggesting that the *Full model* does explain participant judgments better at least in these experiments, although the advantage was small.

This suggests that, in contexts like the ones we considered, attending only to how people behave (as done in the *Recommendation only model*) is a strong cue to people's prior beliefs. If this is representative of other contexts, this opens the possibility that attending to people's behavior might be a simple way to determine when to begin to consider whether someone is biased or not. Such a possibility feels intuitive: rather than constantly attempting to infer other people's priors beliefs, we might only do so when we notice that their behavior appears biased (then leading us to consider what they know about us, and decide if such behavior implies a bias). This is a direction we hope to explore in future work.

Our finding that social bias detection can rely on Theory of Mind has several immediate implications. First, our results predict that this capacity might be, to a large extent, uniquely human. Although non-human primates share some Theory of Mind capacities (Krupenye, Kano, Hirata, Call, & Tomasello, 2016), the frequency with which they deploy these social inferences might be highly limited (Berke, Horschler, Jara-Ettinger, & Santos, 2023). If we are correct, then the experience of feeling that someone is unfairly biased towards you might be an experience that is not shared across the animal kingdom. Second, our results also hint at how the capacity to draw such inferences may develop in childhood. Previous work has suggested that 4 year-olds have the inferential machinery to reason about others' beliefs about the self (Asaba & Gweon, 2022) and that they themselves hold prior expectations about what others know (VanderBorght & Jaswal, 2009). Thus, young children's belief-reasoning abilities and their knowledge of social biases may enable them to detect biases relatively early in childhood.

At the same time, our work does not imply that detecting social biases inevitably relies on Theory of Mind inferences. First, it is possible that, after repeatedly inferring a social bias through Theory of Mind, people learn to associate certain behaviors with underlying bias, no longer requiring mental-state inference. For example, imagine that someone asks you where you come from, even after they have learned that you were born in the United States. If you receive this question repeatedly, then you might eventually become suspicious that these questions on their own suggest a prior belief that you do not belong in a particular space. So, this would be a case where our *Recommendation only model* would accurately capture people's inferences. Second, it is also possible that there are other routes to social bias inference that rely on other cognitive processes. People may infer biases by relying on covariation information——for instance, if your friend is on a dating website and tends to "swipe right" (i.e., show that you like) people of a certain background and "swipe left" (i.e., show that you don't like) people from another background, you might infer that they have a bias towards one group over another. Thus, our work only shows that people *can* infer social biases through Theory of Mind, but this does not imply that this is the only mechanism.

Our work prompts questions about extending and generalizing the model to capture how bias inferences work in the real world (see Figure 6 for a list of limitations). First, our experiments and model focused on situations where the space of possible prior beliefs is highly structured and constrained: observers were limited to one of three categorical biases (negative bias, neutral/no bias, positive bias). In the real world, however, 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"). Second, we frequently make inferences not just about the bias itself (e.g.: "this person has a bias that I'm bad at math"), but also about the specific identity marker(s) or stereotypes on which that bias is based (e.g.: "this person thinks I'm bad at math *because* I'm a woman"). In Experiments 3a-c, participants were explicitly told about a single demographic feature on which the observer's biases were based (race, gender, or age). However, such information is often not readily available and further, people's biases often stem from stereo-

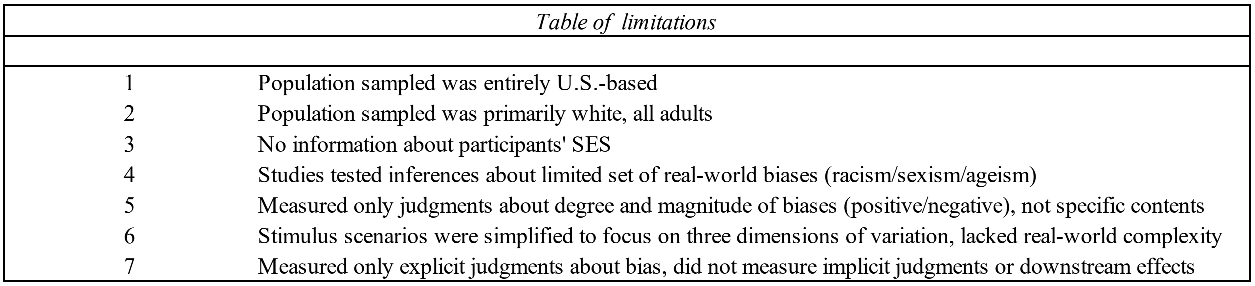


Table of limitations

1	Population sampled was entirely U.S.-based
2	Population sampled was primarily white, all adults
3	No information about participants' SES
4	Studies tested inferences about limited set of real-world biases (racism/sexism/ageism)
5	Measured only judgments about degree and magnitude of biases (positive/negative), not specific contents
6	Stimulus scenarios were simplified to focus on three dimensions of variation, lacked real-world complexity
7	Measured only explicit judgments about bias, did not measure implicit judgments or downstream effects

Figure 6: Limitations on the generalizability of our results.

types about multiple identities (e.g., being a Black woman). Thus, future work should extend our model to consider how people infer the magnitude and direction of others’ biases, as well as the source and stereotype(s) on which that bias in based.

Our work also prompts questions about inferences about others’ biases directed at the self, and how these affect one’s own behaviors. Running a first-person version of our experiments would enable us to test for individual differences in people’s expectations about social biases. Specifically, many people likely have their own priors about other people’s priors. For example, we might think that someone is more or less likely to hold, say, a certain racial bias depending on where that person is from, which may be informed by our own personal experiences of receiving bias. In some cases, this may make bias inference easier, allowing us to make more accurate inferences from even less data. Furthermore, people may have different reactions to being the victim of bias, specifically with regards to how they try to counteract others’ biases. 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. 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.

At the beginning of his basketball career, Jeremy Lin was granted very little, despite others’ observations of his clear achievements. Our work lends scientific credence to our (and many others') intuitions about how to explain this discrepancy: Officials and coaches should have treated him better given what they saw him do, but their actions imply pre-existing biases that shaped how they viewed him on the court. From blatant cases like Lin's mistreatment to everyday microagressions, humans can accurately pick up on social biases by reasoning about others' minds.

References

38 at the garden. (2022). New York: HBO.

Aboud, F. E. (2003). The formation of in-group favoritism and out-group prejudice in young children: Are they distinct attitudes? *Developmental psychology*, *39*(1), 48.

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.

Awad, E., Levine, S., Anderson, M., Anderson, S. L., Conitzer, V., Crockett, M., … others (2022). Computational ethics. *Trends in Cognitive Sciences*, *26*(5), 388–405.

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. PsyArXiv.

Berke, M., Horschler, D., Jara-Ettinger, J., & Santos, L. (2023). Differences between human and non-human primate theory of mind: Evidence from computational modeling. *bioRxiv*, 2023–08.

Boysen, G. A., Vogel, D. L., Cope, M. A., & Hubbard, A. (2009). Incidents of bias in college classrooms: Instructor and student perceptions. *Journal of Diversity in Higher Education*, *2*(4), 219.

Cao, J., Kleiman-Weiner, M., & Banaji, M. R. (2017). Statistically inaccurate and morally unfair judgements via base rate intrusion. *Nature Human Behaviour*, *1*(10), 738–742.

Cao, J., Kleiman-Weiner, M., & Banaji, M. R. (2019). People make the same bayesian judgment they criticize in others. *Psychological Science*, *30*(1), 20–31.

Coffman, K., Collis, M., & Kulkarni, L. (2019). *Stereotypes and belief updating*. Harvard Business School Cambridge, MA.

Cushman, F. (2023). Computational social psychology. *Annual Review of Psychology*, *75*.

Davis, I., Carlson, R. W., Jara-Ettinger, J., & Dunham, Y. (2023). Identifying social partners through indirect prosociality: a computational account. *Cognition*, *240*, 105580.

Fritzsche, B. A., DeRouin, R. E., & Salas, E. (2009). The effects of stereotype threat and pacing on older adults' learning outcomes. *Journal of Applied Social Psychology*, *39*(11), 2737–2755.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in cognitive sciences*, *7*(7), 287–292. doi: https://doi.org/10.1016/S1364-6613(03)00128-1

Gershman, S. J., Pouncy, H. T., & Gweon, H. (2017). Learning the structure of social influence. *Cognitive Science*, *41*, 545–575.

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

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: https://doi.org/10.1016/j.tics.2016.05.011

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

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110–114.

Lamont, R. A., Swift, H. J., & Abrams, D. (2015). A review and meta-analysis of age-based stereotype threat: negative stereotypes, not facts, do the damage. *Psychology and aging*, *30*(1), 180.

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.

Marcus, G., Gross, S., & Seefeldt, C. (1991). Black and white students' perceptions of teacher treatment. *The Journal of Educational Research*, *84*(6), 363–367.

Monahan, C., Macdonald, J., Lytle, A., Apriceno, M., & Levy, S. R. (2020). Covid-19 and ageism: How positive and negative responses impact older adults and society. *American Psychologist*, *75*(7), 887.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences*, *109*(41), 16474–16479.

Muzzatti, B., & Agnoli, F. (2007). Gender and mathematics: attitudes and stereotype threat susceptibility in italian children. *Developmental psychology*, *43*(3), 747.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, research, and practice*, *6*(1), 101.

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature human behaviour*, *2*(10), 750–756.

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

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of african americans. *Journal of personality and social psychology*, *69*(5), 797.

Stone, J., Lynch, C. I., Sjomeling, M., & Darley, J. M. (1999). Stereotype threat effects on black and white athletic performance. *Journal of personality and social psychology*, *77*(6), 1213.

VanderBorght, M., & Jaswal, V. K. (2009). Who knows best? preschoolers sometimes prefer child informants over adult informants. *Infant and Child Development: An International Journal* *of Research and Practice*, *18*(1), 61–71.

Wayman, J. C. (2002). Student perceptions of teacher ethnic bias: Implications for teacher preparation and staff development.

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