Racial Bias Shapes Social Reinforcement Learning

Björn Lindström, Ida Selbing, Tanaz Molapour and Andreas Olsson

Psychological Science published online 23 January 2014
DOI: 10.1177/0956797613514093

Research Article

Psychological Science
1–9

DOI: 10.1177/0956797613514093
pss.sagepub.com

Racial Bias Shapes Social Reinforcement Learning

Björn Lindström, Ida Selbing, Tanaz Molapour, and Andreas Olsson
Department of Clinical Neuroscience, Karolinska Institutet

Abstract

Both emotional facial expressions and markers of racial-group belonging are ubiquitous signals in social interaction, but little is known about how these signals together affect future behavior through learning. To address this issue, we investigated how emotional (threatening or friendly) in-group and out-group faces reinforced behavior in a reinforcement-learning task. We asked whether reinforcement learning would be modulated by intergroup attitudes (i.e., racial bias). The results showed that individual differences in racial bias critically modulated reinforcement learning. As predicted, racial bias was associated with more efficiently learned avoidance of threatening out-group individuals. We used computational modeling analysis to quantitatively delimit the underlying processes affected by social reinforcement. These analyses showed that racial bias modulates the rate at which exposure to threatening out-group individuals is transformed into future avoidance behavior. In concert, these results shed new light on the learning processes underlying social interaction with racial-in-group and out-group individuals.

Keywords

learning, racial and ethnic attitudes and relations, social influences, emotions

Received 3/21/13; Revision accepted 10/25/13

The social world contains a multitude of decision-making problems with important behavioral consequences—for example, whom should you interact with and whom should you avoid? Reinforcement learning (RL), in which actions are based on the averaged positive or negative experiences previously associated with each option, provides a simple and powerful solution to such problems (Rangel, Camerer, & Montague, 2008). For example, you might learn to dislike and avoid your new neighbor because of a previous aversive experience, such as an unfriendly gaze (Blair, 2003). Social forms of RL like this, in which your behaviors toward others are based on the emotional significance of their previous actions toward you, are likely to be ubiquitous in social interaction.

Facial expressions constitute one such class of reinforcers that can play an important role in social RL by signaling benign and malign behavioral intentions (Blair, 2003; Heerey, 2013). The face also conveys information about an individual's racial- and ethnic-group belonging, and such markers of social-group membership might affect learning from social reinforcement. In support of this conjecture, previous research has shown superior memory for angry, relative to neutral or happy, racial-out-group faces (Ackerman et al., 2006) and more persistent learning of fear in response to racial-out-group faces compared with racial-in-group faces (Navarrete et al., 2009; Olsson, Ebert, Banaji, & Phelps, 2005). It is, however, unknown whether the group membership (race) of a face can modulate the value of its emotional expression and, thus, shape behavior during social interaction.

In addition, little is known about the possible mechanisms underlying the impact of social reinforcement on behavior. Previous research has shown that individual differences in racial bias affect the development of interracial interactions (Baron & Banaji, 2006; Stanley, Sokol-Hessner, Banaji, & Phelps, 2011), which suggests that individual differences in racial bias might modulate the

Corresponding Author:
Björn Lindström, Department of Clinical Neuroscience, Karolinska Institutet, Nobelsväg 9, Stockholm 171 77, Sweden
E-mail: bjorn.lindstrom@ki.se

way we learn about others through their reinforcing actions. For example, your attitudes toward the racial group of a neighbor might determine how that neighbor's unfriendly gaze affects your future decisions to interact or avoid him or her. In spite of its importance for understanding social behavior in racially heterogeneous milieus, available research does not speak to whether and, if so, how racial bias interacts with basic aspects of social RL, such as the reinforcing value of others' emotional expressions.

To address these issues, we examined how racial bias shapes avoidance behavior reinforced by threatening (angry) and friendly (happy) facial expressions posed by members of racial in- and out-groups, which is analogous to how one's behavior is shaped as a function of the facial feedback received from others in response to one's actions (Blair, 2003). Understanding the mechanics of learning through social reinforcers, and how these mechanisms are modulated by individual differences in racial bias, is of key importance for unraveling the motivational basis of social influence. Our approach goes beyond the existing literature on learning about racial-out-group faces by describing a novel experimental model of a dynamic social situation in which facial race and emotion reinforce behavior. We used computational modeling to differentiate between two distinct hypotheses about the underlying computational learning mechanisms influenced by social reinforcement, and we tied these learning mechanisms to individual differences in racial bias.

Individual differences in racial bias can theoretically be examined on (at least) two different levels of racial attitudes: the explicit, conscious level and the implicit, presumably more unconscious, level (Greenwald & Banaji, 1995). In the present study, we used two standard tests to assess explicit and implicit racial bias—the Modern Racism Scale (MRS) and the Implicit Association Test (IAT), respectively. The MRS measures self-reported attitudes toward a racial out-group (McConahay, 1986), and the race IAT assesses the associations among concepts related to racial groups and emotional values (e.g., Black and "bad") by measuring response latency (Greenwald & Banaji, 1995).

Previous research has shown that both emotional expressions and markers of racial-group belonging can modulate learning (Navarrete et al., 2009; Öhman & Mineka, 2001; Olsson et al., 2005). For example, learned fear responses to images of faces belonging to members of a racial out-group relative to an individual's own race are more resistant to extinction (Olsson et al., 2005). These results resemble fear learning in response to angry in-group faces (Öhman & Mineka, 2001), which suggests that both racial-out-group faces and angry in-group faces are perceived as potentially threatening. Along the same lines, other research has shown that angry out-group faces can reverse the *same-race bias* in recognition memory (Ackerman et al., 2006), so that although recognition memory usually tends to be worse for neutral out-group relative to in-group faces, the opposite is true for angry faces. The lines of research reviewed here have focused on either classically conditioned associations between faces and aversive stimuli (typically electric shocks) or recognition memory for faces in the absence of any reinforcement. We reasoned that of equal, or perhaps even greater, relevance for today's interactive social world is to understand how one's behavior is modulated by social reinforcers, such as threatening and friendly faces posed by individuals from one's racial in- or out-group.

To address this issue, we devised a probabilistic RL task in which neutral stimuli were arbitrarily associated with social reinforcers (threatening and friendly facial expressions). Similar tasks, typically using monetary rewards, have been successfully used to analyze the psychological and neural mechanisms underlying learning and decision making in the field of neuroeconomics (e.g., Glimcher, 2011; Rangel et al., 2008). In the research presented here, we aimed to elucidate how social reinforcers and markers of racial-group belonging interact to affect behavior through RL by (a) addressing how individual differences in racial bias modulate social RL from facial expressions posed by members of racial in- and out-groups and (b) investigating, using computational modeling, the underlying learning mechanisms and how these are shaped by racial bias. First, we reasoned that because angry out-group faces (Black faces presented to a White experimental group) were potentially perceived as more threatening than were angry in-group and happy out-group faces, learning to avoid them should be easier. It is important to note that we predicted that this learning bias would be directly related to individual differences in racial bias. The inclusion of friendly faces was an important experimental control that allowed us to determine specifically how racial bias influences learning from in- and out-group individuals by estimating both main and interaction effects of racial group, emotion, and racial bias.

Second, we aimed to understand not only whether, but also how, social reinforcers (here, emotional expressions) and racial belonging together would affect RL by elucidating the underlying computations. The brain needs to perform a range of computations to choose the best action at any moment, and these computations can be parsed into several subprocesses, of which two of the most important are *outcome evaluation* (OE; ascertaining the positive or negative value of the outcome) and *outcome learning* (OL; determining how fast, i.e., at what rate, an unexpected outcome affects subsequent behavior; Rangel et al.,

2008). We used standard computational RL models to quantitatively test two contrasting hypotheses—the OE hypothesis and the OL hypothesis—to learn which of these computational subprocesses social reinforcement would primarily influence. Furthermore, we predicted that individual differences in racial bias would be correlated with the RL-model parameter involved in learning from threatening out-group faces.

These hypotheses were evaluated by fitting a set of parameters commonly used in RL to decompose overt behavior into separable subprocesses (Kable & Glimcher, 2007; Katahira, Fujimura, Okanoya, & Okada, 2011). The OE hypothesis was modeled by fitting different *outcome-value* parameters for the different experimental conditions (see the Results section in this article and the Model Descriptions section in Computational Modeling in the Supplemental Material available online). Outcome value is conceptually similar to monetary value and should drive behavior in the same way (e.g., a cue that predicts a $10 reward motivates actions to a greater degree than does a cue that predicts a $1 reward; Dayan & Balleine, 2002). The obvious difference between monetary and social reinforcement is that the latter has no fixed currency and, therefore, its value must be estimated directly from the decision maker's behavior (Katahira et al., 2011). Thus, if the outcome value assigned to social feedback delivered by in-group and out-group individuals differs, this class of models should provide the best account of the data.

In contrast, the OL hypothesis was modeled with different *learning-rate* parameters. The learning rate describes what impact any given discrepancy between expected and experienced outcomes (the prediction error) will have on the subsequent value of the action that led to that outcome (e.g., Glimcher, 2011). With a high learning rate, the most recent outcome will have a large impact on the subsequent value of that action, whereas a low learning rate gives a slow integration of values that results in a more stable behavior (Glimcher, 2011). Following this logic, if social reinforcement delivered by in-group and out-group individuals differs in the rate of subsequent avoidance of the action that led to the reinforcement, this class of models should provide the best fit to the data. For the sake of completion, we also considered hybrid models incorporating both outcome-value and learning-rate parameters (see the Results section).

Method

Participants

Thirty volunteers of European decent (20 female, 10 male) were recruited and provided written informed consent to participation in the experiment. The participants received two movie vouchers in exchange for their participation.

Stimuli

Eighty-four faces were selected from the NimStim Set of Facial Expressions (Tottenham et al., 2009) and the Center for Vital Longevity Face Database (Minear & Park, 2004). The stimuli selection balanced race (42 Black and 42 White), emotional expression (48 neutral, 20 happy, and 20 angry), and gender (42 male and 42 female). The images were transformed to gray scale and cropped to the same size. All stimuli were presented in frontal view and at a standardized viewing size (all images were scaled to a unitary height of 556 pixels) against a uniform white background. The choice stimuli were a set of 50 abstract fractal images.

Task and procedure

Participants performed a probabilistic two-choice learning task (see Fig. 1 for overview and design details). We instructed the participants (by presenting a 1-s text message prior to each of the 16 blocks of the experiment) to choose which of two abstract fractal stimuli led to the least exposure ("avoid") to the emotional face (angry or happy, depending on the block). The rational for contrasting angry and happy faces was to assess the specificity of the predicted interaction of racial bias, emotion, and race. Each fractal differed in terms of the probability that it was followed by an emotional face (instrumentally optimal choice = .3; instrumentally suboptimal choice = .7). The presentation of the neutral or emotional face, subsequent to the participant's choice, served as the only feedback on the choice made. By repeatedly choosing between these two fractals, the participants learned across trials which choice was optimal in each block. No instructions regarding the racial-group belonging of the stimuli were provided. The gender of facial stimuli was randomized for each block. Because the racial group of the facial stimuli was irrelevant to the instructed goal of the task, this design allowed us to probe the indirect effects of emotion and race in interaction on instrumental learning. After the experiment, we had the participants complete a validated Swedish version of the MRS to assess their explicit negative attitudes toward non-European immigrants in Sweden (Akrami, Ekehammar, & Araya, 2000) and the race IAT (Greenwald & Banaji, 1995) to measure their implicit racial bias.

Results

In this section, we first present the empirical results, which showed that individual differences in racial bias

critically influenced social RL. Second, we address the computational basis of this influence by testing, through model comparison, two competing hypotheses about the underlying learning mechanisms. Finally, we demonstrate a predicted relationship between the estimated model parameters and individual differences in racial bias.

Statistical analysis

The data were analyzed using logistic generalized linear mixed models with by-subject random intercept and slopes for the fixed effects (Baayen, Davidson, & Bates, 2008). Hierarchical logistic regression is the preferred statistical method for repeated measures of a binary dependent variable (optimal = 1, suboptimal = 0; Jaeger, 2008). The generalized linear mixed model approach allowed us to directly estimate the trial-by-trial probability of choosing the instrumentally optimal choice while controlling for between-subjects variability. Reported main and interaction effects are based on model comparison using the Wald test (Fox & Weisberg, 2011).

We addressed our hypothesis that racial bias would specifically influence learning to avoid threatening out-group members by formulating a regression model. The model included the terms Racial Group × Emotion × MRS and Racial Group × Emotion × IAT; that is, all main effects and interactions except the IAT × MRS interaction (i.e., similar to a factorial analysis of covariance). This analysis strategy allowed us to draw conclusions about the separate effects of explicit and implicit racial bias while controlling for overlapping variance between the two (for additional analyses, see Figs. S1–S3 and the Two-Way Interaction and Response Times sections in Statistical Analyses in the Supplemental Material).

As predicted, the analysis revealed a three-way Emotion × Racial Group × MRS interaction, $\chi^2(1) = 19.88$,[1] $p < .001$, which showed that the effect of out-group faces on behavior was markedly different for individuals with higher, relative to lower, explicit racial bias (see Fig. 2). Individuals with low explicit racial bias performed best when reinforced by happy out-group faces, whereas individuals with high explicit racial bias performed best when reinforced by threatening out-group faces. Simple effects showed that explicit racial bias modulated behavior only when reinforced by out-group faces (see the Simple Effects section in Statistical Analyses in the Supplemental Material for an explication of the interaction).

Furthermore, the regression analysis showed a significant Racial Group × IAT interaction, $\chi^2(1) = 6.82$, $p = .009$ (see Fig. S2 in the Supplemental Material). This interaction showed that implicit racial bias was negatively related to the probability of avoiding in-group faces, $\beta = -1.02$, $SE = 0.08$, and that this effect was attenuated for out-group faces—simple interaction: $\beta = 0.86$, $SE = 0.27$. The IAT and MRS scores were not significantly correlated across participants, $r(28) = .3$, $p = .12$.

Taken together, these analyses showed, as predicted, that individual differences in racial bias strongly modulated the effect of social reinforcement and racial-group belonging on behavior. Explicit racial bias interacted with both the emotion of the facial expressions and their racial-group belonging, which resulted in better avoidance of threatening out-group faces in proportion to the level of explicit racial bias. Implicit racial bias was related to better avoidance of both threatening and friendly racial-out-group faces relative to same-race faces.

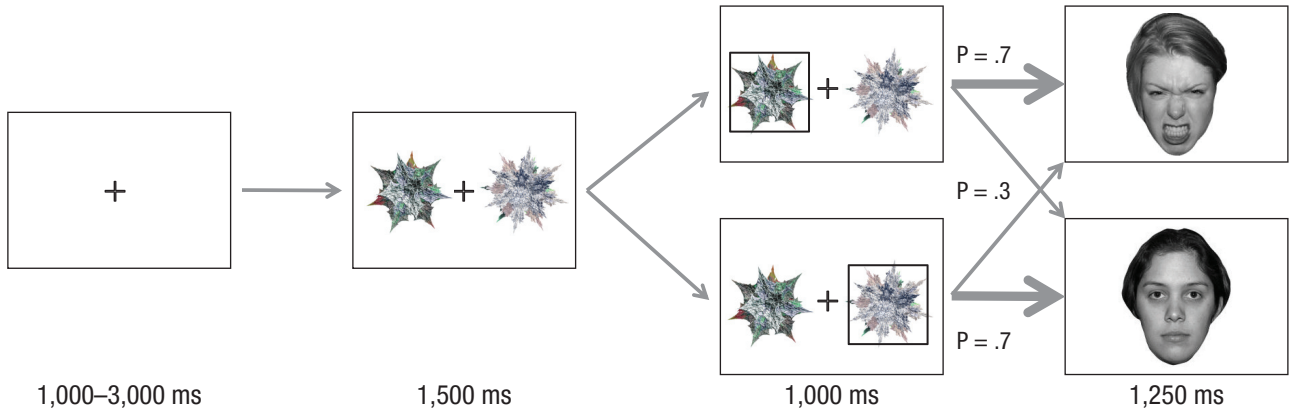

Fig. 1. Schematic illustrating the experimental reinforcement-learning task. The participants' task was to choose between pairs of stimuli that were probabilistically followed by neutral or emotional faces and to learn by trial and error which choice led to the least exposure to the emotional face (friendly or threatening, depending on block). The location of the choice stimuli varied randomly between the left and right positions across trials to prevent spatial-choice strategies. The task had a 2 (racial group: in-group vs. out-group) × 2 (emotion: friendly vs. threatening) factorial design in which every combination of factors was repeated for four blocks, each of which comprised 30 trials. The figure illustrates the optimal (right choice) and suboptimal (left choice) performance in an "avoid-angry" block. Stimuli are not drawn to scale. P = probability.

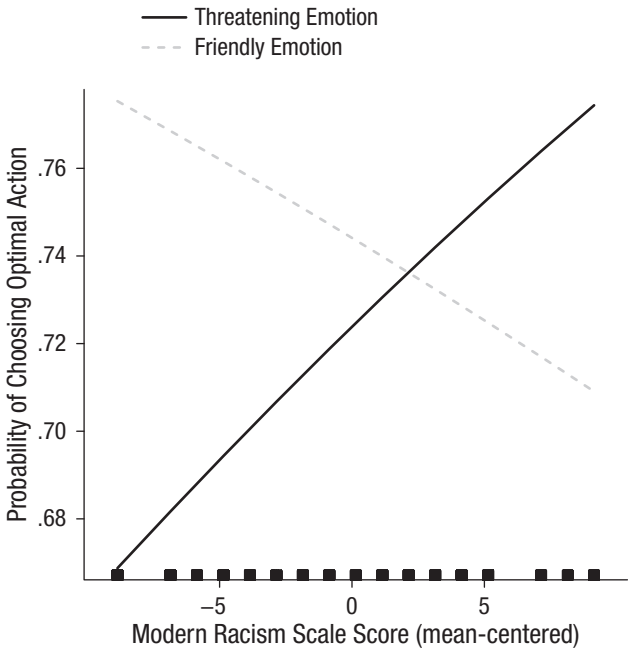

Fig. 2. Estimated probability of choosing the optimal action when avoiding friendly or threatening out-group faces as a function of individual differences in explicit racial bias (as measured by the Modern Racism Scale) and the emotion of the reinforcing facial expression. The one-dimensional "scatter plot" on the *x*-axis depicts the approximate distribution of mean-centered Modern Racism Scale scores. Note that several participants could have the same score.

Computational model analysis

Model selection. We used a standard Q-learning model to analyze task performance. Similar models have previously been used to decompose performance into different latent components (e.g., Frank, Moustafa, Haughey, Curran, & Hutchison, 2007). The model states that the expected value $Q_i(t+1)$ of the chosen action i at trial $t+1$ is the sum of the previous value of the action $Q_i(t)$ and the prediction error $\delta(t)$ multiplied with the learning rate α:

$$Q_i(t+1) = Q_i(t) + \alpha \times \delta(t).$$

The prediction error is the difference between the expected outcome value of the action, $Q_i(t)$, and its actual outcome value, $R(t)$:

$$\delta(t) = R(t) - Q_i(t).$$

We implemented the two competing hypotheses (OE vs. OL) by generalizing the basic Q-learning model to include separate outcome-value or learning-rate parameters for the different experimental conditions (see Fig. 3 and the Model Specification and Model Description sections in Computational Modeling in the Supplemental Material). Several formulations of the OE and OL hypotheses, varying in complexity, were compared (see the Parameter Fitting section in Computational Modeling in the Supplemental Material) with the Akaike information criterion (AIC), which punishes model complexity with a penalty term (because adding complexity to a model often improves model fit; Daw, 2011). The results of the model comparison are presented in Figure 3 (see also Computational Modeling and Table S1 in the Supplemental Material).

The winning model (OL2) included six learning-rate (α) parameters. Of the six parameters, four regulated how much the Q value for each action should be updated after exposure to the incorrect emotional outcome for each experimental condition (friendly in-group, friendly out-group, threatening in-group, and threatening out-group) and two regulated Q-value updating on the basis of the correct neutral racial-out-group and racial-in-group outcomes, respectively. This model thereby delimits the learning driven by incorrect outcomes (i.e., exposure to the emotional facial expressions) from the learning driven by correct outcomes (i.e., exposure to the neutral facial expressions).

Most important, the superior fit of the OL2 model relative to the OE models, which estimated the outcome value of the social reinforcement but held the learning rate constant across conditions, provided strong support for the OL hypothesis relative to the OE hypothesis. This result indicated that the behavioral results were driven by an OL process rather than by an OE process (Rangel et al., 2008). This conclusion was further corroborated by the better fit of the pure learning-rate model (OL2), compared with the hybrid models, which included parameters for both learning rate and outcome value.

Model parameters and individual differences in racial bias. To relate the winning model (OL2) to individual differences in explicit and implicit racial bias, we correlated the parameter estimates with the MRS and IAT scores across participants. As predicted, both the MRS scores, $r(28) = .44$, $p = .015$, and the IAT scores, $r(28) = .45$, $p = .012$, were related to only one parameter estimate—the learning rate for threatening other-race faces (see Table S2 in the Supplemental Material for all correlations).

In summary, the model comparison provided strong support for the OL hypothesis relative to the OE hypothesis. Thus, the different types of social reinforcement were similarly valued but differed in the rate at which they were transformed into future avoidance behaviors. Racial bias was specifically related to the estimated learning rate associated with threatening other-race faces, which suggests a computational mechanism for how

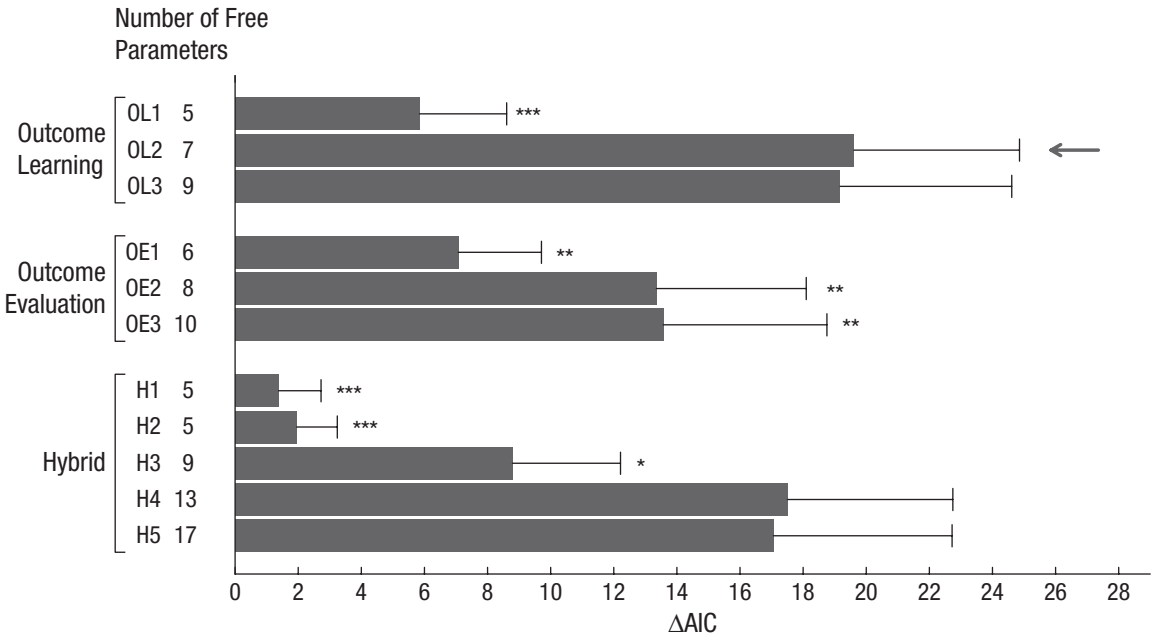

Fig. 3. Model comparison: the statistical fit of the candidate reinforcement-learning models expressed as the difference in Akaike information criterion (AIC) from a simple baseline model ($\Delta = AIC_{baseline} - AIC_{candidate\ model}$), which had the same fixed parameters across conditions (i.e., did not take the experimental design into account). The OL2 model, indicated by the arrow, had higher ΔAIC and fewer free parameters than did the competing models and, thus, provided the best and most parsimonious explanation of the data. The most successful alternative models (OL3, H4, and H5) had 2 to 10 more free parameters than did the OL2 model but still failed to explain the data significantly better (see Computational Modeling and Table S1 in the Supplemental Material for full model specifications). The average ΔAIC between the OL2 model and the OE models was 8.3, which indicated strong support for the OL hypothesis over the OE hypothesis (Burnham & Anderson, 2002). All models included a free parameter (β) that regulates how deterministically Q-value differences are translated into choices (see the Q-Learning and the Softmax Function section in Computational Modeling in the Supplemental Material). Error bars indicate +1 *SD*. Asterisks denote significant pairwise AIC differences relative to the OL2 model (*$p < .05$; **$p < .01$; ***$p < .001$). OL = outcome learning; OE = outcome evaluation; H = hybrid.

individual differences in racial bias affect learning from social reinforcement delivered by racial-out-group individuals (see the Statistical Analysis section).

Discussion

The research presented here demonstrates that individual differences in racial bias can modulate basic aspects of social RL; more precisely, our findings show how threatening and friendly in- and out-group faces shape future behaviors. Furthermore, findings obtained by testing two contrasting hypotheses about the computations mediating these social RL processes illustrate a possible mechanism of social learning. Taken together, these results shed new light on the learning processes that underlie social interaction with racial-in-group and racial-out-group individuals.

Racial bias can theoretically be separated into two different but related constructs—explicit bias and implicit bias (Greenwald & Banaji, 1995). Both types of racial bias proved to be crucial modulators of social RL; the explicit (MRS) and implicit (IAT) measures of racial bias independently modulated the effect of social reinforcement and racial-group belonging on behavior when incorporated into the same regression analysis (i.e., when we controlled for the other measure). Explicit racial bias strongly, and uniquely, modulated the impact of out-group faces on behavior, such that the higher the explicit racial bias, the better the learned avoidance of threatening out-group individuals (see Fig. 2). The reverse pattern was also true; individuals with low MRS scores showed the best instrumental learning in the out-group conditions when reinforced by friendly faces. Previous research has shown that individuals low in explicit racial bias, compared with individuals high in explicit racial bias, recruit more effortful processing in response to angry out-group faces relative to happy out-group faces (Chiu, Ambady, & Deldin, 2004). This effortful processing is thought to indicate downregulation of automatic negative attitudes motivated by egalitarian values. We speculate that if individuals with low MRS scores in the present experiment similarly recruited more effortful processing to downregulate automatic

negative attitudes during the "angry out-group" blocks, this could have resulted in fewer resources devoted to learning the actual choice-outcome contingencies and, thus, worse instrumental learning.

An alternative, but mutually nonexclusive, explanation is the influence of a conformation-bias-induced salience effect. Accordingly, individuals low in explicit racial bias would be most attentive to faces that confirm their expectations about the out-group (i.e., friendly out-group faces that would support superior learning). Conversely, threatening out-group faces might have been more salient to participants high in explicit racial bias, thus promoting better learning in this group. Our data do not allow us to differentiate between these and other possible explanations.

Implicit racial bias was related to better learned avoidance of both threatening and friendly racial-out-group faces relative to racial-in-group faces. However, explicit and implicit racial bias also shared common variance. Across participants, the two measures showed a moderate, but nonsignificant, correlation (r = .3), similar to what is typically reported (Nosek, 2007). This result suggests that individual differences in explicit and implicit racial bias function as separable, but partially overlapping, modulators of social RL.

Previous findings have shown more persistent fear conditioning in response to racial-out-group faces, compared with racial-in-group faces (Navarrete et al., 2009; Olsson et al., 2005), and to angry in-group faces, compared with neutral in-group faces (Öhman & Mineka, 2001). The present results provide an important extension of previous research by showing that reinforcing facial expressions and markers of racial-group belonging interact in their impact on instrumental learning as a function of individual differences in racial bias (see Fig. 2). In contrast to our finding of a critical modulatory impact on learning by both explicit and implicit racial bias, Olsson et al. (2005) did not find that racial biases were related to conditioning to racial-out-group faces. This difference in sensitivity to individual differences in racial bias might hint at important differences in the mechanisms that underlie classical conditioning to, relative to instrumental RL from, out-group facial expressions.

Race is a salient, but nonexclusive, marker of group membership. Our results do not allow us to conclusively determine whether the observed learning bias generalizes to all types of out-groups or whether it is specific to Black faces shown to White participants. However, several studies using fear conditioning have documented learning biases related to non-Black out-group faces (Mallan, Sax, & Lipp, 2009; Navarrete et al., 2012), which suggests that learning biases might be a general characteristic of interactions with salient out-groups. Notably, results from a recent study, which used a "minimal-group" paradigm that induced group membership merely by the color of the T-shirts worn by participants and the depicted in- and out-group members, showed facilitated fear learning to out-group faces (Navarrete et al., 2012).

By applying an RL-model-based analysis, we distinguished between two competing hypotheses, OE and OL, about the computational processes underlying social reinforcement (Lin, Adolphs, & Rangel, 2011; Rangel et al., 2008). The model comparison provided strong support for the OL hypothesis (see Fig. 3). We found that social reinforcement and racial-group belonging together affected how rapidly unexpected outcomes were transformed into future behaviors. Individual differences in both explicit and implicit racial bias were positively and selectively correlated with the learning rate associated with threatening racial-out-group faces. These correlations indicated that compared with individuals with lower racial bias, individuals with higher racial bias more rapidly updated the value of each action to adjust their subsequent behavior when exposed to threatening racial-out-group faces. As shown by the behavioral results (see Fig. 2), this rapid value updating was functional: Individuals high in explicit racial bias exhibited enhanced performance when their actions were reinforced by threatening out-group faces. Taken together, these results suggest a computational mechanism for how racial bias affects social reinforcement that depends on the racial belonging of the reinforcing individual.

Previous research has shown that when presented to White participants, unknown Black faces tend to recruit brain areas linked to automatic valuation processes, such as the amygdala (Kubota, Banaji, & Phelps, 2012). If this automatic difference in evaluating in-group and out-group faces significantly contributes to social RL, one would expect the computational modeling analyses to support the OE hypothesis, which was not the case. Our findings suggest that such automatic valuation processes do not directly translate into choice behavior. Instead, higher-level cognitive processes sensitive to individual differences in racial bias might mediate the mapping from social reinforcement to action by adjusting the impact that exposure to threatening racial-out-group individuals will have on subsequent behavior. Providing a possible neural correlate of the present findings, recent research has shown that connectivity between the striatum, commonly implicated in the computation of prediction errors, and the medial frontal cortex correlate with estimates of learning rate (van den Bos, Cohen, Kahnt, & Crone, 2012).

To conclude, the present results demonstrate that individual differences in racial bias play a key role in social RL: Higher racial bias was associated with better avoidance of racial-out-group faces. It is important to note that individuals high in racial bias updated the value of

actions reinforced by threatening out-group individuals more rapidly than did individuals low in racial bias. These findings have implications for our understanding of learning from social interaction and of how this learning is modulated by the racial-group belonging of the people with whom we interacts.

Author Contributions

B. Lindström developed the study concept. All authors contributed to the study design. B. Lindström and I. Selbing analyzed and interpreted the data under the supervision of A. Olsson. B. Lindström and A. Olsson drafted the manuscript, and I. Selbing and T. Molapour critically revised the manuscript. All authors approved the final version of the manuscript for submission

Acknowledgments

We gratefully thank the three anonymous reviewers for their valuable suggestions, Nazar Akrami and Robin Bergh for providing the measures of racial bias, and Kristian Ekeroth for assistance with data collection.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This research was supported by a grant to A. Olsson from the Swedish Science Council (Vetenskapsrådet; 421-2010-2084).

Supplemental Material

Additional supporting information may be found at http://pss.sagepub.com/content/by/supplemental-data

Note

1. The degrees of freedom refer to the difference in the number of parameters between the models compared by the Wald tests.

References

Ackerman, J. M., Shapiro, J. R., Neuberg, S. L., Kenrick, D. T., Becker, D. V., Griskevicius, V., . . . Schaller, M. (2006). They all look the same to me (unless they're angry): From out-group homogeneity to out-group heterogeneity. *Psychological Science*, *17*, 836–840.

Akrami, N., Ekehammar, B., & Araya, T. (2000). Classical and modern racial prejudice: A study of attitudes toward immigrants in Sweden. *European Journal of Social Psychology*, *30*, 521–532.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes. Evidence of race evaluations from ages 6 and 10 and adulthood. *Psychological Science*, *17*, 53–58.

Blair, R. J. R. (2003). Facial expressions, their communicatory functions and neuro-cognitive substrates. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*, 561–572.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York, NY: Springer.

Chiu, P., Ambady, N., & Deldin, P. (2004). Contingent negative variation to emotional in- and out-group stimuli differentiates high- and low-prejudiced individuals. *Journal of Cognitive Neuroscience*, *16*, 1830–1839.

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In M. R. Delgado, E. Phelps, & T. W. Robbins (Eds.), *Decision making, affect, and learning: Attention and Performance XXIII* (pp. 1–26). New York: Oxford University Press.

Dayan, P., & Balleine, B. W. (2002). Reward, motivation, and reinforcement learning. *Neuron*, *36*, 285–298.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression*. Thousand Oaks, CA: Sage.

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences, USA*, *104*, 16311–16316.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences, USA*, *108*, 15647–15654.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4–27.

Heerey, E. A. (2013). Learning from social rewards predicts individual differences in self-reported social ability. *Journal of Experimental Psychology: General*. Advance online publication. doi:10.1037/a0031511

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, *10*, 1625–1633.

Katahira, K., Fujimura, T., Okanoya, K., & Okada, M. (2011). Decision-making based on emotional images. *Frontiers in Psychology*, *2*, Article 311. Retrieved from http://www.frontiersin.org/Journal/10.3389/fpsyg.2011.00311/full

Kubota, J. T., Banaji, M. R., & Phelps, E. A. (2012). The neuroscience of race. *Nature Neuroscience*, *15*, 940–948.

Lin, A., Adolphs, R., & Rangel, A. (2011). Social and monetary reward learning engage overlapping neural substrates. *Social Cognitive and Affective Neuroscience*, *7*, 274–281.

Mallan, K. M., Sax, J., & Lipp, O. V. (2009). Verbal instruction abolishes fear conditioned to racial out-group faces. *Journal of Experimental Social Psychology*, *45*, 1303–1307.

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–125). San Diego, CA: Academic Press.

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, *36*, 630–633.

Navarrete, C. D., McDonald, M. M., Asher, B. D., Kerr, N. L., Yokota, K., Olsson, A., & Sidanius, J. (2012). Fear is readily associated with an out-group face in a minimal group context. *Evolution and Human Behavior*, *33*, 590–593.

Navarrete, C. D., Olsson, A., Ho, A. K., Mendes, W. B., Thomsen, L., & Sidanius, J. (2009). Fear extinction to an out-group face: The role of target gender. *Psychological Science*, *20*, 155–158.

Nosek, B. A. (2007). Implicit-explicit relations. *Current Directions in Psychological Science*, *16*, 65–69.

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, *108*, 483–522.

Olsson, A., Ebert, J. P., Banaji, M. R., & Phelps, E. A. (2005). The role of social groups in the persistence of learned fear. *Science*, *309*, 785–787.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*, 545–556.

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences, USA*, *108*, 7710–7715.

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., . . . Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, *168*, 242–249.

van den Bos, W., Cohen, M. X., Kahnt, T., & Crone, E. A. (2012). Striatum–medial prefrontal cortex connectivity predicts developmental changes in reinforcement learning. *Cerebral Cortex*, *22*, 1247–1255.