Race effects on impression formation in social interaction: An instrumental learning account

Iris J. Traast, David T. Schultner, Bertjan Doosje, and David M. Amodio
University of Amsterdam

Word count: 10116 words

Author note

Iris J. Traast (ORCID: 0000-0002-2132-4632), David T. Schultner (ORCID: 0000-0003-2253-4065), Bertjan Doosje (ORCID: 0000-0002-2479-5405), and David M. Amodio (ORCID: 0000-0001-7746-0150).

Preregistration of study design, hypotheses and analyses can be found at https://aspredicted.org/LGA_BHZ (Study 1) and https://aspredicted.org/31S_ZH6 (Study 2). Supplemental information, data, and analysis scripts can be found at https://osf.io/rnjgh.

This research was funded by the Netherlands Organisation for Scientific Research (VICI 016.185.058), awarded to DMA. We thank Björn Lindström and members of the Amsterdam-NYU Social Neuroscience Lab for their assistance with this research and feedback on earlier versions of this manuscript.

Correspondence regarding this article should be addressed to Iris J. Traast or David M. Amodio, Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129, REC G, 1001 NK Amsterdam, NL. Email: i.j.traast@uva.nl or d.m.amodio@uva.nl.

Abstract

How does race influence the impressions we form through direct interaction? In two preregistered experiments (*N*s=239/179), White American participants played a money sharing game with Black and White players, based on a probabilistic reward reinforcement learning task, in which they chose to interact with players and received feedback on whether a player shared. We found that participants formed stronger reward preferences for White relative to Black players despite equivalent reward feedback between groups—a pattern that was stronger among participants with low internal motivation to respond without prejudice and high explicit prejudice. This race effect in reward learning was evident in participants' behavioral choice preferences, but not in their self-reported perceptions of group members' reward rates. Computational modeling suggested two mechanisms through which race affected instrumental learning: race (a) influenced White participants' initial expectancies (i.e., priors) about Black compared with White players' behavior and (b) led participants to update reward representations of Black and White players according to separate learning rates. These findings demonstrate that race can influence the formation of impressions through direct social interaction and introduce an instrumental learning framework to understand the effects of bias in intergroup interactions.

Keywords: prejudice, reinforcement, learning, social, computational

Public Significance Statement

Race influences how people form impressions of others in direction social interactions through a process of instrumental learning, particularly among perceivers high in prejudice. Computational modeling reveals that this bias occurs because race influences perceivers' initial expectancies and how they learn from their partner's behavior.

In social exchanges, we can form an impression of our interaction partner based on how they act: by observing how they respond to our actions, we learn to prefer people who treat us well over those who treat us poorly (Brewer, 1988; Hackel et al., 2015; Neuberg, 1989). Yet interaction partners may belong to a variety of social categories, such as race, and these categories can influence and bias our impressions (Chen, 2019; Macrae & Bodenhausen, 2000; Maddox, 2004; Shelton & Richeson, 2006). Given the pervasive effects of race on social perception and judgment (Devine, 1989; Eberhardt et al., 2006; Kawakami et al., 2017), we asked whether an interaction partner's race can bias how social preferences are formed through direct social-interactive learning. By examining this question using a social reinforcement learning paradigm in conjunction with computational modeling, we further sought to understand the mechanisms through which race influences initial perceptions of a person's reward value and how this impression develops with repeated interactions.

Forming impressions through social interactive learning

People form impressions about others in multiple ways (Amodio, 2019; Uleman & Kressel, 2013): for example, we can learn about a person by observing their behavior, hearing gossip about them, drawing inferences based on their social networks, and, perhaps most importantly, by engaging with them in direct social interaction. In social interactions, we form an impression of a person based on how they respond to us (Hackel et al., 2015). We learn to value those who respond positively (e.g., in a cooperative, fair, helpful, kind, or generous manner) over those who respond negatively and, by updating these impressions over repeated interactions, we learn whom to approach or avoid in the future. This kind of learning—through action and feedback—characterizes the process of instrumental learning through reward reinforcement (Hackel & Amodio, 2018; Sutton & Barto, 1998).

Although relatively little research has examined instrumental learning in impression formation (e.g., Hackel et al., 2015, 2020), this approach offers a theoretical framework for how impressions are formed and expressed through interactive behavior. Following reinforcement learning theory (Sutton & Barto, 1998), the reward value of a choice is updated when reward feedback deviates from the expected reward. The degree of this deviation is referred to as the *prediction error*. A prediction error is positive when reward feedback exceeds expectations or negative when reward feedback is less than expected. The updating of a reward value depends on the size of a prediction error as a function of one's *learning rate,* which represents the degree to which an expected value is changed in response to new feedback.

Unlike previously-studied impression formation processes, which generally involve the semantic learning of concepts and evaluations, instrumental learning encodes the reward value of a behavior which is acquired through action and feedback and updated across interactions (Amodio, 2019; Cone et al., 2017). These characteristics of semantic and instrumental learning reflect their respective neural substrates: whereas semantic learning of social concepts and evaluations is generally supported by the anterior temporal lobes (Gilbert et al., 2012; Olson et al., 2013; Wang et al., 2017), instrumental learning is supported by the striatum and its interplay with the ventromedial prefrontal cortex and motor cortex (Averbeck & O'Doherty, 2022; Hackel et al., 2015). As such, instrumental learning is expressed in approach or avoid behavioral choices, in contrast to trait judgments and attitudes examined in most prior studies of impression formation (Amodio, 2019; Hackel et al., 2020). In line with these distinctions, research has shown that striatum-based instrumental learning may be formed and expressed implicitly (i.e., without awareness), in comparison with semantic or episodic learning supported by the temporal lobes (Foerde & Shohamy, 2011; Knowlton et al., 1996). Thus, according to an instrumental

learning account, social-interactive learning is rooted in reward reinforcement, is updated incrementally in response to feedback, should be more directly evident in behavior than in self-reports, and may form and be expressed implicitly.

Race effects on social interactive learning

Can the race of an interaction partner influence how one forms an impression via social interaction? Race is known to profoundly affect how we perceive, judge, and interact with people (Allport, 1954; Fiske, 1998; Kawakami et al., 2017): racial stereotypes shape people's assumptions about a group member's characteristics and expectations for how they will act (Darley & Gross, 1983; Kunda & Sherman-Williams, 1993), and prejudice drives people's willingness to accept and act on such biases (Devine, 1989). Indeed, White American stereotypes portray Black Americans as less friendly and competent (Fiske et al., 2002; Hass et al., 1991), and greater prejudice among White people predicts avoidance tendencies in interracial interactions (Amodio & Devine, 2006; Dovidio et al., 1997; 2002; Fazio et al., 1995; McConnell & Leibold, 2001). Furthermore, in interracial interactions, race can influence the interpretation of a partner's behavior (Shelton & Richeson, 2006), such as when White Americans judge a Black person's performance as inferior to a White person's even when their actual performance is equated (Biernat et al., 2010; Gaertner & Dovidio, 2000).

These findings suggest that in the context of instrumental learning, race may influence two components of the learning process. First, it may bias one's behavioral expectations at the beginning of an interaction, such that the expected value of a racial outgroup member is initially set lower than the expected value of an ingroup member. In computational models of reinforcement learning, this initial expectancy can be modeled as a *prior*. Second, race may affect the degree to which a reward value is updated in response to an interaction partner's

feedback (i.e., when there is a prediction error). That is, the same reward feedback may be experienced differently when it comes from a racial ingroup member than outgroup member, leading the perceiver to update their impressions of ingroup and outgroup members according to different updating rules. In computational models, these updating rules can be modeled as *learning rates*.

Evidence for separate learning rates for two different groups would indicate that a learner maintains separate representations of each group's value, even if the learning rate value does not differ in magnitude. A difference in learning rate magnitude would indicate that representations are updated more readily for members of one group compared to another. In the context of existing biases (i.e., priors), separate learning rates of similar value would function to maintain an existing prejudice (e.g., pro-White and anti-Black bias), whereas a difference in learning rate would suggest that a pre-existing prejudice toward one group is revised more easily than prejudice toward the other group. Regardless of whether learning rates differ in magnitude, the existence of separate learning rates for racial ingroup and outgroup members would indicate that race influences impression updating.

Whereas much previous research has examined the effects of expectancies in intergroup bias (e.g., Hamilton et al., 1990; Shelton & Richeson, 2006), this proposed effect on learning suggests a new biasing mechanism that emerges in the context of direct interactions. Together, these two biasing effects of race—priors and learning rates—could result in a learner forming more negative impressions of a racial outgroup member than a racial ingroup member, even when members of each group respond in equally-rewarding ways.

Although the effect of race on instrumental learning has not been previously reported, research by Schultner et al. (2022) found that positive and negative stereotypes of novel groups

influenced instrumental learning from members of each group. Despite equated reward feedback between groups, participants formed more positive reward associations with members of the positively stereotyped group than with the negatively stereotyped group. This stereotype effect on instrumental learning emerged in behavior but was not evident in participants' self-reported perceptions of group members' reward rates. Further, computational modeling suggested this effect involved a combination of stereotype-based priors and separate learning rates for each group. It is unclear, however, whether a similar pattern would emerge in the context of race—a real, if constructed (Cikara et al., 2022), social category that may elicit prejudiced or egalitarian responses.

Individual differences in race effects on learning

Despite the pervasiveness of racial discrimination in social structures and by individuals, many people reject racial prejudice and strive to respond without prejudice (Devine, 1989; Devine & Monteith, 1993). Research on *internal motivation to respond without prejudice*—the desire to respond without prejudice for personal reasons, as opposed to normative reasons (Plant & Devine, 1998)—shows that White people high in internal motivation respond more carefully to racial cues, engage more respectfully with Black interaction partners, and respond without prejudice more consistently across situations relative to people low in internal motivation (Krosch et al., 2017; LaCosse & Plant, 2020; Plant & Devine, 2009).

In a social-interactive learning context, high internally-motivated individuals may strive to learn accurately from a racial outgroup member's feedback. Indeed, highly internally-motivated individuals are more likely to approach interracial interactions, to control their racial biases, to rely less on monoracial category perceptions, and to be more receptive to outgroup-positive associations (Amodio et al., 2008; Chen et al., 2014; Li et al., 2016; Plant et al., 2010).

As a result, learners with strong internal motivation may be less susceptible to race effects on instrumental social learning compared to learners who are weakly motivated. Hence, in social interactions where race is salient, internal motivation may be an important factor in race effects on learning. External motivation, by comparison, is engaged when responses are public and one's expressions of prejudice would be met with disapproval. Although the present research examined social-interactive learning in a private context, and thus EMS was not a focus, it is possible that high external motivation to respond without prejudice might also lead to individuated learning in public contexts.

Research Overview

In two experiments, we examined the effect of race on impressions formed through social instrumental learning. These studies were conducted in the context of the United States, in which the majority racial group comprises White people (with European heritage) and in which there is a long history of White's discrimination toward Black people (i.e., those with African or Caribbean heritage). Therefore, these studies included self-identified White, non-Hispanic American participants who interacted with Black and White American partners.

In both studies, participants completed a social reinforcement learning task (adapted from Schultner et al., 2022; also Hackel et al., 2015) with partners who presented as White or Black. We hypothesized that White participants would form more negative instrumental reward associations with Black partners compared with White partners, and that this effect would be larger among participants low in internal motivation to respond without prejudice.

We further hypothesized that race would influence the formation of instrumental preferences through a combination of two processes: (1) race would bias initial expectancies (i.e., a *group-based prior*), such that participants would begin the task with a tendency to choose

White over Black players, and (2) participants would update their impressions of Black and White players at different rates (*group-based learning rates*). These processes, in combination, would result in higher learned reward values for White over Black players. We tested these mechanistic hypotheses using computational modeling, which quantified the fit of this formalized model to participants' task behavior and compared it with several alternatives (described in Results and SI).

Study 1

Method

Participants

Participants were 305 self-identified White Americans born in the USA, recruited via Amazon Mechanical Turk (Mturk), who were compensated with $2.00 plus a performance-based bonus (ranging from $1.00 to $2.00). Participants indicated their race/ethnicity on the question "Please select all categories that apply to you" and chose one or more of the listed categories (*White*; *Hispanic, Latino or Spanish origin*; *Black or African American*; *Asian*; *American Indian or Alaska Native*; *Middle Eastern or North African*; *Native Hawaiian or other Pacific Islander*; and *some other race, ethnicity, or origin*). Participants indicated their gender as female, male, or other, or they could choose not to respond. Following exclusions for below-chance learning (under 50% choice accuracy on trials comparing the highest- vs. lowest-reward players for Black *and* White players during test; 24 participants) or extremely fast reaction times (median RT<300 ms in training or test phase; 41 participants), the final sample for analysis included 239 participants (112 female-identified, 116 male-identified, and 11 did not indicate gender; M_{age} =39.91, SD_{age}=11.60). The preregistered stopping goal for this initial test was N=200 with the aim of obtaining valid data from at least 160 participants, based on research using a similar task

(Schultner et al., 2022). However, this goal was exceeded when recruiting to replace exclusions and to test exploratory hypotheses involving higher-order interactions (see SI); we report analyses using the full sample here and, in the SI, report additional analyses based on the preregistered sample size, which yielded the same hypothesized results.

Procedure

The study was conducted and hosted on a computer via the open-source framework psiTurk (v3.3.0; Gureckis et al., 2016; Eargle et al., 2020). Online data collection occurred during October 2019. After providing informed consent and reading the instructions, participants completed the main learning task, followed by a series of post-task questionnaires assessing their perceptions of player reward rates and the internal and external motivation to respond without prejudice scales (IMS/EMS).

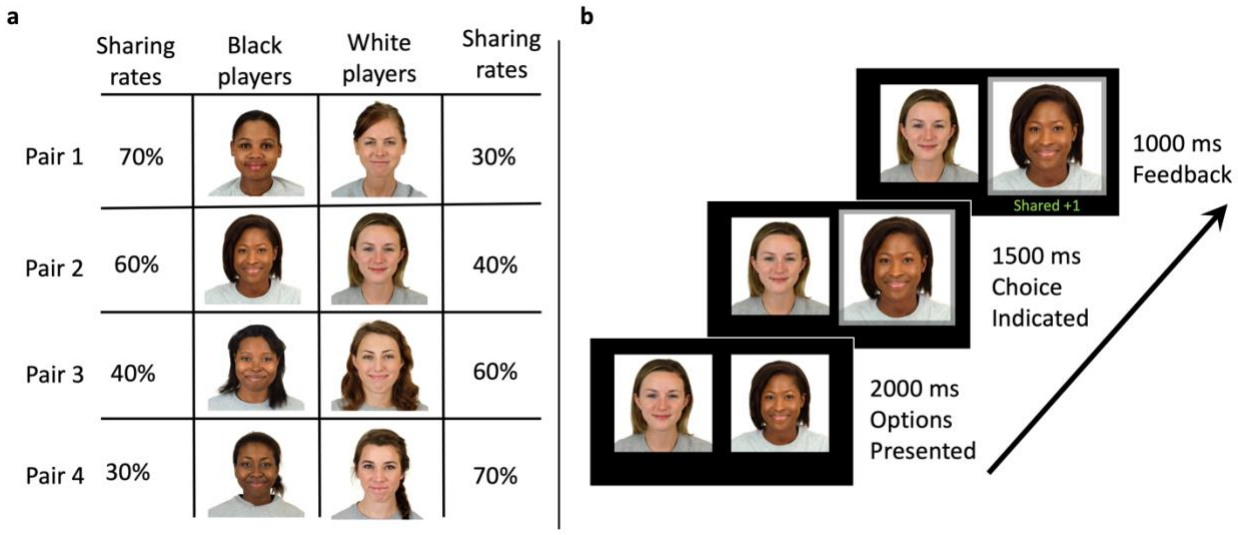
Social reinforcement learning task. Participants completed an interactive money sharing task based on a widely-used probabilistic reward reinforcement paradigm (Frank et al., 2004), adapted for social-interactive learning (Schultner et al., 2022; Hackel et al., 2015, 2022). Participants were told that they would play a point-sharing game with eight players. The participants' explicit goal was to learn to choose players most likely to share points, in order to earn as many points as possible which would be exchanged for cash at the end of the study. Players were described as participants from a previous version of the study, whose sharing responses on each trial were recorded. In reality, players were fictitious and responded with fixed reward rates (Figure 1a). Of the eight players, four were White and four were Black in appearance; players were either all men or all women, with player gender counterbalanced across participants. Faces representing players were selected from the Chicago Face Database based on norming data, such that they were distinguishable by race but did not differ in attractiveness or

trustworthiness (Ma et al., 2015; see SI), and all were smiling in line with the cover story that they were past participants who posed for a picture.

The task included two phases: a *training phase*, in which participants chose between players and learned from the reward feedback they received, and a *test phase*, in which choices were made without feedback so that learning could be assessed. The training phase included two blocks of 80 trials. On each trial, participants viewed and chose between one Black player and one White player, and then received immediate feedback on whether the chosen player shared 1 or 0 points (Figure 1b). If no response was given within 2.5 s, the trial ended without feedback, followed by a "too slow" message, and then proceeded to the next trial. During training, participants viewed four fixed pairs of faces that varied in the reward rates between the Black and White player (70/30, 60/40, 40/60, and 30/70, Figure 1a). The presentation order of pairs and their feedback was randomized within participants and assignment of face to reward rate was randomized across participants.

Figure 1

Trial order and player reward rates



Note. Panel a displays reward rates for player pairs during the training phase. Player images were randomized, and gender was counterbalanced. Panel b shows a sample trial sequence of the training phase. Participants viewed two player faces, chose one to interact with (player on the right in the current trial), and then received feedback ('Shared: +1' or 'Shared: 0').

The test phase was designed to provide a readout of learned reward values that generalized beyond the context of the fixed pairs encountered during training. Thus, during the test phase, participants viewed and chose between every possible pairing of a Black and White player, without feedback. Participants were informed they would continue to earn money by choosing high-sharing players, which would be added to their end-of-study bonus. These choices provided an index of learned instrumental reward associations for each player. Reward learning was indicated by the degree at which players with high reward rates (i.e., sharing), relative to the other player in the pair during training, were chosen during the test phase. A race effect was indicated by the degree of preference for players from one race over another, collapsing across individual player reward rates (which were equated between race groups).

Perceived reward rates. Following task completion, participants reported their estimate of each player's reward rate. Participants viewed each player's face one at a time, in randomized order, and were asked "What percent of the time did this player share with you?" Perceived reward rate was indicated by typing in a number between 0 to 100%.

Internal and external motivation to respond without prejudice scales (IMS and EMS). Internal and external motivation to respond without prejudice were measured using the IMS and EMS, which each comprised five items (Plant & Devine, 1998). A sample IMS item is "I am personally motivated by my beliefs to be unprejudiced toward Black people," and a sample EMS item is "I try to hide any negative thoughts about Black people in order to avoid negative reactions from others." Participants rated their agreement with each item on a scale of 1 (strongly disagree) to 9 (strongly agree). Because our hypotheses did not concern EMS, results involving EMS are reported in the SI.

Transparency and openness

In this paper, we have reported how our sample size was determined, all data exclusions and all manipulations and measures in the study. All data, analysis code and research materials for Study 1 are available at osf.io/rnjgh. Data were analyzed using R Statistical Software (v3.6.1; R Core Team 2019). The study design, hypotheses, and analyses were preregistered at https://aspredicted.org/LGA_BHZ. This study complies with TOP level 2 guidelines.

Results

Descriptive analyses

The IMS (α=.88) and EMS (α=.87) scales were reliable and produced distributions similar to those of previous studies (e.g., Plant & Devine, 1998). Mean IMS score was 7.23 (*SD*=1.80, range:1.40–9.00) and mean EMS score was 4.70 (*SD*=2.27, range:1.00–9.00).

Race effects on instrumental learning

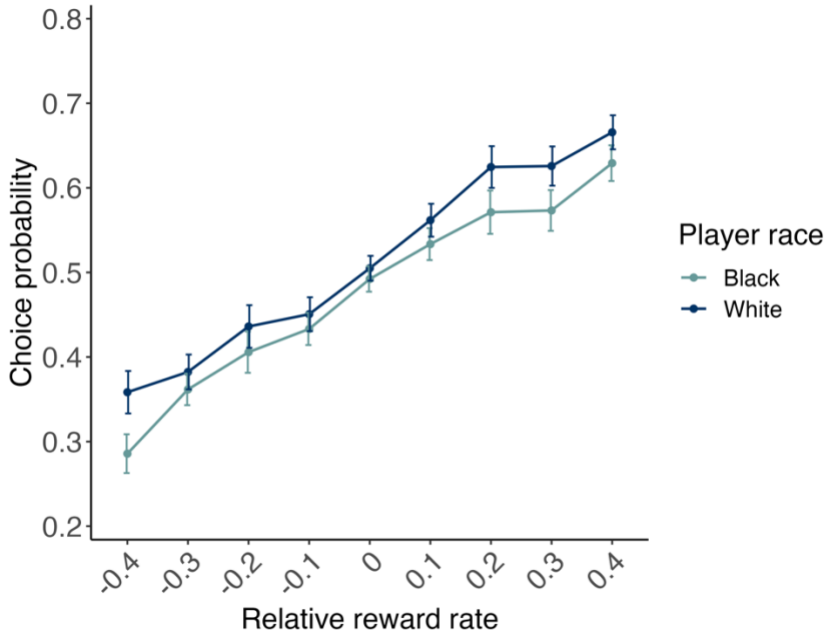
We hypothesized that social instrumental reward learning would be influenced by the race of an interaction partner, such that White Americans would learn more positive reward values for White than Black partners, on average, and that this effect would be moderated by internal motivation. Hence, we predicted that participants would show a choice preference for White over Black players, in addition to a preference for players with higher reward rates. We further predicted a stronger effect of race among participants with lower IMS scores.

We tested these main predictions with a logistic regression using the R lme4 package (v1.1-26; Bates et al., 2015). Our main analysis consisted of a generalized linear mixed effects model, with relative reward rate, race, IMS, and the Race x IMS interaction as predictors and choice behavior as the outcome. Variables with repeated measures (relative reward rate and race) were clustered within participants using random slopes and intercepts.

This analysis produced a significant effect of relative reward rate (OR=9.08, 95%CI=[5.89,13.99], p<.001), demonstrating that participants successfully learned from feedback to choose higher-sharing players over lower-sharing players (Figure 2). Critically, this analysis also produced the predicted effect of race (OR=6.43, 95%CI=[2.55,16.19], p<.001), such that White players were over six times more likely to be chosen than Black players. The effect of race was not moderated by reward rate (Race x Reward interaction: OR=1.13, 95%CI=[0.84,1.51], p=.429).

Figure 2

Effects of race and reward on choice

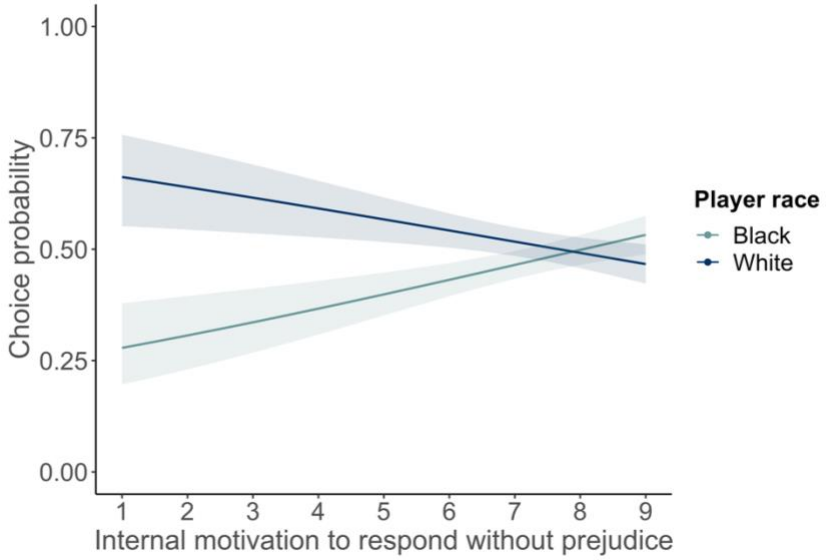


Note. Effects of relative reward rate and race on choice during test phase in Study 1, showing a preference for higher-rewarding players, and White players over Black players across relative reward rates. Relative reward rate (difference between training-phase reward rates of a choice pair) is displayed on the x-axis, and choice probability displayed on the y-axis. Error bars represent standard errors.

We additionally hypothesized that the effect of race would be moderated by IMS. Indeed, the effect of race was qualified by a Race x IMS interaction (OR=0.79, 95%CI=[0.70,0.89], p<.001), such that participants with lower IMS scores showed higher odds of choosing White players compared with participants with higher IMS scores (Figure 3; see SI for full regression output). To interpret this interaction, simple slope analyses were conducted at high (M+1SD) and low (M-1SD) levels of IMS. These analyses confirmed that the effect of race was significant among low IMS participants ($M_{low\text{-}IMS}$=5.47, β=0.57, t=3.59, p<.001), but not among high-IMS participants ($M_{high\text{-}IMS}$=9.09, β=-0.28, t=-1.75, p=.079). Together, these results supported our main hypotheses that race influences social instrumental reward learning and that this effect is strongest among low IMS individuals.

Figure 3

Race x IMS interaction effect on choice



Note. Race x IMS interaction effect on choice in Study 1, showing a stronger effect of race on choice preference among participants with lower internal motivation. Shaded areas represent the 95% confidence interval.

Finally, to provide a preliminary test of whether participants entered the task with pre-existing racial preferences (i.e., priors), we tested whether a preference for White over Black players was already evident in the first 50 trials of training. Indeed, this analysis revealed an initial preference for White over Black players (race effect: OR=2.78, 95%CI=[1.44,5.34], p=.002). Moreover, this initial preference differed as a function of internal motivation (Race x IMS interaction: OR=0.88, 95%CI=[0.81,0.96], p=.005), such that low IMS participants showed a pre-existing pro-White bias (β=0.33, t=2.92, p=.004), but high-IMS participants did not (β=-0.13, t=-1.16, p=.247). These results provide preliminary behavioral evidence for an effect of race on expectancies—a pattern we examine further with computational modeling.

Race effects on perceived reward rates

Did participants explicitly perceive a difference in reward rates between Black and White players, as suggested by their choice behavior? To test this question, self-reported perceived reward rates were submitted to a multilevel linear regression with race and actual player reward rate as predictors. Participants' reported reward rates correspond significantly with players' actual reward rates, β=0.54, 95%CI=[0.46,0.61], p<.001, suggesting participants were aware of individual variation in reward feedback. However, participants perceptions of reward rate did not differ significantly by race (Black players: M=43.26, SD=18.69; White players: M=41.41, SD=18.80), β=1.83, 95%CI=[-0.28,3.93], p=.089, and the Race x IMS interaction did not reach significance, β=1.03, 95%CI=[-0.15,2.21], p=.086.

Next, we tested whether the effect of race on participants' choice behavior was independent of their explicit perceptions of reward rate. To this end, we repeated the regression predicting choice behavior described above while statistically adjusting for race differences in perceived reward rate (i.e., by including the Race x Perceived Reward interaction as covariate). Although perceived reward rate was significantly associated with choice preference (OR=1.06, 95%CI=[1.05,1.08], p<.001), such that those who perceived higher reward rates from White players were also more likely to choose White players, the effects of race (OR=4.01, 95%CI=[1.92,8.38], p<.001) and the Race x IMS interaction (OR=0.84, 95%CI=[0.76,0.93], p=.001) remained significant predictors of choice. These results suggest that the race effect on instrumental learning was independent of participants' explicit perceptions of rewards.

Computational modeling of learning mechanisms

The regression analyses reported above suggest that race influenced the instrumental preferences of low internal motivation participants. However, these regression analyses could not discern whether and how race influenced the formation of preferences. To test our mechanistic

hypothesis—that race influences the learning of reward-based associations, in addition to expectancies—we used a computational modeling approach, in which we examined the fit of trial-by-trial behavioral data to a model that specifies an effect of race on both learning (i.e., separate learning rates for White and Black players) and on initial expectancy (i.e., a prior). According to this model, participants hold different initial reward representations for White and Black players, which are then updated via separate updating rules in response to prediction errors. This model does not specify a difference in the magnitude of learning rates—that is, it tests whether reward representations of Black and White players are maintained and updated independently, but not whether the change in reward value to a prediction error is greater for one group than the other. To the extent this combined model, which includes a race prior and separate learning rates, provides a better fit to behavioral data than alternative models, then our hypothesized mechanism would be supported.

To test this hypothesis, we fit participants' behavioral choice data to the combined computational model which specified opposing expectancies (modeled as a prior):

$$Q_{white}^{t=0} = P, \text{ and } Q_{black}^{t=0} = -P$$

Where *P* denotes a prior for either White or Black players, such that participants show a race-based preference before interactions occur.

Reward representations were updated using the Rescorla-Wagner learning rule:

$$Q_{i,race}^{t+1} = Q_{i,race}^{t} + a_{race}(R^t - Q_{i,race}^{t})$$

Where $Q_{i,race}$ is the action value of selecting player i with a specific race in trial t, R is the reward received in trial t, and a_{race} denotes the learning rate parameter, which determines the extent to which the prediction error affects subsequent reward estimates, and which differs by

target race, such that prediction errors experienced when choosing Black players may be processed differently than prediction errors experienced when choosing White players.

We then compared our hypothesized *combined model* with three main plausible alternative models (see SI for additional alternatives):

(a) An *unbiased model*, which contained no prior and one learning rate for both racial groups; in this basic reinforcement learning model, race has no effect on expectancies or learning.

(b) A *group-based prior model*, containing a prior for White vs. Black players; in this model, participants begin with different expectancies for Black and White players. Although these expectancies anchor the updating of reward values for players, updating rules do not differ by race. This model corresponds to classic stereotyping models in which stereotypes guide expectancies but are replaced by individuated learning.

(c) A *group-based learning model*, containing no prior but separate learning rates for White and Black players; in this model, participants do not begin with different expectations for Black and White players, but they form and update separate representations of Black and White players' reward value.

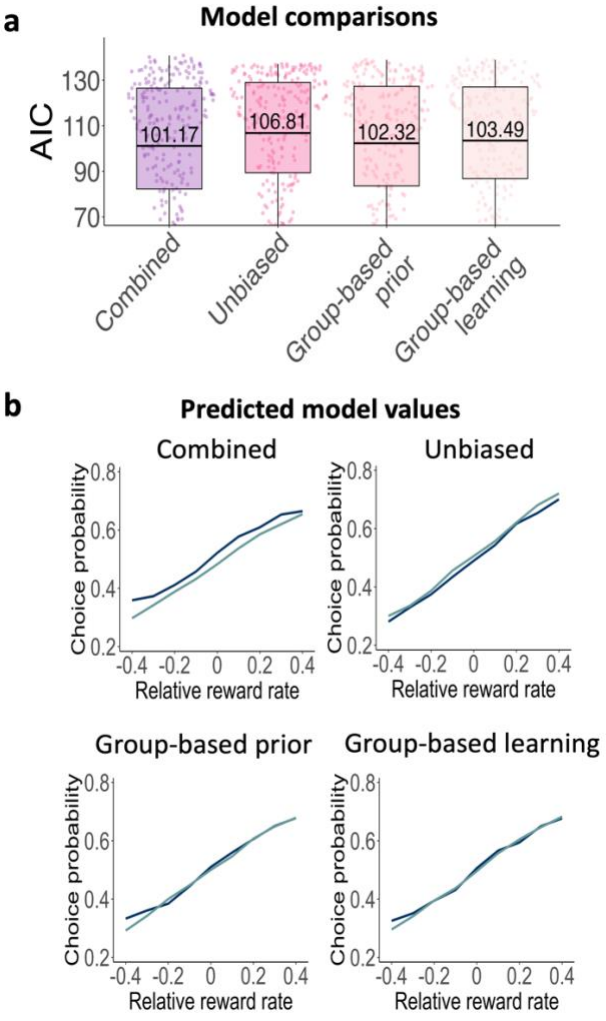
In addition to these main alternative models, we considered a *prior+valence group-based learning* model, which included a prior as well as four separate learning rates for positive vs. negative feedback from Black and White players, and an *ingroup reward model*, in which ingroup choices were always associated with a bonus, simulating a reward experience for ingroup favoritism (see these models and other alternatives in SI).

These models were tested and compared with our hypothesized model based on their Akaike Information Criteria (AIC): the lowest AIC indicates best fit to the data, as it explains the

greatest amount of variation with the fewest possible parameters (complete model specifications are provided in the SI). Model comparison indicated that the hypothesized combined model provided the best fit to the data compared with all other models (Figure 4a). Although the differences in AIC were relatively small between the combined model (AIC=101.17) and the competing group-based prior (AIC=102.32; ΔAIC=1.15) and group-based learning (AIC=103.49; ΔAIC=2.32) models, these competing models nonetheless provided worse fits. Because the AIC penalizes additional parameters, the improved fit for that combined model was unlikely to be a result of overfitting. Thus, according to the best-fitting model, participants had divergent initial expectancies for Black and White players (*prior*) and updated their impressions at separate rates depending on partner race (*group-based learning rates*).

Figure 4

Computational model comparisons and simulated data



Note. **a,** Model comparisons between the hypothesized combined model with the unbiased model, group-based prior model, and group-based learning model in Study 1. **b,** Predicted values simulated by each model in Study 1.

To further assess model fit, we examined patterns of data simulated by these competing models. The combined model produced simulated data that closely matched participants choice behavior (Figure 4b), whereas simulated data from the alternative models failed to produce the group effect observed in behavior.

Finally, given that the effect of race on choice preference was most evident among participants with lower internal motivation, we tested whether the fit advantage for the hypothesized combined model was better for these participants. As expected, the correlation between IMS and AIC model fit was significant, $t(237)=2.22$, $p=.027$, such that low IMS participants showed better model fit, and thus lower AIC, compared with high IMS participants.

It is notable that we refrained from interpreting participant-level parameter estimates representing the prior or group-specific learning rates. This is because each parameter is dependent on others in the model and thus cannot be interpreted independently. Furthermore, in part because of this nonindependence, participant-level parameter estimates derived using the present modeling approach may be highly variable (Wiecki et al., 2013; Piray et al., 2019) and unstable (Schaaf et al., 2023), and thus difficult to interpret (Eckstein et al., 2021). We report these estimates in the SI for transparency but caution against their interpretation.

Discussion

Study 1 provided initial support for the hypothesis that race influences the process of impression formation in the context of direct social interaction through its effect on instrumental learning. We found that participants formed more positive associations with White than Black players, despite their equivalent reward rates. Moreover, this effect was moderated by IMS, such that it was stronger for participants with low levels of internal motivation. Participants with relatively high internal motivation were not significantly influenced by race when learning player's reward rates. Because these choices were incentivized with real cash bonuses and, given participants' explicit goal to learn about the individual players' tendencies to share and choose

accordingly, the effect of race on choice behavior was interpreted as reflecting participants' ability to accurately learn group members' reward associations.

Despite the effect of race on choice behavior (for low IMS participants), participants' subjective perceptions of player's reward rates did not vary by race. Furthermore, a covariate analysis (reported above), which adjusted for perceived reward values when predicting behavioral choice preferences, showed that explicit perceptions did not account for the effect of race on choice behavior. These results are consistent with the possibility that group-based preferences in social instrumental learning may be formed and expressed in behavior without explicit awareness.

Finally, computational modeling results suggested that the effect of race on instrumental learning could be explained by two complementary processes. According to the best-fitting "combined" model, these White American participants began the task with a different reward-expectancy for White compared with Black players, and then updated reward representations of Black and White players according to separate learning rates. The effect of race on priors was consistent with behavioral evidence for a pro-White preference at the beginning of the task. The effect of race on learning rates suggests that participants maintained separate reward representations for Black and White players and updated these representations according to different rules (i.e., adjusting reward associations to different degrees based on prediction errors). This overall pattern was stronger for participants with low internal motivation to respond without prejudice, such that relative model fit was greater among participants with lower IMS scores.

A limitation of our approach is that it could not reliably determine whether learning rates for White and Black players differed in magnitude—a prediction that might follow from research showing greater individuation in judgments of ingroup members (Kawakami et al., 2014;

Meissner & Brigham, 2001; Vingilles-Jaremko, 2020). Nevertheless, model comparison results favored a model with separate learning rates for Black and White players over a model that did not distinguish learning by race, indicating that, regardless of potential differences in learning rate magnitude, race affected how participants learned from feedback. Together, these results provided preliminary support for the hypothesis that race affects instrumental learning in direct interactions and that it does so through mechanisms of biased expectancies and updating.

Study 2

Study 2 was conducted with two major aims: First, we sought to replicate the results of Study 1, particularly given its deviation from its preregistered sample size. Our second aim was to probe the moderators of the race effect on social instrumental learning, beyond IMS, as these may inform potential interventions to mitigate race bias in social-interactive learning.

Internal motivation to respond without prejudice represents one's personal egalitarian goals and intentions (Plant & Devine, 1998), and it has been associated with a variety of egalitarian responses, such as the control of stereotypes, multicategory perception, and engagement in prejudice-reducing activities, in addition to low-prejudiced attitudes (Amodio et al., 2008; Chen et al., 2014; Plant & Devine, 2009). Nevertheless, it is possible that the effect of IMS observed in Study 1 merely reflected an effect of prejudice, or that IMS and prejudice operated in concert to influence interracial effects on social instrumental learning. To this end, participants in Study 2 completed post-task measures of implicit and explicit prejudice in addition to IMS.

Furthermore, to shed light on the lack of race effects in participants' self-reported perceptions of player reward rate, we included a self-report measure of player liking. If the lack of a race difference in perceived reward rate in Study 1 was due to participants' effort to conceal

their prejudice, then we would also expect no race effect in ratings of player liking. However, if participants report race differences in liking but not in reward perceptions, then we could more strongly infer that they truly perceived no difference in the reward rates of Black and White players.

Method

Participants

Participants were 197 self-identified White Americans born in the US, recruited via CloudResearch (formerly TurkPrime; see Litman et al., 2017), who were compensated with $7.00 as reward plus a performance-based bonus (ranging from $1.00 to $2.00). Participants were prescreened such that they were all currently living in the US, born in the US, had English as their first language or learned it before the age of seven, and identified themselves as only non-Hispanic White. Despite this prescreen procedure, participants were asked to indicate their race/ethnicity on the question "Please select all categories that apply to you" and chose one or more of the listed categories (*White*; *Hispanic, Latino or Spanish origin*; *Black or African American*; *Asian*; *American Indian or Alaska Native*; *Middle Eastern or North African*; *Native Hawaiian or other Pacific Islander*; and *some other race, ethnicity or origin*; see SI). Participants indicated their gender as female, male, non-binary, or not included, or they could choose not to respond. Following the preregistration, data collection stopped once we tested 200 participants, with the goal of obtaining valid data from at least 160. After exclusions for below-chance learning on 70-30 pairs during test (13 participants), extremely fast median reaction times (8 participants) or over 80% missed trials (0 participants), the final sample for analysis consisted of 179 participants (61 female-identified, 74 male-identified, and 40 did not report gender; M_{age} =43.321, SD_{age}=11.272).

Procedure

The study was conducted and hosted on a computer via psiTurk (v3.3.0; Gureckis et al., 2016; Eargle et al., 2020). Online data collection occurred in April 2022. After providing consent and reading the instructions, participants completed the same social learning task as in Study 1, followed by an Implicit Association Task (IAT) and post-task questionnaires including: assessments of perceived rewards for each player participants interacted with, ratings of liking toward each player, a race-based Feeling Thermometer, and the IMS/EMS, followed by a set of items exploring factors that may have influenced participants' choices which were not analyzed and thus not discussed here (see SI). The perceived reward questions and IMS/EMS measure was also the same as in Study 1. The order of questionnaires was designed to minimize awareness of our hypotheses, with more reflective questionnaires positioned later in the sequence.

Tasks and measures

Social reinforcement learning task. The social reinforcement learning task was the same as in Study 1.

Implicit Association Test (IAT). Participants completed a standard seven-block evaluative IAT (based on Greenwald et al., 2003), in which they classified positive and negative words as "good" or "bad" and Black and White face images as "Black" or "White." Block order (compatible-first vs. incompatible-first) was counterbalanced. Using natural log transformed reaction times for correct responses, *D* scores were computed for each participant as in Amodio & Devine (2006): compatible block RTs were subtracted from incompatible block RTs and divided by the pooled *SD* separately for practice and test blocks, and these were averaged for the final *D* score.

Perceived reward for each player. As in Study 1, participants were asked "What percent of the time did this player share with you?" and answered on a scale from 0 to 100%.

Player liking ratings. Participants indicated how much they liked each player on a scale of 0 to 100. Participants viewed the face of each player, in random order, and responded to the item, "On a scale of 0 - 100, where 0 means 'did not like at all' and 100 means 'liked very much', how much did you like this player?"

Feeling thermometers. Participants' explicit prejudiced attitudes were measured with race-based feeling thermometers. Participants indicate their feelings toward four major American racial/ethnic groups—White Americans, Black Americans, Hispanic Americans, and Asian Americans—on a scale of 0 (very cold) to 100 (very warm) degrees. Because it may be difficult to interpret absolute ratings toward groups, explicit prejudice was scored as the difference between ratings for White and Black Americans, such that higher scores represented more pro-White/anti-Black attitudes.

Internal and external motivation to respond without prejudice scales. The IMS and EMS were used as in Study 1. Results involving EMS are reported in the SI.

Transparency and openness

In this paper, we have reported all data exclusions and all manipulations and measures in the study. All data, analysis code and research materials for Study 2 are available at osf.io/rnjgh. Data were analyzed using R Statistical Software (v3.6.1; R Core Team 2019). The study design, hypotheses, and analyses were preregistered at https://aspredicted.org/31S_ZH6. This study complies with TOP level 2 guidelines.

Results

Descriptive analyses

The IMS (α=0.90) and EMS (α=0.94) scales were reliable and produced distributions similar to Study 1. Descriptives and intercorrelations involving these and other key variables are shown in Table 1.

Table 1

Means, standard deviations, and correlations for key variables in Study 2.

Variable	1	2	3	4	5	6	7
1. Race difference in choice preference	–						
2. IMS	-0.13 [-.27, .02]	–					
3. EMS	0.01 [-.14, .16]	-0.06 [-.21, .08]	–				
4. Race difference in perceived reward	0.67*** [.58, .74]	-0.06 [-.20, .09]	0.13 [-.02, .27]	–			
5. Race difference in player liking	0.61*** [.50, .69]	-0.23** [-.36, -.09]	0.07 [-.08, .22]	0.68*** [.59, .75]	–		
6. Explicit prejudice	0.23** [.08, .36]	-0.58*** [-.67, -.48]	0.10 [-.04, .25]	0.18* [.04, .32]	0.32*** [.19, .45]	–	
7. Implicit prejudice	0.10 [-.05, .24]	-0.04 [-.18, .11]	0.14 [.00, .28]	0.04 [-.11, .19]	-0.04 [-.18, .11]	0.13 [-.01, .27]	–
M	0.52	7.50	4.18	-1.83	-4.26	4.58	0.21
SD	0.17	1.77	2.35	14.41	16.38	18.13	0.41

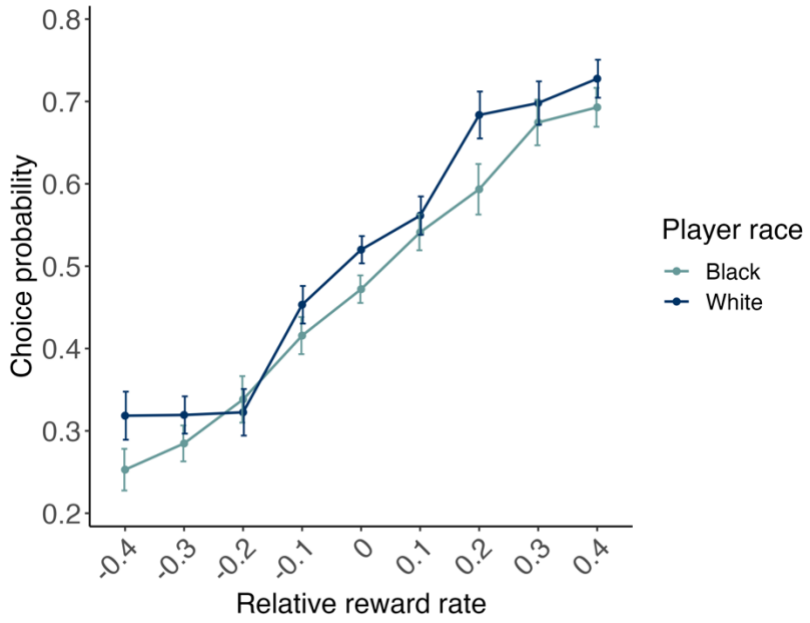
Note. Race difference in choice preference = proportion White over Black player choices in test phase, from 0 (choosing only Black players) to 1 (choosing only White players). IMS = internal motivation scale. EMS = external motivation scale (see SI for EMS results). Race difference in perceived reward = perceived reward rate for White – Black players (scored -100 to 100). Race difference in player liking = liking for White – Black players (scored-100 to 100). Explicit prejudice = feeling thermometer difference score for White – Black Americans; higher scores represent preferences for White over Black people. Implicit prejudice = IAT D score; higher scores represent stronger preference for White over Black faces. 95% confidence intervals for correlations shown in brackets.
*p<.05, **p<.01, ***p<.001.

Effects of race and IMS on instrumental learning

We first tested our primary hypothesis that race would influence participants' learning of players' reward value, as indicated by test phase choice behavior, and that this effect would be moderated by internal motivation, using the logistic regression described in Study 1. As in Study 1, this analysis produced an effect of race (OR=4.33, 95%CI=[1.17,16.06], p=.029; Figure 5), such that participants preferred to choose White players over Black players, as well as an effect of reward rate, such that higher rewarding players were chosen more often (OR=38.41, 95%CI=[22.02,66.99], p<.001).

Figure 5

Effects of race and reward on choice



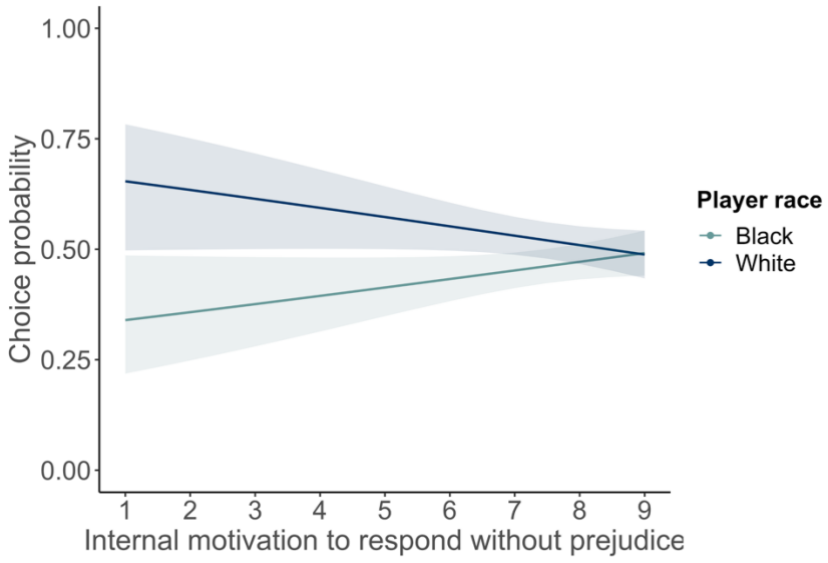
Note. Effects of relative reward rate and race on choice during test phase in Study 2, showing a preference for higher-rewarding players, and White players over Black players across relative reward rates. Relative reward rate is displayed on the x-axis, and choice probability displayed on the y-axis. Error bars represent standard errors.

The predicted Race x IMS interaction race was marginally significant (OR=0.85, 95%CI=[0.72,1.01], p=.058; see SI for full regression output). Nevertheless, given our a priori hypotheses and Study 1 results, we proceeded to test our specific predictions with simple slopes analyses. Replicating the pattern observed in Study 1, participants with relatively low IMS showed a choice preference for White over Black players ($M_{low\text{-}IMS}$=5.74, β=0.53, t=2.42, p=.015), whereas relatively high IMS participants showed no effect of race on preferences ($M_{high\text{-}IMS}$=9.25, β=-0.05, t=-0.25, p=.801; Figure 6).

As in Study 1, we examined choice preferences during the first 50 trials of training to test whether participants begin the task with a pro-White expectancy. Indeed, this analysis revealed a preference for White over Black players (race effect: OR=3.08, 95%CI=[1.30,7.31], p=.011), which was moderated by internal motivation (Race x IMS interaction: OR=0.86, 95%CI=[0.77,0.96], p=.009): low-IMS participants showed a pre-existing pro-White bias (β=0.28, t=1.94, p=.0.044), whereas high-IMS participants did not (β=-0.25, t=-1.77, p=.076).

Figure 6

Interaction effect of race and IMS on choice



Note. Interaction effect of race and IMS on choice in Study 2. IMS displayed on the x-axis and choice probability (probability of player being chosen) on the y-axis. The shaded areas represent the 95% confidence interval.

Implicit prejudice effects

On average, IAT *D* scores were greater than zero, indicating pro-White implicit bias (*M*=.21, *SD*=.41; *t*(178)=6.88, *p*<.001). When IAT scores were included as a predictor in the regression testing effects of relative reward and race on choice behavior, IAT was not a significant moderator of the race effect (Race x IAT interaction: OR=1.60, 95%CI=[0.78,3.30], *p*=.198). Thus, the effect of race on instrumental learning was not associated with individual differences in implicit prejudice, as measured by the IAT.

Explicit prejudice effects

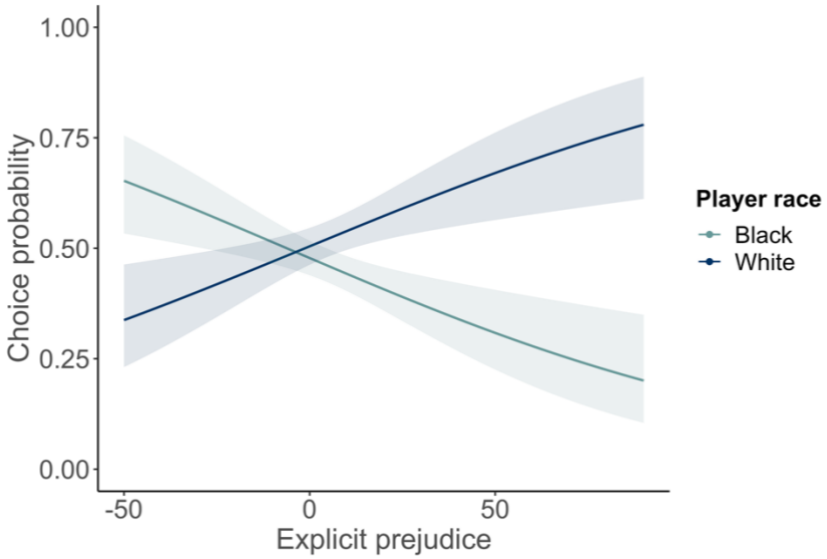
Participants reported more positive feelings toward White Americans (*M*=77.20, *SD*=17.86) than Black Americans (*M*=72.62, *SD*=21.53), *t*(178)=3.38, *p*<.001. A logistic regression that investigated main effects of race, actual reward rate, explicit prejudice, and the Race x Explicit Prejudice interaction produced a significant effect of reward rate (OR=38.49, 95%CI=[22.07,67.13], *p*<.001) and a significant Race x Explicit Prejudice interaction (OR=1.03, 95%CI=[1.01,1.05], *p*=.001; Figure 7; see SI for full regression output). Similar to the effects of IMS, relatively high-prejudice participants showed a choice preference for White over Black players ($M_{high\text{-}prejudice}$=22.73, β=0.88, *t*=0.25, *p*<.001), whereas choices of relatively low-prejudice participants were not affected by race ($M_{low\text{-}prejudice}$=-13.42, β=-0.27, *t*=-1.31, *p*=.191).

Furthermore, the pattern of pro-White expectancy during the first 50 trials of training, observed for low IMS participants, was also found for high prejudice participants (Race x

Explicit Prejudice interaction: OR=1.02, 95%CI=[1.01,1.03], p=.002; $M_{high\text{-}prejudice}$=22.73, β=0.54, t=3.23, p=.001; $M_{low\text{-}prejudice}$=-13.42, β=-0.19, t=-1.13, p=.260).

Figure 7

Interaction effect of Race x Explicit prejudice on choice.



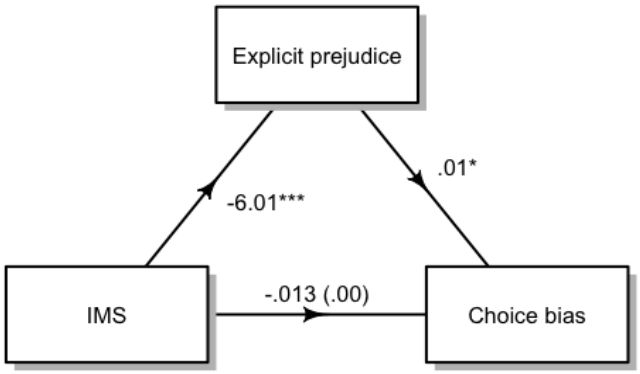
Note. Interaction effect of race and explicit prejudice on choice in Study 2. Explicit prejudice represents the different score between feeling thermometer ratings for White and Black Americans (higher values = more positive feelings toward White than Black people). Choice preference (y-axis) represents the probability of player being chosen. The shaded areas represent the 95% confidence interval.

To determine whether explicit prejudice could explain the observed effect of IMS, reported above, a regression analysis was conducted that investigated effects of race, reward rate, IMS, explicit prejudice, and both the Race x IMS and Race x Explicit Prejudice interactions. A significant Race x Explicit Prejudice interaction emerged (OR=1.03, 95%CI=[1.01,1.05], p=.005) but the Race x IMS interaction was no longer significant (OR=1.00, 95%CI=[0.82,1.23], p=.983).

Given the correlation between IMS and explicit prejudice, and in light of their theoretical relationship, such that IMS is conceptualized as a more trait-like construct, whereas the feeling thermometer could assess a more state-like attitude, this pattern was suggestive of mediation. To explore this possibility, we conducted a mediation analysis using a bootstrapping procedure with 1000 samples in R (v4.5.0; Tingley et al., 2014). This analysis indicated that the effect of IMS on choice bias was indeed fully mediated by explicit prejudice (indirect effect: β=-0.012, 95%CI=[-0.03,0.00], *p*=.044; Figure 8), consistent with the possibility that the effect of internal motivation on instrumental learning from Black vs. White players was driven by participants' explicit prejudice.

Figure 8

Meditation model showing effect of IMS and explicit prejudice on choice preference



Note. Meditation model showing effect of IMS and explicit prejudice on choice preference for White players compared to Black. Estimates indicate the unstandardized coefficients of the total effect. Numbers in brackets indicate the unstandardized coefficients of averaged direct effects.
*p<.05, **p<.01, ***p<.001.

Perceived reward rates

As in Study 1, participants' reported perceptions of reward rate corresponded with players' actual reward rates (β=0.54, 95%CI=[0.47, 0.60], p<.001), but did not vary by player race (Black players: M=45.22, SD=16.08; White players: M=43.39, SD=15.77; β=1.83, 95%CI=[-0.28,3.93], p=.089), nor the Race x IMS interaction (β=0.48, 95%CI=[-0.72,1.68], p=.436).

Despite the lack of a race effect on perceived rewards, variation in this perception was nevertheless associated with race differences in behavioral choice preferences (OR=1.10, 95%CI=[1.08, 1.12], p<.001). Importantly, however, when this perception was included as a covariate in a regression predicting choice behavior, significant effects remained for race (OR=3.91, 95%CI=[1.49,10.26], p=.006) and the Race x IMS interaction (OR=0.88, 95%CI=[0.78,1.00], p=.047) when IMS was a predictor and, in a separate regression, for race (OR=1.40, 95%CI=[1.12,1.75], p=.003) and the Race x Explicit Prejudice interaction (OR=1.02, 95%CI=[1.00,1.03], p=.014) when explicit prejudice was a predictor. Thus, the effect of race on choice preference, as a function of both IMS and prejudice, was independent of participants' subjective perceptions of players' sharing behavior.

Self-reported Player Liking

A multilevel regression analysis indicated that participants liked players who shared at higher rates, β=0.45, 95%CI=[0.38,0.53], p<.001. However, liking was also influenced by race, β=-10.67, 95%CI=[-20.03,-1.30], p=.026, such that participants indicated higher liking for Black players (M=63.29, SD=19.56) compared with White players (M=59.03, SD=19.92)—a pattern opposite to the choice preferences observed in behavior. The effect of race on liking was moderated by IMS, β=1.99, 95%CI=[0.78,3.21], p=.001, such that high IMS participants

reported greater liking for Black than White players ($M_{high\text{-}IMS}$=9.25, β=7.72, t=4.80, p<.001), whereas low IMS participants reported no race difference in liking ($M_{low\text{-}IMS}$=5.74, β=0.65, t=0.39, p=.700).

Participants' reported liking was also moderated by explicit prejudice (Race x Explicit Prejudice interaction: β=-0.29, 95%CI=[-0.41,-0.16], p<.001), such that low prejudice participants reported greater liking for Black players than White players ($M_{low\text{-}prejudice}$=-13.50, β=9.50, t=5.63, p<.001), whereas high prejudice participants reported no race difference in liking ($M_{high\text{-}prejudice}$=22.66, β=-0.95, t=-0.59, p=.558). This pattern of player liking was somewhat different than that observed for player choice: whereas race differences in player liking emerged among participants with lower prejudice/higher internal motivation participants, race effects on behavioral choice preferences emerged for participants with higher prejudice/lower internal motivation.

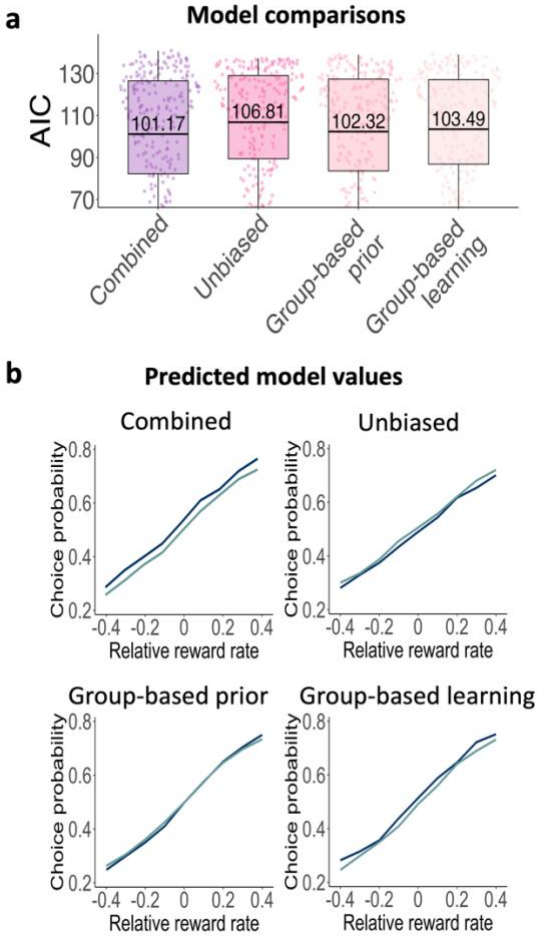
Computational modeling of learning mechanisms

To investigate the mechanisms involved in the observed race effects on instrumental learning, we conducted the same set of model comparisons tested in Study 1. Replicating Study 1, the best fitting model was the combined model, which specifies opposing expectancies (modeled as a prior) and separate learning rates for White and Black players (Figure 9a). The differences in AIC between the hypothesized combined model (AIC=94.33) and both the competing group-based prior model (AIC=98.62; ΔAIC=6.60) and the group-based learning model (AIC=99.44; ΔAIC=7.89) were more substantial here than in Study 1. Further, as in Study 1, the combined model produced simulated data that closely matched participants choice behavior (Figure 9b), and better model fit was associated with lower IMS, t(177)=3.38, p<.001, and higher explicit prejudice, t(177)=-2.16, p=.032. These results again suggest that participants

acquired and maintained a group bias through a combination of group-based initial expectancies and the updating of separate representations for Black and White players.

Figure 9

Computational model comparisons and simulated data



Note. **a,** Model comparisons between the hypothesized combined model with the unbiased model, group-based prior model, and group-based learning model in Study 2. **b,** Predicted values simulated by each model in Study 2.

Discussion

Study 2 replicated and extended the findings of Study 1. First, race was again found to influence social instrumental learning: White American participants expressed a choice preference for White players over Black players, and this effect was moderated by internal motivation. Although the moderating effect of IMS was only marginally significant in Study 2, *a priori* simple slopes analyses showed, as in Study 1, a significant choice preference for White over Black players among relatively low IMS participants, but no race preference among relatively high IMS participants. It is notable that, in Study 2, IMS was administered toward the end of an expanded set of questionnaires, and this may have influenced the size of its effect. Second, the effects of race on choice preference remained when adjusting for participants' explicit perceptions of player reward rates, suggesting that the effect of race on choice preferences may have operated implicitly. Finally, computational modeling again indicated that the best fitting model included an effect of race on initial expectancies (*prior*) as well as updating (*learning rates*). Together, these results replicated those of Study 1, demonstrating again that the race of one's interaction partner can influence how one forms impressions through direct social interaction via instrumental learning.

Study 2 also included measures of prejudiced attitudes and player liking to further illuminate the effects of race on social instrumental learning. Implicit prejudice, measured by the IAT, did not relate to the race effect in participants' choice preferences. However, explicit prejudice, measured by self-reported feeling thermometers, did moderate this effect, such that more explicitly prejudiced participants showed a strong preference for White over Black players, whereas lower explicitly prejudiced participants showed no race preference. This pattern was similar to that of IMS in direction but stronger in magnitude, suggesting that prejudice might

reflect a more proximal effect on learning associated with IMS. Supporting this idea, we found that explicit prejudice fully mediated the effect of IMS on bias in choice preferences.

In contrast to their choice behaviors, which tended to favor White players, participants' self-reported liking was greater for Black than White players—an effect that also depended on their internal motivation and explicit prejudice. Participants with relatively low IMS and high explicit prejudice reported greater liking for White players, whereas participants with relatively high IMS and low prejudice reported greater liking for Black players. It is notable that, on average, participants preferred Black over White players despite reporting more negative group-based attitudes toward Black Americans relative to White Americans—a pattern that may reflect a divergence in perceptions of abstract groups and their intergroup hierarchies compared with specific individuals with whom one interacts. Nevertheless, despite this divergence in average preferences, greater reported liking for Black over White players was correlated with higher IMS and lower explicit prejudice.

General discussion

The formation of impressions through direct social interaction and feedback relies on instrumental learning (Amodio, 2019; Hackel et al., 2015). Here, we asked whether race could affect this process of social-interactive impression formation. We found, in two studies, that race significantly influenced White Americans' formation of reward associations with Black compared with White interaction partners, even though the average sharing behaviors of Black and White players were equivalent. Importantly, this effect was moderated by internal motivation: participants with low internal motivation showed a preference for White players over Black players, whereas those with high internal motivation did not show a race preference. In

Study 2, this race bias in choice preference was also moderated by explicit prejudice, with greater bias among more highly-prejudiced participants, and explicit prejudice fully mediated the effect of IMS on choice behavior. These results demonstrate that race influences how people form impressions of others through direct social interaction, via instrumental learning—an effect that is pronounced among people with lower internal motivation and higher explicit prejudice.

This effect of race on instrumental learning, while expressed in participants' choice behavior, was not evident in their self-reported perceptions of player feedback. That is, while participants exhibited more positive reward associations with White than Black players in their choice behavior, this preference was not evident in their explicit reports of player reward rates. This pattern is consistent with the possibility that instrumental learning effects are expressed implicitly, independent of one's explicit awareness of reward contingencies (Amodio, 2019; Amodio & Ratner, 2011; Knowlton et al., 1996). Considered in the context of participants' explicit task goals—to accurately learn player reward rates and to choose the most frequently-sharing players in order to earn the most possible money—these results suggest an implicit effect of race on instrumental learning that countered their explicit intentions.

This instrumental learning account of race-biased reward learning was further supported by computational modeling analyses. The best fitting model (*combined model*) suggests that participants held race-biased initial expectancies (modeled as a prior) as well as separate learning rates for Black and White players. This combination of a group-based prior and group-based learning rates suggests a two-step process, whereby race induces different initial expectancies for Black and White players, which are then maintained by holding and updating separate representations based on player race. By identifying separate learning rates for Black and White players, this finding further supports the conclusion that race influenced how perceivers learned

from players, as opposed to merely biasing their decisions without affecting their learning. More broadly, this combined model provides an explanation for how prejudices can persist despite the absence of actual differences in group members' behavior.

Together, these studies demonstrate that race can influence impression formation in the context of direct social interaction, and they introduce a mechanism through which this effect occurs. We also observed individual differences in the effect of race on instrumental learning, such that it was pronounced among individuals with relatively low internal motivation and high explicit prejudice. These findings advance our understanding of how prejudice forms and persists in the context of intergroup interactions.

An instrumental learning account of racial prejudice

An enduring critique of research on prejudiced attitudes is that attitudes are often poor predictors of behavior (Duckitt, 1992; Lai & Wilson, 2021). Yet, to date, research has not considered the potential role of instrumental learning as a component of a prejudiced attitude. Our findings, which demonstrate the role of instrumental learning in the formation of group-based preferences, suggests a novel perspective on prejudice and its relation to behavior. Instrumental learning forms through action and feedback, and it is supported by neurocognitive systems involved in reward processing and motor function (Liljeholm & O'Doherty, 2012). Hence, a racial bias in instrumentally-learned reward associations represents a behavior-based prejudiced attitude, akin to the conative component of attitudes in the classic tripartite model (Amodio, 2019; Breckler, 1984).

Our use of computational modeling further clarified the behavioral basis of prejudice in these studies. These analyses showed that participants' trial-by-trial behavioral responses fit with classic reinforcement learning patterns, such that they were sensitive to probabilistic feedback

and were updated incrementally according to a prediction error. In particular, the finding that behavior was best explained by a model with separate learning rates for Black and White players, in addition to a group-based prior, supports our hypothesis that race influenced the incremental formation of behavioral-based reward associations. These findings support our proposal that an instrumental learning component of prejudice complements traditional attitude components, contributing to a more comprehensive account whereby prejudice is formed and expressed via multiple cognitive processes and behavioral channels (Amodio, 2014; Amodio & Cikara, 2021).

A limitation of the current modeling approach is that it did not permit reliable analyses of participant-level parameter estimates. It remains possible that learning rates could differ by race; for example, White learners might update their representations of White players more readily than Black players in response to prediction errors. However, it is also possible that learning rates, as operationalized in the reinforcement learning framework used here, would not differ in magnitude. That is, given existing pro-White priors, learners might be similarly slow to update representations of both White and Black interaction partners, thereby maintaining their initial prejudices. A third possibility is that updating depends on the valence of prediction errors, with greater ingroup updating for positive predictions errors and greater outgroup updating for negative prediction errors; however, this model (see model e in SI)—which includes separate positive and negative learning rates for Black and White players—did not provide the best fit to behavioral data. Although we declined to interpret participant-level parameter estimates, these estimates are reported in the SI and show mixed results: in Study 1, there were no differences, and in Study 2, learning rate estimates were higher for White than Black players. In future research, other computational procedures may be used to obtain valid participant-level parameter estimates to address potential differences. Nevertheless, the consistent finding that a model with

two separate learning rates fit the data better than a model with one learning rate indicates that participants maintained separate representations of reward association for Black and White players and updated them according to different learning rules. This in itself was a novel contribution of the present computational analysis.

Individual differences in race effects on social instrumental learning

The effect of race on social instrumental learning was moderated by participants' internal motivation to respond without prejudice and explicit prejudice: participants with relatively low internal motivation (and high prejudice in Study 2) formed stronger reward associations with White than Black players, whereas participants with relatively high internal motivation and low prejudice showed no learning bias. However, the specific processes through which internal motivation and prejudice modulated this learning remains unclear. Prior research suggests multiple possibilities which could function in combination to produce our observed effects.

First, people with high internal motivation tend to approach interracial interactions and seek to express their egalitarian values, whereas people with low internal motivation tend to avoid interracial interactions (Plant & Devine, 2009; Plant et al., 2010). This tendency could have influenced participants' choice preferences in the task, despite their explicit goal to choose players based on their individual sharing rates (which did not vary by race). Indeed, prior research shows that the avoidance of outgroup interactions can lead to the formation of negative associations (Fazio et al., 2004; Allidina & Cunningham, 2021; Bai et al., 2022). Although the learning task used in our studies were designed to minimize the effects of avoidance across learning trials—participants knew that only one player would share on each trial, and reward rates were equated between groups—computational modeling results indicated that race influenced participants race-based prior, in addition to their updating, and that this model

provided fit better to the behavior of participants with lower IMS. Thus, the observed effects of internal motivation could in part be due to its implications for intergroup approach tendencies.

Second, internal motivation has been linked to more effective self-regulation (Devine et al., 2002; Plant & Devine, 2009). For example, high internal motivation (combined with low external motivation) has been associated with more sensitive neural detection of automatic stereotypes and more effective inhibition of these biases in behavior (Amodio et al., 2008). It is possible that, in our studies, high internal motivation participants were also more effective at directing their attention to task-relevant information such as the actual reward feedback provided by a player—a form of proactive control (Amodio & Swencionis, 2018).

Third, high internal motivation has been associated with more open-minded categorizations of race and ethnicity, such that high internally motivated people are more likely to view a person as multiracial (Chen et al., 2014). Low internal motivation individuals, by comparison, are more likely to view people rigidly in terms of single racial or ethnic categories. This could have led low IMS participants in the present studies to view players more categorically (i.e., monoracially), which would have strengthened the degree of racial associations with reward feedback. For high IMS participants, race may have been less mono-categorically perceived and thus had a weaker impact on reward learning.

Finally, although internal motivation suggests a rich set of processes that may influence social instrumental learning, Study 2 revealed that explicit prejudiced attitude had a stronger, more proximal effect on group-based learning. Prejudiced attitude should directly lead to a racial difference in initial expectancies, consistent with our computational modeling evidence for a group-based prior. Prejudiced attitude may also mute the impact of positive prediction errors from outgroup members and negative prediction errors from ingroup members. These effects

may explain how explicit prejudice, expressed as self-reported feelings, could lead to a behavior-based instrumental bias. Moreover, it is possible that internal motivation influenced learning via these effects of prejudiced attitude.

Each of these potential effects of internal motivation and explicit prejudice could have played a role in the observed results. An important goal of future research is to determine specifically how these, and other aspects of internal motivation and prejudice can modulate intergroup instrumental learning. An understanding of these processes will enhance our understanding of prejudice formation in social interactions and inform interventions to reduce this form of prejudice.

Using an instrumental learning approach to study interracial interactions

In the present work, the process of forming an impression through social interaction was characterized in terms of instrumental learning. Whereas prior research has examined passive modes of impression formation, such as through passive observation of behavior, communication of verbal descriptions, or exposure to conceptual or evaluative associations, an instrumental learning approach conceptualizes impression formation as resulting from the interplay of actions and feedback. Moreover, by incorporating models of reinforcement learning, an instrumental learning approach can bring ideas and methods from cognitive science and neuroscience to bear on questions about impression formation. For example, by considering the basic parameters of reinforcement learning and developing corresponding computational models, we were able to test new hypotheses about the specific ways in which group membership influences impression formation in social interaction. This instrumental learning framework has been used recently to characterize social-interactive impression formation (Hackel et al., 2015, 2020) and to illuminate how this process contributes to automatic responses (Hackel et al., 2019) and context-specific

impressions (Hackel et al., 2022). Our studies extend this approach to understand how race affects socio-interactive learning and contributes to racial discrimination. In future research, an integration of instrumental learning with existing impression formation approaches may yield a more comprehensive account of race effects on social interaction.

A limitation of this research, however, is that to examine instrumental learning mechanisms in a social context, we used an experimental task that reduced social interactions to their most essential elements: action and feedback. This approach prioritized experimental control at the expense of ecological validity. Yet, there may be many additional factors in real-life social interactions that could further influence the observed effects of race on instrumental learning. Having demonstrated a basic effect of race on social instrumental learning in a relatively minimal interaction context, subsequent work may build on these findings to understand how this effect will generalize to more complex forms of interaction.

Constraints on Generality

Our theoretical question concerns the effect of race on dominant group members' impressions of minoritized individuals, in an effort to understand the influence of prejudice on social learning and its expression in behavior. We examined this question in the American context, in White American participants' perceptions of Black and White people (Remedios, 2022). We focused on this context because of the history of White Americans' oppression of Black people and its persisting effects in systemic and individual discrimination, and because this is the context with which the authors are most familiar. This focus limits our ability to generalize these findings to other contexts. It is possible that the observed patterns of bias in instrumental learning may generalize to other social and cultural settings, with variation in these effects linked to context-specific factors. For example, gender or ethnic stereotypes within a

culture could have unique effects on expectancies and learning rates during instrumental learning. Additionally, while our questions focused on dominant group members' perceptions of a minoritized racial group, it is also crucial to study processes from the perspective of minority group members (Shelton, 2000). To address the generalization of these findings, additional research will need to explore potential variations in these processes in other contexts and from other perspectives.

Conclusions

We have demonstrated that race influences impression formation in direct social interaction through the process of instrumental learning. This effect of race on instrumental learning is pronounced among people with low internal motivation to respond without prejudice and high explicit prejudice, and computational modeling suggests it emerges from a combination of racial bias in initial interaction expectancies and in the updating of impressions over the course of repeated interaction. Together, these findings identify a social-interactive mode of prejudice formation that is expressed in behavior which complements existing theories of prejudiced attitudes and may inform new approaches to prejudice reduction.

References

Allidina, S., & Cunningham, W. A. (2021). Avoidance begets avoidance: A computational account of negative stereotype persistence. *Journal of Experimental Psychology: General*, *150*(10), 2078–2099. https://doi.org/10.1037/xge0001037

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience*, *15*(10), 670-682. https://doi.org/10.1038/nrn3800

Amodio, D. M. (2019). Social cognition 2.0: An interactive memory systems account. *Trends in Cognitive Sciences*, *23*(1), 21-33. https://doi.org/10.1016/j.tics.2018.10.002

Amodio, D. M., & Cikara, M. (2021). The social neuroscience of prejudice. *Annual Review of Psychology*, *72*, 439-469. https://doi.org/10.1146/annurev-psych-010419-050928

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91*(4), 652–661. https://doi.org/10.1037/0022-3514.91.4.652

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2008). Individual differences in the regulation of intergroup bias: the role of conflict monitoring and neural signals for control. *Journal of Personality and Social Psychology*, *94*(1), 60-74. https://doi.org/10.1037/0022-3514.94.1.60

Amodio, D. M., & Ratner, K. G. (2011). A memory systems model of implicit social cognition. *Current Directions in Psychological Science*, *20*(3), 143-148. https://doi.org/10.1177/0963721411408562

Amodio, D. M., & Swencionis, J. K. (2018). Proactive control of implicit bias: A theoretical model and implications for behavior change. *Journal of Personality and Social Psychology*, *115*(2), 255-275. https://doi.org/10.1037/pspi0000128

Averbeck, B., & O'Doherty, J. P. (2022). Reinforcement-learning in fronto-striatal circuits. *Neuropsychopharmacology*, *47*(1), 147-162. https://doi.org/10.1038/s41386-021-01108-0

Bai, X., Fiske, S. T., & Griffiths, T. L. (2022). Globally Inaccurate Stereotypes Can Result From Locally Adaptive Exploration. *Psychological Science*, *33*(5), 671–684. https://doi.org/10.1177/09567976211045929

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Biernat, M., Fuegen, K., & Kobrynowicz, D. (2010). Shifting standards and the inference of incompetence: Effects of formal and informal evaluation tools. *Personality and Social Psychology Bulletin*, *36*(7), 855-868. https://doi.org/10.1177/0146167210369483

Breckler, S. J. (1984). Empirical validation of affect, behavior, and cognition as distinct components of attitude. *Journal of personality and social psychology*, *47*(6), 1191-1205. https://doi.org/10.1037/0022-3514.47.6.1191

Brewer, M. B. (1988). "A dual process model of impression formation," in Advances in Social Cognition, Vol. 1, eds T. K. Srull, and R. S. Wyer, Jr. (Hillsdale, NJ: Lawrence Erlbaum Associates), 1–36.

Chen, J. M. (2019). An integrative review of impression formation processes for multiracial individuals. *Social and Personality Psychology Compass*, *13*(1), e12430. https://doi.org/10.1111/spc3.12430

Chen, J. M., Moons, W. G., Gaither, S. E., Hamilton, D. L., & Sherman, J. W. (2014). Motivation to control prejudice predicts categorization of multiracials. *Personality and Social Psychology Bulletin*, *40*(5), 590-603. https://doi.org/10.1177/0146167213520457

Cikara, M., Martinez, J. E., & Lewis Jr, N. A. (2022). Moving beyond social categories by incorporating context in social psychological theory. *Nature Reviews Psychology*, *1*, 537-549.

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised. *Advances in experimental social psychology* (Vol. 56, pp. 131-199). Academic Press. https://doi.org/10.1016/bs.aesp.2017.03.001

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20-33. https://doi.org/10.1037/0022-3514.44.1.20

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5–18. https://doi.org/10.1037/0022-3514.56.1.5

Devine, P. G., & Monteith, M. J. (1993). The role of discrepancy-associated affect in prejudice reduction. *Affect, cognition and stereotyping*, 317-344. Academic Press. https://doi.org/10.1016/B978-0-08-088579-7.50018-1

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: the role of motivations to respond without prejudice. *Journal of Personality and Social Psychology*, *82*(5), 835-848. https://doi.org/10.1037/0022-3514.82.5.835

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of personality and social psychology*, *82*(1), 62-68. https://doi.org/10.1037/0022-3514.82.1.62

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of experimental social psychology*, *33*(5), 510-540. https://doi.org/10.1006/jesp.1997.1331

Duckitt, J. (1992). Prejudice and behavior: A review. *Current Psychology, 11*(4), 291-307. https://doi.org/10.1007/BF02686787

Eargle, D., Gureckis, T., Rich, A.S., McDonnell, J., & Martin, J.B. (2020). PsiTurk: An open platform for science on Amazon Mechanical Turk (Version v3.3.0).

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychological science*, *17*(5), 383-386. https://doi.org/10.1111/j.1467-9280.2006.01716.x

Eckstein, M. K., Wilbrecht, L., & Collins, A. G. E. (2021). What do Reinforcement Learning Models Measure? Interpreting Model Parameters in Cognition and Neuroscience. *Current opinion in behavioral sciences*, *41*, 128–137. https://doi.org/10.1016/j.cobeha.2021.06.004

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology, 87*(3), 293–311. https://doi.org/10.1037/0022-3514.87.3.293

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of personality and social psychology*, *69*(6), 1013-1027. https://doi.org/10.1037/0022-3514.69.6.1013

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), Handbook of social psychology: Vol. 2 (4th ed., pp. 357-411). New York: McGraw-Hill.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). Competence and warmth scales. *Journal of Personality and Social Psychology*, *82*(6). https://doi.org/10.1037/t35954-000

Foerde, K., & Shohamy, D. (2011). The role of the basal ganglia in learning and memory: insight from Parkinson's disease. *Neurobiology of learning and memory*, *96*(4), 624-636. https://doi.org/10.1016/j.nlm.2011.08.006

Frank, M. J., Seeberger, L. C., & O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, *306*(5703), 1940-1943. https://doi.org/10.1126/science.1102941

Gaertner, S. L., & Dovidio, J. F. (2000). The aversive form of racism. In C. Stangor (Ed.), *Stereotypes and prejudice: Essential readings* (pp. 289–304). Psychology Press.

Gilbert, S. J., Swencionis, J. K., & Amodio, D. M. (2012). Evaluative vs. trait representation in intergroup social judgments: Distinct roles of anterior temporal lobe and prefrontal

cortex. *Neuropsychologia*, *50*(14), 3600-3611. https://doi.org/10.1016/j.neuropsychologia.2012.09.002

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of personality and social psychology*, *85*(2), 197-216. https://doi.org/10.1037/0022-3514.85.2.197

Gureckis, T.M., Martin, J., McDonnell, J., Rich, A.S., Markant, D., Coenen, A., Halpern, D., Hamrick, & J.B., Chan, P. (2016) psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavioral Research Methods, 48*(3), 829-842. https://doi.org/10.3758/s13428-015-0642-8

Hackel, L. M., & Amodio, D. M. (2018). Computational neuroscience approaches to social cognition. *Current opinion in psychology*, *24*, 92–97. https://doi.org/10.1016/j.copsyc.2018.09.001

Hackel, L. M., Berg, J. J., Lindström, B. R., & Amodio, D. M. (2019). Model-Based and Model-Free Social Cognition: Investigating the role of habit in social attitude formation and choice. *Frontiers in Psychology, 10*, 2592. https://doi.org/10.3389/fpsyg.2019.02592

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nature Neuroscience, 18*(9), 1233-1235. https://doi.org/10.1038/nn.4080

Hackel, L. M., Mende-Siedlecki, P., & Amodio, D. M. (2020). Reinforcement learning in social interaction: The distinguishing role of trait inference. *Journal of Experimental Social Psychology, 88*, 103948. https://doi.org/10.1016/j.jesp.2019.103948

Hackel, L. M., Mende-Siedlecki, P., Loken, S., & Amodio, D. M. (2022). Context-dependent learning in social interaction: Trait impressions support flexible social choices. J*ournal of Personality and Social Psychology, 123*, 655-675. https://doi.org/10.31234/osf.io/symrj

Hamilton, D. L., Sherman, S. J., & Ruvolo, C. M. (1990). Stereotype-based expectancies: Effects on information processing and social behavior. *Journal of Social Issues*, *46*, 35-60. https://doi.org/10.1111/j.1540-4560.1990.tb01922.x

Hass, R. G., Katz, I., Rizzo, N., Bailey, J., & Eisenstadt, D. (1991). Cross-racial appraisal as related to attitude ambivalence and cognitive complexity. *Personality and Social Psychology Bulletin*, *17*(1), 83-92. https://doi.org/10.1177/0146167291171013

Kawakami, K., Amodio, D. M., & Hugenberg, K. (2017). Intergroup perception and cognition: An integrative framework for understanding the causes and consequences of social categorization. In *Advances in experimental social psychology* (Vol. 55, pp. 1-80). Elsevier Academic Press. https://doi.org/10.1016/bs.aesp.2016.10.001

Kawakami, K., Williams, A., Sidhu, D., Choma, B. L., Rodriguez-Bailón, R., Cañadas, E., Chung, D., & Hugenberg, K. (2014). An eye for the I: Preferential attention to the eyes of in-group members. *Journal of Personality and Social Psychology*, **107**, 1–20. https://doi.org/10.1037/a0036838

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, *273*(5280), 1399-1402. https://doi.org/10.1126/science.273.5280.1399

Krosch, A. R., Tyler, T., & Amodio, D. M. (2017). Race and recession: The effect of economic scarcity and egalitarian motivation on racial discrimination. *Journal of Personality and Social Psychology, 113*, 892-909. https://doi.org/10.1037/pspi0000112

Kunda, Z., & Sherman-Williams, B. (1993). Stereotypes and the construal of individuating information. *Personality and Social Psychology Bulletin, 19*(1), 90-99. https://doi.org/10.1177/0146167293191010

LaCosse, J., & Plant, E. A. (2020). Internal motivation to respond without prejudice fosters respectful responses in interracial interactions. *Journal of Personality and Social Psychology, 119*(5), 1037–1056. https://doi.org/10.1037/pspi0000219

Lai, C. K., & Wilson, M. E. (2021). Measuring implicit intergroup biases. Social and *Personality Psychology Compass, 15*(1), e12573. https://doi.org/10.1111/spc3.12573

Li, T., Cardenas-Iniguez, C., Correll, J., & Cloutier, J. (2016). The impact of motivation on race-based impression formation. *Neuroimage, 124*, 1-7. https://doi.org/10.1016/j.neuroimage.2015.08.035

Liljeholm, M., & O'Doherty, J. P. (2012). Contributions of the striatum to learning, motivation, and performance: an associative account. *Trends in cognitive sciences, 16*(9), 467-475. https://doi.org/10.1016/j.tics.2012.07.007

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(2), 433-442. https://doi.org/10.3758/s13428-016-0727-z

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods, 47*, 1122-1135. https://doi.org/10.3758/s13428-014-0532-5

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual review of psychology, 51*(1), 93-120. https://doi.org/10.1146/annurev.psych.51.1.93

Maddox, K. B. (2004). Perspectives on racial phenotypicality bias. *Personality and Social Psychology Review*, *8*(4), 383-401. https://doi.org/10.1207/s15327957pspr0804_4

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37*(5), 435–442. https://doi.org/10.1006/jesp.2000.1470

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, **7**, 3–35. https://doi.org/10.1037/1076-8971.7.1.3

Neuberg, S. L. (1989). The goal of forming accurate impressions during social interactions: Attenuating the impact of negative expectancies. *Journal of Personality and Social Psychology*, *56*(3), 374–386. https://doi.org/10.1037/0022-3514.56.3.374

Piray P., Dezfouli A., Heskes T., Frank M.J., Daw N.D. (2019) Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies. *PLOS Computational Biology 15*(6): e1007043. https://doi.org/10.1371/journal.pcbi.1007043

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology, 75*, 811–832. https://doi.org/10.1037/0022-3514.75.3.811

Plant, E. A., & Devine, P. G. (2009). The active control of prejudice: unpacking the intentions guiding control efforts. *Journal of personality and social psychology*, *96*(3), 640-652. https://doi.org/10.1037/a0012960

Plant, E. A., Devine, P. G., & Peruche, M. B. (2010). Routes to positive interracial interactions: Approaching egalitarianism or avoiding prejudice. *Personality and Social Psychology Bulletin*, *36*(9), 1135-1147. https://doi.org/10.1177/0146167210378018

Olson, I. R., McCoy, D., Klobusicky, E., & Ross, L. A. (2013). Social cognition and the anterior temporal lobes: a review and theoretical framework. *Social Cognitive and Affective Neuroscience, 8*(2), 123-133. https://doi.org/10.1093/scan/nss119

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Remedios, J. D. (2022). Psychology must grapple with Whiteness. *Nature Reviews Psychology, 1*(3), 125-126. https://doi.org/10.1038/s44159-022-00024-4

Schaaf, J. V., Weidinger, L., Molleman, L., & van den Bos, W. (2023). Test-Retest Reliability of Reinforcement Learning Parameters. *Behavior Research Methods.* https://doi.org/10.3758/s13428-023-02203-4

Schultner, D., Stillerman, B., Lindström, B., Hackel, L. M., Hagen, D., Jostmann, N., & Amodio, D. M. (2022, May 20). Societal stereotypes shape learning to produce group-based preferences. https://doi.org/10.31234/osf.io/mwztc

Shelton, J. N. (2000). A reconceptualization of how we study issues of racial prejudice. *Personality and Social Psychology Review, 4*(4), 374-390. https://doi.org/10.1207/S15327957PSPR0404_6

Shelton, J. N., & Richeson, J. A. (2006). Interracial interactions: A relational approach. *Advances in Experimental Social Psychology, 38,* 121-181. https://doi.org/10.1016/S0065-2601(06)38003-3

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 135, pp. 223-260). Cambridge: MIT press.

Tingley D, Yamamoto T, Hirose K, Keele L, Imai K (2014). "Mediation: R Package for Causal Mediation Analysis." *Journal of Statistical Software*, 59(5), 1–38. http://www.jstatsoft.org/v59/i05/.

Uleman, J. S., & Kressel, L. M. (2013). A brief history of theory and research on impression formation. In D. E. Carlston (Ed.). *Oxford handbook of social cognition* (pp. 53-73): New York: Oxford University Press.

Vingilis-Jaremko, L., Kawakami, K., & Friesen, J. P. (2020). Other-groups bias effects: Recognizing majority and minority outgroup faces. *Social Psychological and Personality Science*, **11**, 908–916. https://doi.org/10.1177/1948550620919562

Wang, Y., Collins, J.A., Koski, J., Nugiel, T., Metoki, A., & Olson, I.R. (2017). Dynamic neural architecture for social knowledge retrieval. *Proceedings of the National Academy of Sciences, 114*(16). E3305-E3314. https://doi.org/10.1073/pnas.1621234114

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in neuroinformatics, 7*(14). https://doi.org/10.3389/fninf.2013.00014.