Racial Bias Drives Social Reinforcement Learning

Björn R. Lindström1,2, Ida Selbing1,2, Tanaz Molapour1,2, Andreas Olsson1,2
1 Department of Clinical Neuroscience, Karolinska Institutet, Sweden 2 Stockholm Brain Institute

Introduction

• Emotional facial expressions function as reinforcers in social interaction, and can affect future approach or avoidance decisions [1].
• Facial markers of racial group-belonging affect memory and behavior [2–3].
• It is unknown how individual differences in explicit and implicit racial bias affect how we learn from the emotional facial expressions of others.
• We asked how individual differences in racial bias affects social reinforcement learning (RL) from social reinforcement: friendly or threatening facial expressions posed by racial in- or out-group individuals.

• We used computational modeling to analyze which learning process was affected by social reinforcement: outcome evaluation (OE) or outcome learning (OL)? [4]

Methods

• Thirty European subjects (20 women)
• Probabilistic two-choice decision making paradigm with NimStim faces as reinforcement.
• Subjects learned by trial-and-error to avoid the choice with highest probability (P = .7) of being reinforced by an emotional face in each block: “Avoid Angry” or “Avoid Happy”
• 2 (Racial Group: In/Out) * 2 (Emotion: Friendly/Threatening) design
• Every combination was repeated for four blocks, each with 30 trials.
• Race Implicit Association Test (IAT; implicit bias) and Modern Racism Scale (MRS; explicit bias)

Reinforcement Learning Models

We used modifications of the Q-learning algorithm to model trial-by-trial behavior.

We differentiated between two hypothesis about the computational mechanisms underlying social RL:

1. Social reinforcement affects behavior through differences in outcome evaluation (OE – hypothesis)
2. Social reinforcement affects behavior through differences in learning from outcomes (OL – hypothesis)

The OE - hypothesis was modeled by fitting different reinforcement (R) parameters for the different conditions:

\[ Q_{ole}(t+1) = Q_{ole}(t) + \alpha \cdot \varepsilon(t) \]

\[ \varepsilon(t) = R^{EMOTION/RACIAL GROUP}(t) - Q_{ole}(t) \]

The OL - hypothesis was modeled by fitting different learning rate parameters (a) for the different conditions:

\[ Q_{ole}(t+1) = Q_{ole}(t) + \alpha^{EMOTION/RACIAL GROUP} \cdot \varepsilon(t) \]

\[ \varepsilon(t) = R(t) - Q_{ole}(t) \]

The MRS [r(28) = .44, p = .01] and IAT(r(28) = .45, p = .012) was selectively correlated with the learning rate of Threatening Out-group faces in the winning model (OL2).

The IAT and MRS scores were not significantly correlated in the sample.

Results

Modern Racism Scale (mean-centered)

Race IAT (mean-centered)

Computational Model Comparison

• We compared the goodness-of-fit of several learning models against a simple baseline model using the Akaike Information Criterion (AIC) which punishes model complexity (larger difference indicates better fit).

• The winning model, OL2 (indicated by an arrow), was an implementation of the OL-hypothesis.

• Model comparison gave strong support for the OL-hypothesis. The group belonging and emotional expression of the social reinforcer affects learning from outcomes rather than the evaluation of outcomes.

Conclusions

• Individual differences in racial bias strongly modulate basic aspects of social reinforcement learning: how emotional facial expressions affects future behavior.

• Higher racial bias was associated with better avoidance of racial out-group faces.

• Computational modeling showed that social reinforcements primarily affects the rate with which social reinforcements were transformed into future actions, rather than directly modulate the value of the outcomes.

• Individual differences in racial bias are linked to these underlying computations: high racial bias subjects learned most rapidly to avoid threatening out-group members.

References