Neural signals for the detection of unintended race bias

David M. Amodio, Eddie Harmon-Jones, Patricia G. Devine, John J. Curtin, Sigan L. Hartley, and Alison E. Covert

University of Wisconsin – Madison

Introduction

Stereotypes of Blacks are so deeply imbedded in American culture that they may be activated automatically (Devine, 1989). Once activated, racial stereotypes can lead to unintentional discriminatory behaviors (Dovidio, Kawakami, & Gaertner, 2002). Indeed, many self-avowed egalitarians report that prejudices often slip through in their behavior, despite their non-prejudiced intentions (Devine, Monteith, Zuwerink, & Elliot, 1991; Monteith, 1993). Although the conditions precluding control have been studied, previous research has not examined the *process* underlying failures to control expressions of prejudice.

Research Question:

- Why does prejudice control sometimes fail?
  - Has the mind not detected that race bias is present?
  - Is the mind aware of the bias, but unable to inhibit prejudiced behavior?

To address these questions, we applied a neural model of cognitive control to the context of race bias:

Dual system model of control

(Botvinick, Braver, Carter, Barch, & Cohen, 2001)

- Evaluation system
  - Continuously monitors ongoing neural activity for conflict between behavioral tendencies
    - Associated with activity in anterior cingulate cortex (ACC)
  - When conflict is detected, second system is signaled
- Regulatory system
  - Organizes behavior to resolve conflict
    - Associated with activity in prefrontal cortex

The dual-system model suggests two explanations for why prejudice control fails:

1) Conflict detection system not activated sufficiently
   - Conflict between automatic race bias and intention to respond without prejudice not detected
2) Regulatory system not activated sufficiently
   - Conflict is detected, but second system fails to regulate behavior

Present Study:

Is the conflict detection system sensitive to the potential for a race-biased response?

- Examined activity of conflict-detection process associated with participants' race-biased responses
- Conflict detection was measured using error-related negativity (ERN) component of the event-related potential

Method

Participants

- 34 White American students

Procedure

- Completed 288 trials of gun-tool task
- EEG: 27 scalp sites, average earlobe reference

ERN derivation

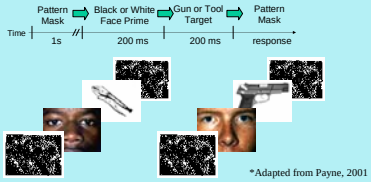
- 1-15 Hz signal at frontocentral midline (Fcz)
- Averaged across error responses within each trial

Error-related negativity (ERN)

- Negative polarity wave
- Occurs concurrent with error response
- Fronto-central scalp distribution
- ACC neural generator
- High temporal resolution

From W. Gehring: www.personal.umich.edu/~wgehring/lab/Learn.html

Gun-tool task*



*Adapted from Payne, 2001

- Participants categorized each target by pressing a computer keyboard button labeled "gun" or "tool"
- Responses were to be made within 500 ms of target
  - Increased error rate to facilitate examination of unintended race-biased responses
- Black face primes were designed to activate "violent" stereotype
  - Black face should facilitate "gun" responses and cause conflict for "tool" responses

Participant instructions:

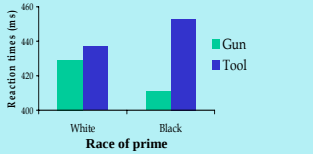
- "Task designed to measure racial prejudice"
- "Errors on certain trials attributed to race bias"
- Responding "gun" instead of "tool" after a Black face suggested influence of stereotype

★Errors on Black-tool trials critical★

Results

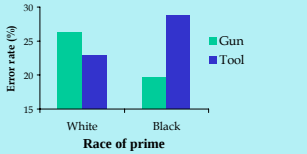
A) Gun-tool task created race-biased response conflict

- Black face primes facilitated "gun" responses and inhibited "tool" responses



Interaction: $F(1, 33) = 14.39$, $p < .001$

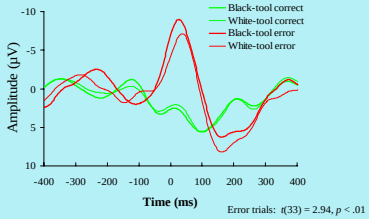
- After seeing a Black face, participants were most likely to make stereotype-consistent errors (e.g., press "gun" when target was "tool")



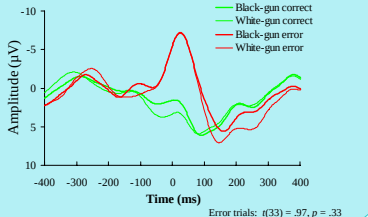
Interaction: $F(1, 33) = 10.84$, $p < .005$

B) Greater conflict detection for race-biased responses

- ERNs were largest for errors attributable to race bias (Black-tool errors), compared to ERNs for all other error types



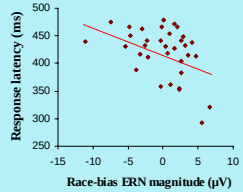
Error trials: $t(33) = 2.94$, $p < .01$



Error trials: $t(33) = .97$, $p = .33$

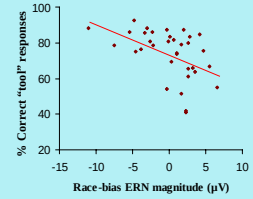
C. Race-biased ERNs predicted greater control

- To examine the behavioral effects of ERN amplitude associated with race-bias–detection, a "race-bias ERN" was computed, representing the ERN to Black-tool errors with White-tool errors covaried
  - Larger race-biased ERNs (negatively valenced) predicted *greater slowing of responses* following errors



$r(32) = -.44$, $p = .01$

  - Larger race-biased ERNs (negatively valenced) predicted *greater post-error accuracy* on "tool" trials but not "gun" trials



$r(32) = -.48$, $p < .005$

Discussion

- Unintentional race bias not due to lack of detection
  - Errors attributable to race bias were associated with larger ERNs than other errors
  - Unintentional race bias most likely associated with failure of PFC-related system to regulate behavior
- Recruitment of race-bias control begins very early in response stream and does not require awareness
  - Suggests revision of predominant models of mental correction, e.g., Wegener & Petty (1997), Wilson & Brekke (1994)
- Greater sensitivity to the potential for race-bias predicted more controlled behavior throughout task
  - Suggests individuals more sensitive to race-biased response conflict are more adept at regulating race-biased behaviors
  - Increasing one's regulatory ability may require enhancing one's implicit sensitivity to race-biased conflict detection