Racial Bias Drives Social Reinforcement Learning

www.emotionlab.se

Björn R. Lindström[1,2], Ida Selbing[1,2], Tanaz Molapour[1,2], Andreas Olsson[1,2]

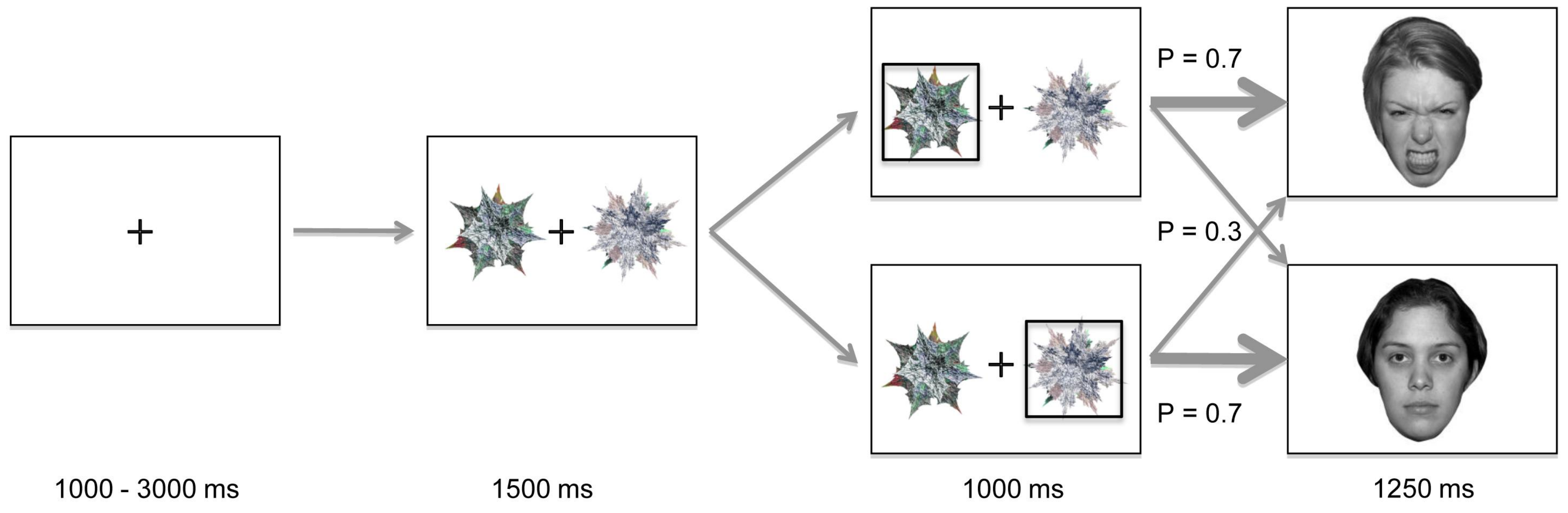
[1] Department of Clinical Neuroscience, Karolinska Institutet, Sweden [2] Stockholm Brain Institute

Introduction

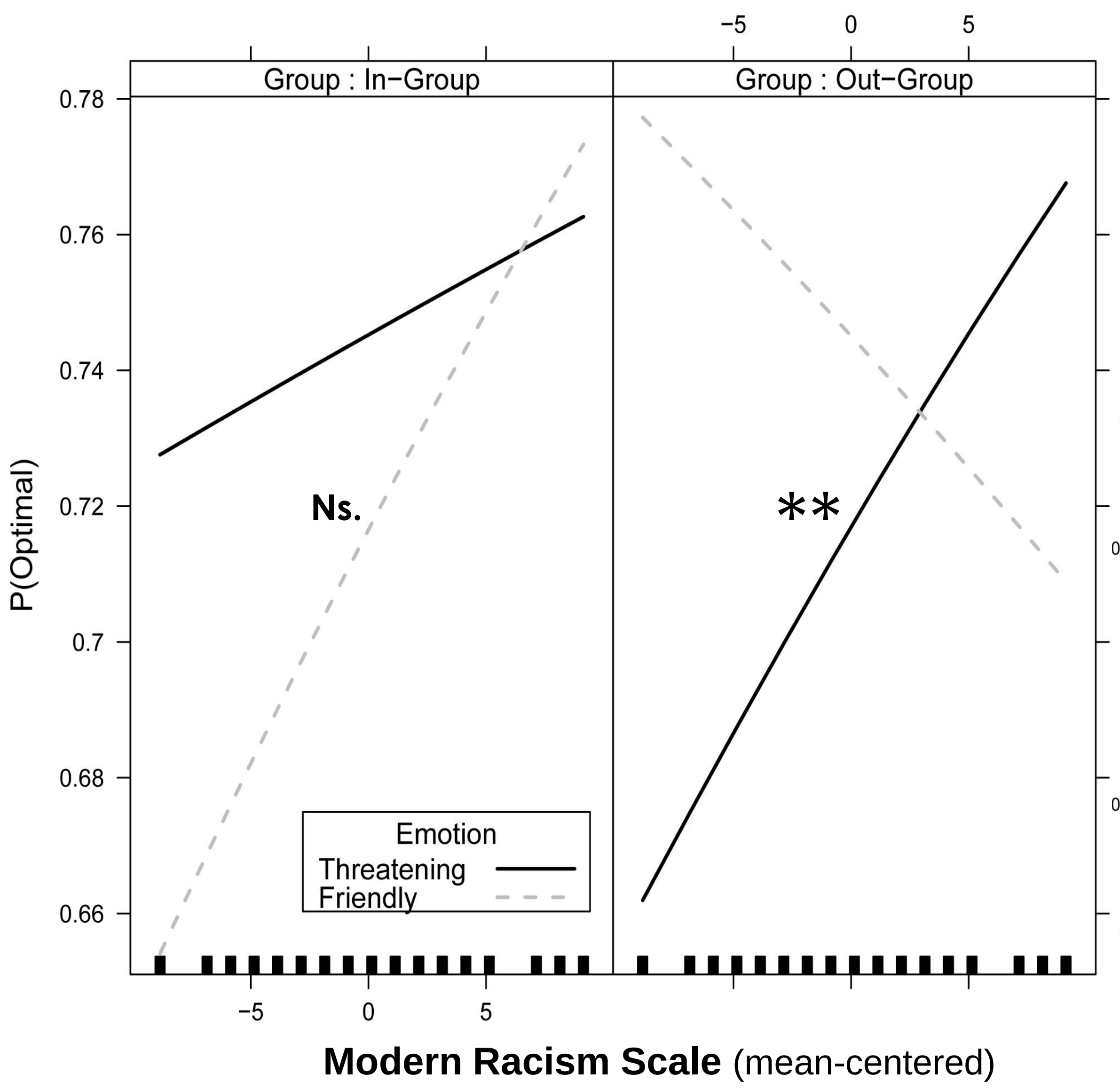
- **Emotional facial expressions** function as **reinforcers** in social interaction, and can affect future approach or avoidance decisions [1]
- Facial markers of **racial group-belonging** affect memory and behavior [2-3]
- It is unknown how individual differences in explicit and implicit **racial 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)

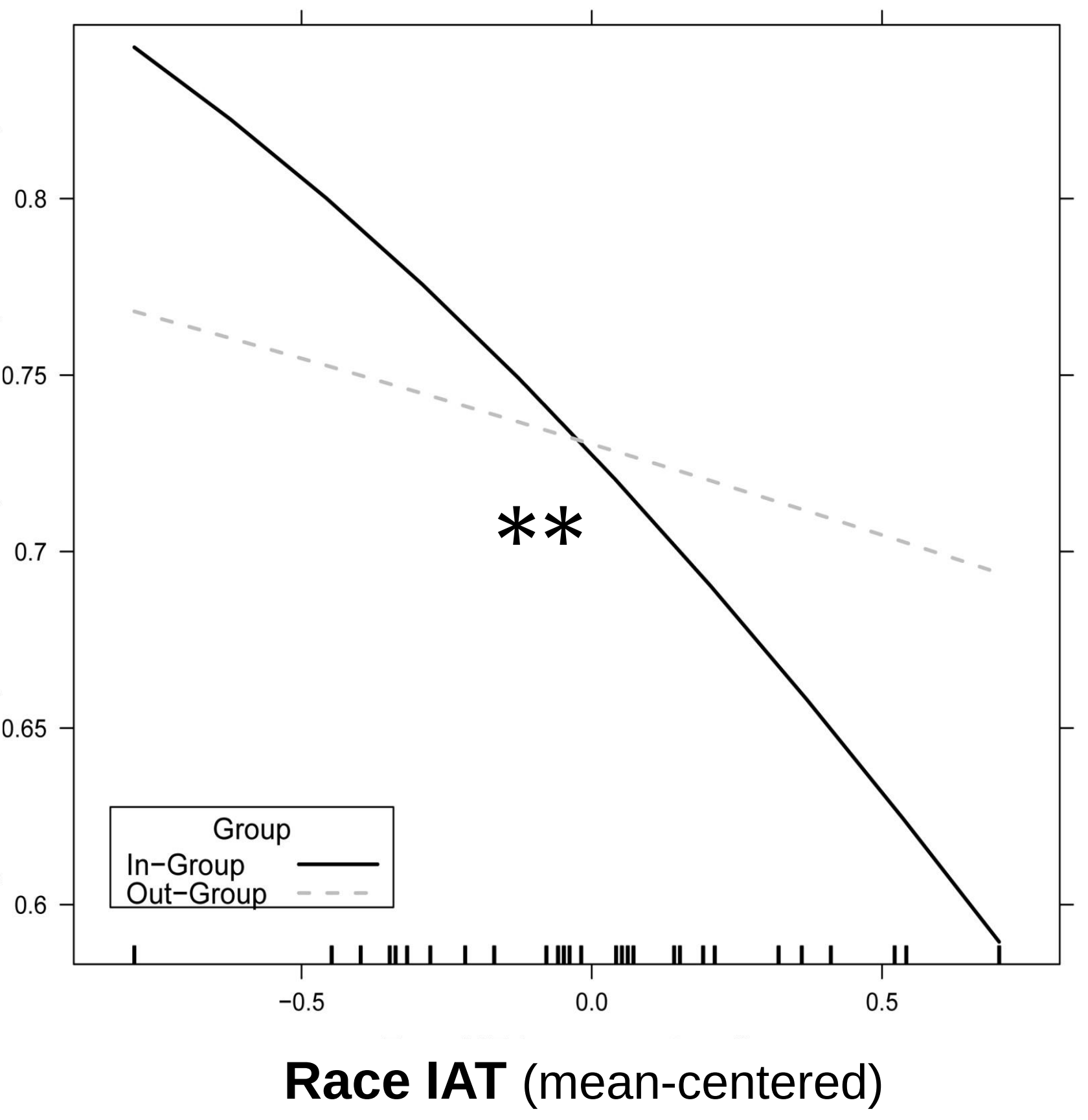


Results



The probability of choosing the optimal action when avoiding friendly or threatening out-group faces was affected by individual differences in MRS in interaction with the emotion of the reinforcing facial expression.

** = p< .01



The probability of choosing the optimal action when avoiding in- or out-group faces was affected by individual differences in IAT , but not in interaction with Emotion.

** = p< .01 .

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 :

I. Social reinforcement affects behavior through differences in *outcome evaluation* (OE – hypothesis)
II. 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:

$$Q_A(t+1) = Q_A(t) + a * \delta(t)$$

$$\delta(t) = R_{EMOTION/RACIAL\ GROUP}(t) - Q_A(t)$$

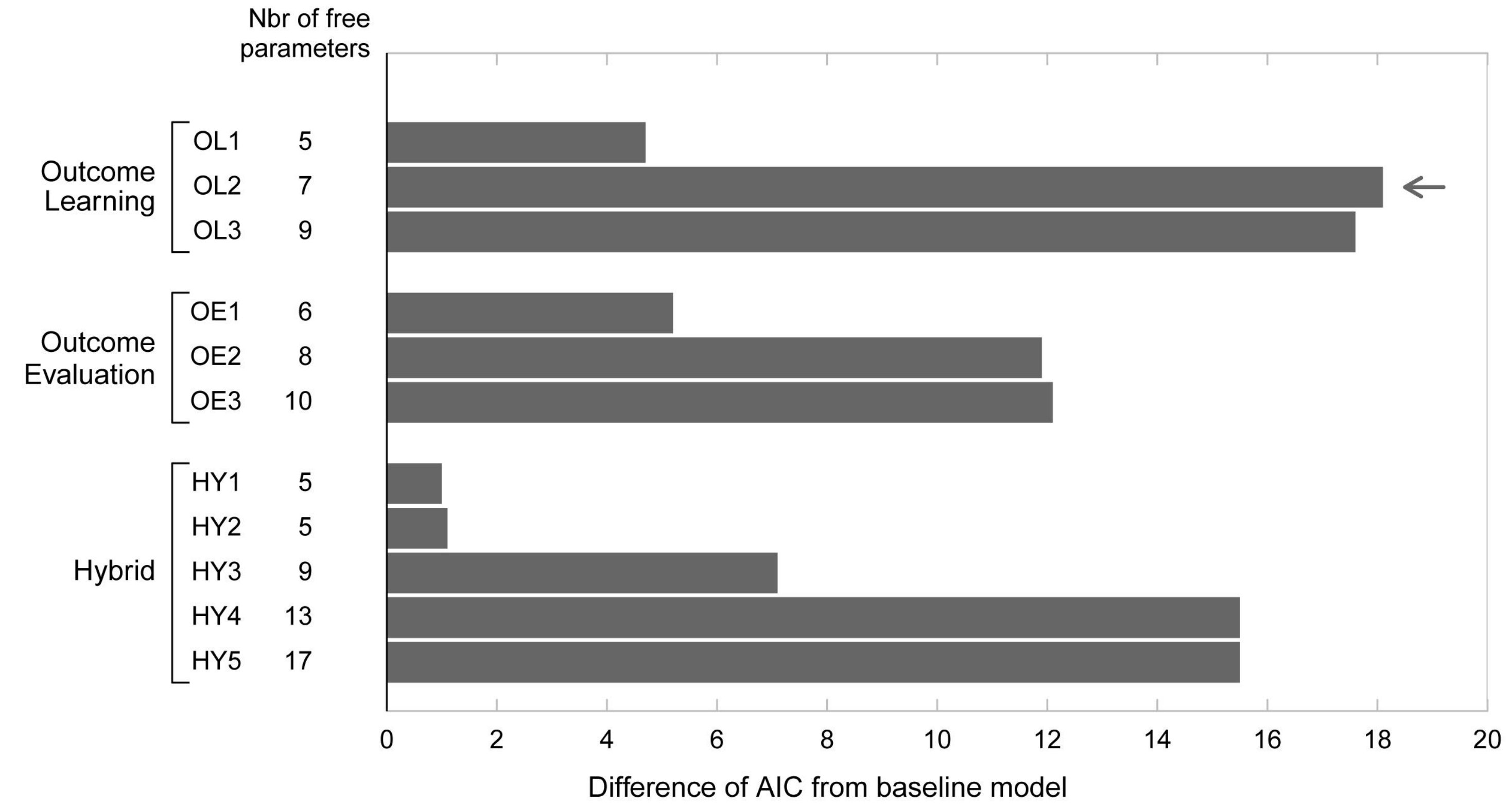
The OL - hypothesis was modeled by fitting different learning rate parameters (a) for the different conditions.

$$Q_A(t+1) = Q_A(t) + a_{EMOTION/RACIAL\ GROUP} * \delta(t)$$

$$\delta(t) = R(t) - Q_A(t)$$

The MRS (*r*(28) = .44, *p* = .015) 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.

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

1. 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*(1), 53-8.

2. Blair, R. J. R. (2003). Facial expressions, their communicatory functions and neuro-cognitive substrates. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *358*(1431), 561-72.

3. 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*(5735), *785-7.*

4. Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature reviews. Neuroscience*, 9(7), 545-56.