Research Article

Psychological Science
1–13
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09567976231215238
www.psychologicalscience.org/PS

Reducing Facial Stereotype Bias in Consequential Social Judgments: Intervention Success With White Male Faces

Youngki Hong, Kao-Wei Chua, and Jonathan B. Freeman
Department of Psychology, Columbia University

Abstract

Initial impressions of others based on facial appearances are often inaccurate yet can lead to dire outcomes. Across four studies, adult participants underwent a counterstereotype training to reduce their reliance on facial appearance in consequential social judgments of White male faces. In Studies 1 and 2, trustworthiness and sentencing judgments among control participants predicted whether real-world inmates were sentenced to death versus life in prison, but these relationships were diminished among trained participants. In Study 3, a sequential priming paradigm demonstrated that the training was able to abolish the relationship between even automatically and implicitly perceived trustworthiness and the inmates' life-or-death sentences. Study 4 extended these results to realistic decision-making, showing that training reduced the impact of facial trustworthiness on sentencing decisions even in the presence of decision-relevant information. Overall, our findings suggest that a counterstereotype intervention can mitigate the potentially harmful effects of relying on facial appearance in consequential social judgments.

Keywords

faces, stereotyping, trustworthiness, criminal sentencing, intervention

Received 8/4/23; Revision accepted 10/14/23

As we navigate the social world, we infer others' personality traits, such as trustworthiness, dominance, or intelligence, based solely on their facial appearances. These judgments are made quickly and often without conscious awareness (Bar et al., 2006; Willis & Todorov, 2006). Data-driven methods have identified sets of features that are most predictive of specific trait impressions (Oosterhof & Todorov, 2008). These features effectively serve as facial stereotypes and have been linked to severe downstream consequences. For example, faces that are perceived as untrustworthy are associated with more negative outcomes in hiring, political elections, and criminal sentencing (Olivola et al., 2014; Wilson & Rule, 2015, 2016). Although these face-based judgments are made very consistently across perceivers, they have little correspondence with actual personality or behavior (Krendl et al., 2014), suggesting that these facial stereotypes can inadvertently influence consequential social judgments.

Given the inaccuracy of facial stereotypes, it is important to explore approaches that can curb their real-world impact. However, face-based judgments have generally been impervious to interventions. Learning that appearances are nonpredictive of actual trustworthiness does not seem to decrease their use (Jaeger et al., 2019), nor does the knowledge of an individual's past behavior (Rezlescu et al., 2012). Even nudging participants not to use facial appearance (Jaeger et al., 2020) or providing more diagnostic social information has not been effective in reducing biased decision-making based on facial appearance (Todorov & Olson, 2008). These studies have either attempted to provide cues to participants that are more useful than facial appearance or they have explicitly educated participants to stop using facial appearance—an approach that has proved unsuccessful.

However, recent research attempting to operate on participants' underlying associations in a more automatic,

Corresponding Author:
Youngki Hong, Columbia University, Department of Psychology
Email: youngkih41@gmail.com

implicit manner has shown promise. Such studies have used a counterstereotype paradigm involving statistical learning: Untrustworthy-looking targets were paired with trustworthy behaviors, and trustworthy-looking targets were paired with untrustworthy behaviors. By countering the associations between facial features and traits, thereby rendering them unreliable, such training was effective in eliminating facial stereotypes associated with trustworthiness (Chua & Freeman, 2021). Tackling the implicit associations that drive these biases may be more effective than interventions relying on more explicit, deliberate processes (e.g., telling participants to stop judging faces). Indeed, the implications of using deliberate, propositional processes versus automatic, associative processes to induce changes in social biases have long been recognized (Gawronski & Bodenhausen, 2006).

Although this counterstereotype paradigm has shown promise, it is unclear whether it can mitigate consequential social judgments of real-world importance, such as life-or-death decisions in the context of criminal sentencing. Here we demonstrate that this paradigm can reduce or eliminate the relationship between facial trustworthiness and criminal-sentencing decisions in the context of real-world prisoners and real-world-like scenarios. Wilson and Rule (2015) found that prisoners in the Florida criminal justice system who were judged to have more untrustworthy-looking faces were more likely to be sentenced to death as opposed to life in prison. They also found that these biases were reflected in naive participants' hypothetical sentencing decisions (Wilson & Rule, 2016). In the current research, we test whether these real-world biases can be mitigated through counterstereotype training.

Study 1 establishes that the training is effective in diminishing the relationship between perceived trustworthiness and courtroom sentencing—that is, whether prison inmates are sentenced to life in prison or death—and that this bias reduction cannot be attributed to merely disengaging one's attention from people's faces. Study 2 extends the findings to sentencing-severity judgments, predicting that the training will reduce participants' biases in hypothetical sentencing judgments. Study 3 uses a sequential priming paradigm to test whether such training is successful also in automatic, implicit trustworthiness evaluations. Finally, Study 4 extends the training to real-world-like scenarios in which participants determine the guilt of a defendant in the presence of realistic decision-relevant information. See Figure 1 for the overview of the studies. Together, the present research shows that reconfiguring the associations between facial appearances and trustworthiness can reliably reduce people's reliance on facial appearance when making consequential social judgments.

Statement of Relevance

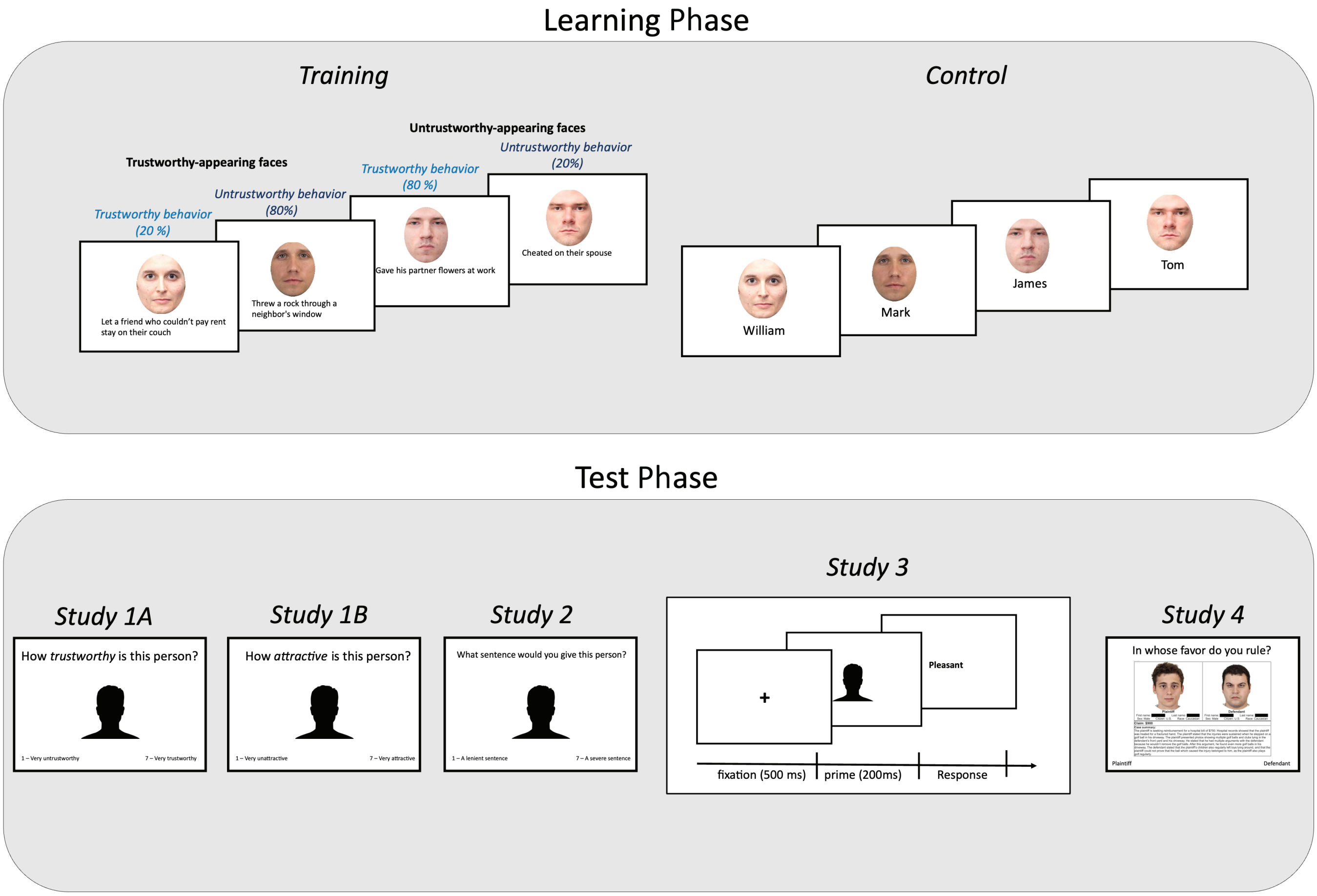
People spontaneously form initial impressions of others based solely on their facial appearance. These impressions, although often inaccurate, can lead to consequential social judgments, including life-or-death decisions like criminal sentencing. The prevailing view is that such biases stem from evolutionarily based approach-avoidance behavior, suggesting that they might be rigid and difficult to change. However, an associative learning paradigm, designed to sever the stereotypical links between specific facial appearances and perceived trustworthiness, was effective in reducing the reliance on facial appearance in social impressions and legal-sentencing decisions involving White male faces. This research highlights the promise of interventions that seek to mitigate the harmful effects of facial stereotypes on consequential social settings by dismantling the automatic associations between facial features and personality traits.

Open Practices Statement

None of the studies reported in this article were preregistered. All data, study materials, and analysis scripts are publicly available at https://osf.io/6m9jv/.

Study 1

Study 1 tested the effectiveness of a counterstereotype training (Chua & Freeman, 2021) in mitigating the relationship between the perceived trustworthiness of real-world prisoners' faces and whether those prisoners are sentenced to death or life in prison. Although the paradigm seeks to dismantle the associations between specific facial appearances (e.g., downward-turned lips) and personality traits (e.g., untrustworthiness) using countervailing exemplars, an alternative explanation could be that these exemplars violate participants' expectations about facial appearance and lead participants in the task to stop paying attention to the face stimuli altogether. To rule out this possibility, we tested the training's effect on the perceived attractiveness of each face. If the training merely diverts attention away from faces in the task, we should observe an extinction of attractiveness-related effects as well. However, if the training selectively affects the associations between facial appearances and trustworthiness (and not attractiveness), then its effects would be isolated to trustworthiness judgments.


Learning Phase
Training
Control
Trustworthy-appearing faces
Untrustworthy-appearing faces
Trustworthy behavior (20 %)
Untrustworthy behavior (80%)
Trustworthy behavior (80 %)
Untrustworthy behavior (20%)
Let a friend who couldn't pay rent stay on their couch
Threw a rock through a neighbor's window
Gave his partner flowers at work
Cheated on their spouse
William
Mark
James
Tom
Test Phase
Study 1A
How *trustworthy* is this person?
1 – Very untrustworthy
7 – Very trustworthy
Study 1B
How *attractive* is this person?
1 – Very unattractive
7 – Very attractive
Study 2
What sentence would you give this person?
1 – A lenient sentence
7 – A severe sentence
Study 3
+
Pleasant
fixation (500 ms)
prime (200ms)
Response
Study 4
In whose favor do you rule?
Plaintiff
First name: [redacted] Last name: [redacted]
Sex: Male Citizen: U.S. Race: Caucasian
Defendant
First name: [redacted] Last name: [redacted]
Sex: Male Citizen: U.S. Race: Caucasian
Claim: $900
Case summary:
Plaintiff
Defendant

The plaintiff is seeking reimbursement for a hospital bill of $750. Hospital records showed that the plaintiff was treated for a fractured hand. The plaintiff stated that the injuries were sustained when he stepped on a golf ball in his driveway. The plaintiff presented photos showing multiple golf balls and clubs lying in the defendant's front yard and his driveway. He stated that he had multiple arguments with the defendant because he wouldn't remove the golf balls. After this argument, he found even more golf balls in his driveway. The defendant stated that the plaintiff's children also regularly left toys lying around, and that the plaintiff could not prove that the ball which caused the injury belonged to him, as the plaintiff also plays golf regularly.

Fig. 1. Schematic overview of the current research.

Method

Participants. Our target sample size was 200 participants (~100 participants per group) and was based on a previous study that demonstrated the effectiveness of the training intervention (Chua & Freeman, 2021). Participants were recruited from Amazon Mechanical Turk (MTurk) and received monetary compensation for Study 1a (trustworthiness judgments) and Study 1b (attractiveness judgments). There were 226 total participants in Study 1a and 225 total participants in Study 1b. After excluding participants who failed attention checks, showed no variance in responses (e.g., chose 4 for all trials), and timed out on more than 50% of the trials during the learning phase (i.e., not paying attention), our final sample for Study 1a was 100 control participants and 106 trained participants (M_{age} = 39.30 years, SD_{age} = 10.51 years; 104 female, 102 male; 134 White, 24 Black, 24 Hispanic, 21 Asian, 3 Native American), and our final sample for Study 1b was 106 control participants and 107 trained participants (M_{age} = 39.25 years, SD_{age} = 10.50 years; 109 female, 104 male; 139 White, 31 Hispanic, 30 Black, 10 Asian, 3 Native American). Exclusions did not differ by training condition in either study (see Supplemental Analysis S4 in the Supplemental Material available online).

Stimuli. For stimuli in the learning phase, we used 20 natural White male faces from the Basel Face Database (Walker et al., 2018). These faces were systematically manipulated on the communion dimension, which is synonymous to the trustworthiness dimension in facial feature space (Chua & Freeman, 2021; Hong & Freeman, 2023). The stimuli consisted of White male faces whose features were increased +2 *SD* or decreased −2 *SD* in the trustworthiness/communion dimension.

For stimuli in the test phase, we used the White male faces from Study 1 of Wilson and Rule's (2015) work, which were obtained from the Florida Department of Corrections website. The researchers identified all White men on death row with a conviction of first-degree murder, which resulted in 226 White men (as of October 2014). In the same period, they also obtained a matched set of 226 White men who were convicted of first-degree murder in Florida but sentenced to life in prison. All face images were gray-scaled to minimize differences in lighting and obscure color cues from inmates' uniforms (which indicate sentencing). To control for influences of racial biases in these judgments (e.g., Blair et al., 2004; Xie et al., 2021), we used only the White male prisoners. We used a randomized set of 400 facial targets (200 sentenced to life in prison, 200 sentenced to death).

Procedure. Participants engaged in a two-part task. The first task involved a learning phase presented to participants ostensibly as a face-memory test. For control participants, the learning phase involved the 20 different faces paired with a name label. The trained participants viewed 20 faces paired with one-sentence trustworthy or untrustworthy behavioral descriptions (e.g., "volunteers at a homeless shelter" vs. "took a bribe to give a student a better grade"). The trustworthy and untrustworthy sentences reliably differed in trustworthiness or untrustworthiness and were rated as equally strong in a pretest. (See the Supplemental Material for full details on the sentences.) Participants were instructed to memorize the face–behavior or face–name pairings for a later test of face memory.

Both control and trained participants saw 10 unique trustworthy-looking faces and 10 unique untrustworthy-looking faces during the learning phase. For the trained group, the 10 trustworthy-looking faces were paired with untrustworthy behaviors 80% of the time, and the 10 untrustworthy-looking faces were paired with trustworthy behaviors 80% of the time. The face–behavior pairings were randomized for each subject. Each trial was self-paced and presented in a randomized order. A timeout warning (meant to ensure that participants carefully read the face–behavior/face–name pairings) remained on screen for 2,000 ms if participants spent less than 500 ms studying the face. Each face–behavior pairing was repeated three times, resulting in 60 learning trials.

Following the training, participants were instructed to rate a series of 400 faces on a given trait (Study 1a, trustworthiness; Study 1b, attractiveness). The 400 faces were shown in a randomized order, in four blocks of 100 faces each. Participants rated the faces on a given trait on a 7-point scale ranging from 1 (*not very trustworthy/attractive*) to 7 (*very trustworthy/attractive*). After the ratings task, participants learned that there would not be a test of face memory and were debriefed about the study's aims.

This study and the subsequent studies were reviewed and approved by the Institutional Review Board (IRB) of Columbia University.

Results

In this study and those that follow, we used logistic mixed-effects regression models to test whether the training affected the relationship between participants' responses to facial targets, whether those targets were sentenced to death or life in prison (Studies 1–3), and whether participants rendered a guilty versus innocent verdict about them (Study 4). Additionally, in Studies 1 through 3, we used analogous linear mixed-effects models that treated real-world sentencing outcome as an independent measure and predicted participants' responses to facial targets and found highly similar results (see Supplemental Analysis S1 in the Supplemental Material). These latter analyses ensured that, across all studies, random effects of both participants and stimuli could be adequately accounted for. To permit standardized comparisons across studies, we *z*-normalized continuous predictors. We report unstandardized regression coefficients (*b*), Wald *z*, and in logistic models, odds ratios (*OR*).

Study 1a: Trustworthiness ratings. First, we used a logistic mixed-effects model to predict sentencing outcome (0 = *life in prison*, 1 = *death*) using trustworthiness rating, training condition (control = −0.5, training = 0.5), and the interaction. The model allowed for random intercepts for participants and random slopes of trustworthiness rating for participants.

The main effect of trustworthiness was significant, *b* = −0.04, *SE* = .01, *z* = 5.27, *p* < .001, *OR* = 0.96, 95% confidence interval (CI) = [.95, .98], indicating that faces that were rated less trustworthy were more likely to belong to individuals who were sentenced to death than life in prison. The main effect of training was not significant, *b* = −0.01, *SE* = .01, *z* = 0.39, *p* = .69, *OR* = 1.00, 95% CI = [.97, 1.02]. Critically, there was a significant interaction, *b* = 0.05, *SE* = .01, *z* = 3.87, *p* < .001, *OR* = 1.06, 95% CI = [1.03, 1.09]. The effect of perceived trustworthiness was highly significant among control participants, simple *b* = −0.06, *SE* = .01, *z* = 6.62, *p* < .001, *OR* = 0.94, 95% CI = [.92, .96], indicating that faces that were rated less trustworthy were more likely to belong to individuals who were sentenced to death than life in prison (Fig. 2a). However, in the trained condition, trustworthiness was no longer predictive of real-world sentencing outcomes, simple *b* = −0.01, *SE* = .01, *z* = 0.96, *p* = .34, *OR* = 0.99, 95% CI = [.97, 1.01].

Study 1b: Attractiveness. Attractiveness judgments tend to be moderately correlated with trustworthiness judgments (Todorov et al., 2008), and Wilson and Rule (2015) found that attractiveness and trustworthiness had

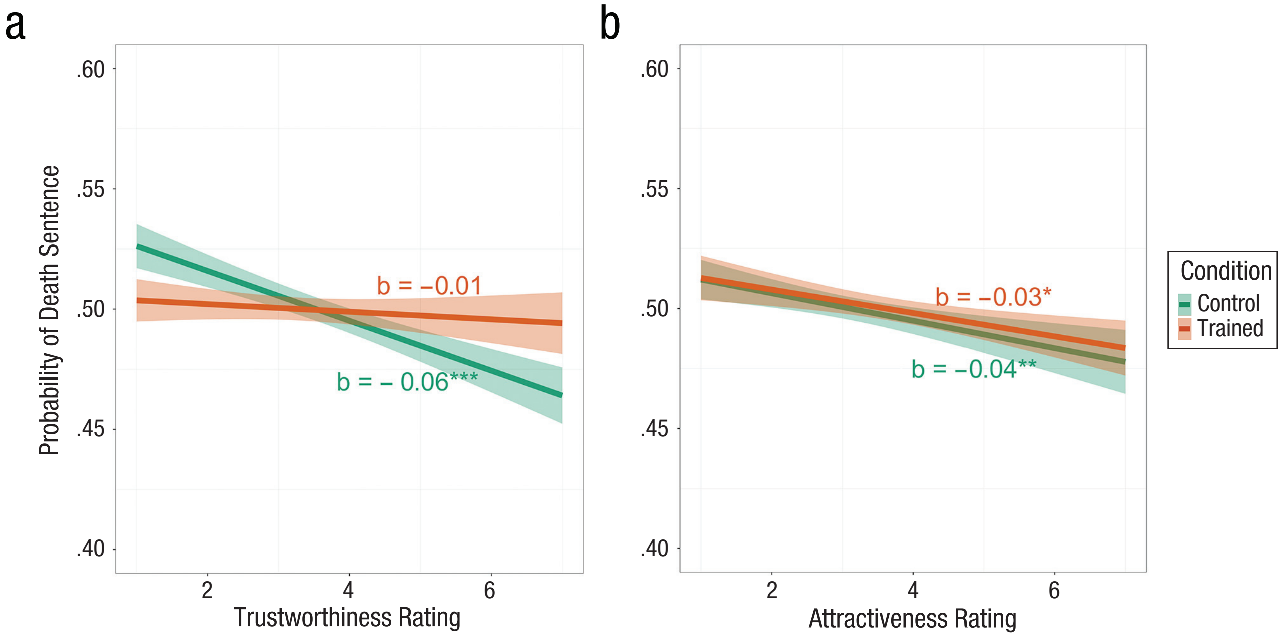


Fig. 2. The probability of a death sentence (vs. life in prison) as a function of (a) trustworthiness ratings in Study 1a and (b) attractiveness ratings in Study 1b for control and trained participants. The shaded areas represent 95% confidence intervals.
* < .05. ** < .01. *** < .001.

a similar relationship with sentencing outcome (although it was not significant when treated as a simultaneous predictor with trustworthiness). Consequently, here we predicted that attractiveress and trustworthiness would have a similar relationship with sentencing outcome, but—critically—that it would not be affected by the training. Using an analogous logistic mixed-effects model, we predicted sentencing outcome (0 = *life in prison*, 1 = *death*) on the basis of attractiveness rating, training condition (control = –0.5, training = 0.5), and the interaction. The model allowed for random intercepts for participants and random slopes of attractiveness rating for participants. The main effect of attractiveness was significant, b = –0.03, SE = .01, z = 3.81, p < .001, OR = 0.97, 95% CI = [.95, .98]; faces that were rated less attractive were more likely to belong to individuals who were sentenced to death than life in prison. The main effect of training was not significant, b = 0.01, SE = .01, z = 0.73, p = .47, OR = 1.01, 95% CI = [.98, 1.04]. Critically, the interaction was also not significant, b = 0.01, SE = .02, z = 0.70, p = .49, OR = 1.01, 95% CI = [.98, 1.05]. The effect of perceived attractiveness on sentencing outcome did not differ between control participants, simple b = –0.04, SE = .01, z = 3.10, p = .002, OR = 0.96, 95% CI = [.94, .99], and trained participants, simple b = –0.03, SE = .01, z = 2.25, p = .02, OR = 0.97, 95% CI = [.95, 1.00]; see Figure 2b.

Collapsing the data of Studies 1a and 1b and testing for a three-way interaction confirmed that the lack of a two-way interaction in attractiveness ratings of Study 1b was statistically different from the significant two-way interaction in trustworthiness ratings of Study 1a (see Supplemental Analysis S2 in the Supplemental Material).

These results demonstrate that a short training was effective in abolishing the relationship between facial trustworthiness judgments and real-world sentencing outcomes. Further, the training did not affect ratings of attractiveness, demonstrating the selectivity of the training in changing the associations between facial appearance and perceived trustworthiness.

Study 2

Although Study 1 demonstrated that the training abolished the relationship between trustworthiness ratings and real-world sentencing outcomes, the impact of the training on perceived trustworthiness alone does not provide insights into how the training may affect sentencing-related decision-making. Therefore, Study 2 extends these findings to hypothetical sentencing-related decisions directly. Following the same training procedure, participants acted as mock jurors and made sentencing recommendations for each target. If perceived trustworthiness contributes to criminal-sentencing decisions, then we would expect the training to similarly moderate the relationship between sentencing recommendations and real-world sentencing outcomes.

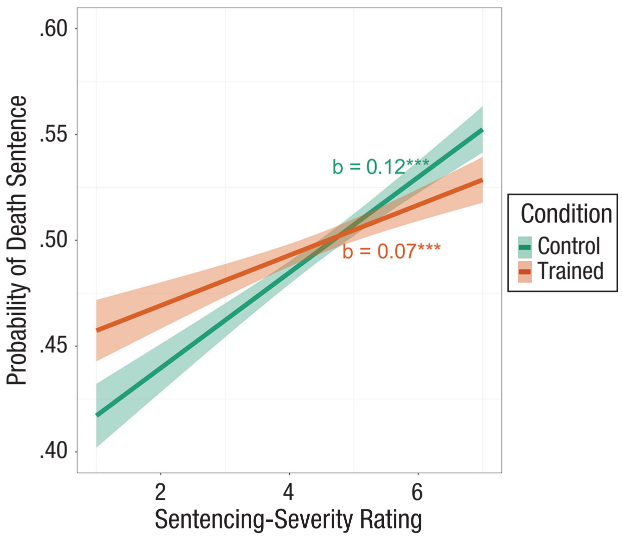


Fig. 3. The probability of a death sentence (vs. life in prison) as a function of sentencing-severity rating for control and trained participants in Study 2. The shaded areas represent 95% confidence intervals. *** < .001.

Method

Participants. A total of 220 participants were recruited from MTurk in exchange for monetary compensation. We followed the same data-exclusion criteria described in Study 1, resulting in the final sample size of 99 control participants and 102 trained participants (M_{age} = 39.26 years, SD_{age} = 11.30 years; 119 female, 80 male, 2 unidentified; 138 White, 21 Black, 20 Hispanic, 13 Asian, 2 Native American, 3 other, 4 unidentified). Exclusions did not differ by training condition (see Supplemental Analysis S4 in the Supplemental Material).

Procedure. The structure of the learning phase was identical to that of Study 1. Following the learning phase, participants acted as mock jurors for criminal cases. They were told on each trial that the defendant was found guilty of murder and that evidence demonstrating guilt is beyond a shadow of a doubt. Their role as a juror was to help make decisions regarding each prisoner's sentencing. Participants made a sentencing recommendation for each face on a 7-point scale (1 = *the most lenient sentence*, 4 = *an average sentence*, 7 = *the harshest possible sentence*). As in Study 1, the 400 faces were split between four blocks of 100 faces each.

Results

As in Study 1, we used a logistic mixed-effects model to predict real-world sentencing outcome (0 = *life sentence*, 1 = *death*) on the basis of participants' sentencing recommendation, the training condition (control = −0.5, training = 0.5), and the interaction. The models allowed for random intercepts for participants and random slopes of sentencing recommendation for participants. The main effect of sentencing recommendation was significant, b = 0.10, SE = .01, z = 10.35, p < .001, OR = 1.10, 95% CI = [1.08, 1.13], indicating that faces that were recommended for harsher sentences were more likely to belong to individuals who were sentenced to death than life in prison. The main effect of training was not significant, b = 0.004, SE = .01, z = 0.30, p = .77, OR = 1.00, 95% CI = [.98, 1.03]. Critically, the predicted interaction was significant, b = −0.05, SE = .02, z = 2.75, p = .006, OR = 0.95, 95% CI = [.92, .99]. Although control participants' sentencing recommendations were strongly predictive of whether inmates were actually sentenced to death versus life in prison, simple b = 0.12, SE = .01, z = 9.22, p < .001, OR = 1.13, 95% CI = [1.10, 1.16], this relationship was substantially attenuated among trained participants, simple b = 0.07, SE = .01, z = 5.51, p < .001, OR = 1.08, 95% CI = [1.05, 1.10] (see Fig. 3).

These results show that the training was effective not only in changing the extent to which participants perceived inmates' faces to be trustworthy or not, but also in changing participants' more direct, criminal-sentencing recommendations of the inmates. In turn, the training was able to mitigate the relationship between participants' hypothetical sentencing recommendations and the inmates' actual sentences (death or life in prison) in the real world.

Study 3

While Studies 1 and 2 demonstrated the training's effects on explicit judgments of faces, Study 3 explores whether the training may also affect more automatic, implicit evaluations of trustworthiness and attenuate their relationship with real-world sentencing outcomes.

Method

As in previous research (Chua & Freeman, 2021), we used sequential priming as an index of implicit trustworthiness evaluations. To the extent that facial targets are automatically evaluated on trustworthiness, trustworthy-looking face primes should facilitate response times when participants categorize trustworthy-related words, just as untrustworthy-looking face primes should facilitate participants' response times when participants categorize untrustworthy-related words.

Participants. Because implicit measures often yield noisier data (Gawronski & Hahn, 2019), we doubled our sample size and recruited 400 participants from Prolific. Participants received monetary compensation for participation. After the same exclusion criteria described in the earlier studies, our final sample size was 192 control and 193 trained participants (M_{age} = 41.79 years, SD_{age} = 13.57 years; 195 male, 182 female, 3 other, 5 unidentified; 262 White, 52 Black, 37 Hispanic, 14 Asian, 11 Multiracial, 3 Native American, 4 other). Exclusions did not differ by training condition (see Supplemental Analysis S4 in the Supplemental Material).

Stimuli. We used the same learning-phase stimuli as in the previous studies. For the test phase, we used the 400 White male inmate photographs used in the previous studies as facial primes for the sequential priming procedure. The priming paradigm requires participants' focal attention on rapidly presented facial primes, cued with a fixation cross. As these are real-world, ambient photographs with faces in different positions, we extracted faces from each photograph using OpenFace's face-extraction tool and applied affine transformation so that each face's eyes, nose, and mouth appear in approximately the same location (Amos et al., 2016). After this procedure, five images were excluded because the OpenFace tool failed to properly extract faces from those images, resulting in a total of 395 target faces (200 death, 195 life in prison). For the target words, we selected five synonymous words denoting trustworthiness versus untrustworthiness. These words were chosen to have similar word length and frequency of usage in the English language: *trustworthiness*—caring, kind, pleasant, trustworthy, warm; *untrustworthiness*—cold, cruel, mean, unpleasant, untrustworthy (average word length: *trustworthiness* = 6.6 characters, *untrustworthiness* = 7.2 characters. The average frequency of usage in the English language according to Brants and Franz (2006) is as follows: trustworthiness = 22,630,510, untrustworthiness = 25,455,201.

Procedure. The structure of the learning phase was identical to that of the previous studies. Following the learning phase (control or training), participants in both conditions completed the priming task. In this task, participants were asked to categorize target words as conveying trustworthiness or untrustworthiness. On each trial, a fixation cross (500 ms) was followed by a facial prime (200 ms), and then the target word remained on screen until a response was received. Participants classified a target word as trustworthy or untrustworthy as quickly and accurately as possible by key press (e.g., press "S" key for trustworthy, press "K" key for untrustworthy; key mapping was counterbalanced across participants).

For each participant, 15 photographs of inmates who were sentenced to life in prison and 15 photographs of those sentenced to death were randomly selected (each from the pool of 195 or 200 photographs in the two sentencing conditions). To ensure that the photographs were as representative of the pool of 395 photographs as possible, we used stratified sampling. That is, we used trustworthiness ratings of the 395 photographs from Wilson and Rule (2015) to divide faces into five categories (from below the 20th percentile to above the 80th percentile, with increments of 20 percentiles), resulting in 40 faces in each category for the death-sentence condition and 39 faces in each category for the life-in-prison condition. The trustworthiness ratings were taken from Wilson and Rule's (2015) second sample, as these ratings demonstrated higher interrater agreement (Cronbach's alpha) than their first sample. We then randomly selected three faces from each category in each sentencing condition across participants. Each face (30 faces total) was paired with each target word (10 words total) once, resulting in 300 trials total. Across 385 participants, each facial identity was presented at least 14 times (M = 29.2, SE = .26).

Results

There were comparable levels of high accuracy in categorization of the target words across conditions (see Supplemental Analysis S3 in the Supplemental Material). Following previous priming studies examining perceived trustworthiness (Chua & Freeman, 2021, 2022), we removed incorrect responses (6% of trials) and trials with response times faster than 250 ms and slower than 3,000 ms (1% of trials after excluding incorrect trials). These exclusions did not differ by condition (see Supplemental Analysis S4). For each participant and for each facial identity, we subtracted the average response time for categorizations of trustworthy words from the average response time for categorizations of untrustworthy words. Thus, following a given facial prime, a positive response-time difference score indicates faster responses to trustworthy words (implicit evaluation as trustworthy), whereas a negative response-time difference score indicates faster responses to untrustworthy words (implicit evaluation as untrustworthy).

We used a logistic mixed-effects model to predict sentencing outcome (0 = *life in prison*, 1 = *death sentence*) from each face's response-time difference score, training condition (control = −0.5, training = 0.5), and the interaction. The model allowed for random intercept for participants and random slopes of response-time difference score for participants. The main effect of response-time difference score was not significant, b = −0.02, SE = .02, z = 1.22, p = .22, OR = 0.98, 95% CI = [.94, 1.01], nor was the main effect of training, b = 0.003, SE = .04, z = 0.09, p = .93, OR = 1.10, 95% CI = [.93, 1.08]. More importantly, there was a significant interaction, b = 0.09, SE = .04, z = 2.51, p = .01, OR = 1.10,

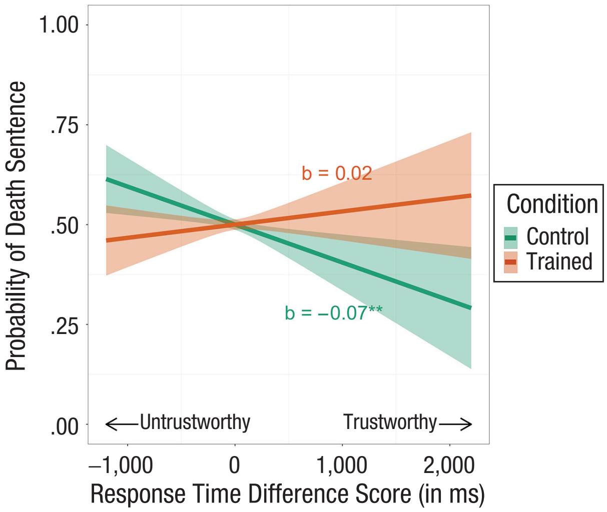


Fig. 4. The probability of a death sentence (vs. life in prison) as a function of RT difference score in the sequential priming task for Control and Trained participants in Study 3. A more negative response-time difference score indicates an implicit evaluation as untrustworthy, whereas a more positive response-time difference score indicates an implicit evaluation as trustworthy. The shaded areas represent 95% confidence intervals.
** < .01.

95% CI = [1.02, 1.18]. Specifically, the relationship between response-time difference score and likelihood of a death sentence was strong among control participants, simple $b = -0.07$, $z = 2.68$, $SE = .03$, $p = .007$, $OR = 0.93$, 95% CI = [.89, .98], indicating that faces that elicited a stronger implicit evaluation as untrustworthy were more likely to belong to individuals who were sentenced to death than life in prison (Fig. 4). However, this relationship was abolished among trained participants, simple $b = 0.02$, $SE = .03$, $z = 0.90$, $p = .37$, $OR = 1.03$, 95% CI = [.97, 1.08].

These results show that the training effects are not limited to explicit judgments of trustworthiness or the harshness of recommended sentences. Following training, even implicitly perceived trustworthiness was no longer predictive of prisoners' real-world life-or-death sentences.

Study 4

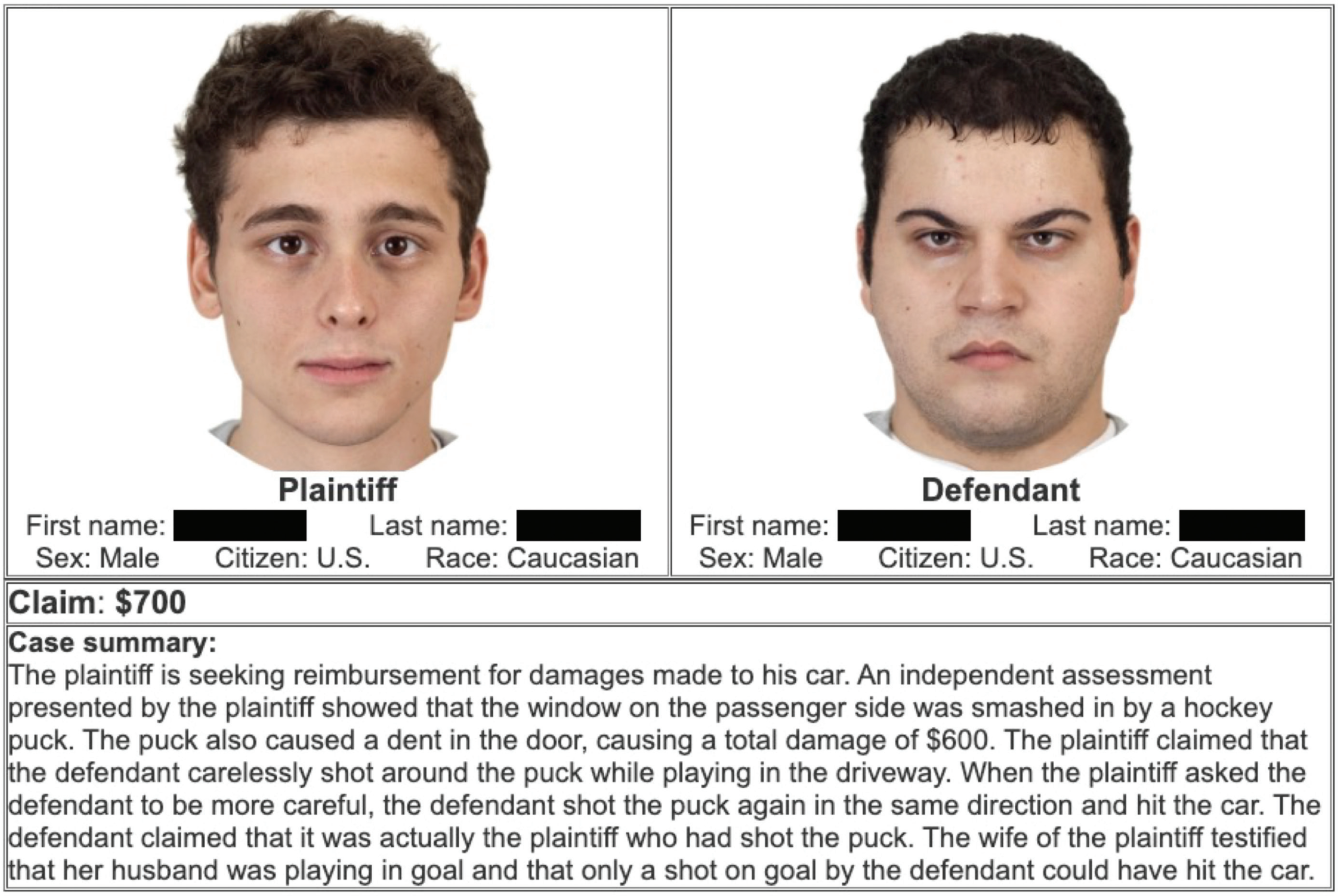
Studies 1, 2, and 3 demonstrated the effectiveness of counterstereotype training on reducing bias in both explicit and implicit responses to the faces of inmates. Study 4 explored whether the training can mitigate such bias in mock sentencing decisions in the presence of realistic, decision-relevant information. Previous research has attempted other interventions with such decision-making that operate on more conscious and deliberate processes, but the effect of facial appearance withstood these interventions (Jaeger et al., 2020). Here we test whether remapping underlying automatic associations between specific facial appearances and perceived trustworthiness may prove more successful.

Method

Participants. As in Study 3, we recruited 400 participants from Prolific. Participants received monetary compensation for participation. Applying the same exclusion criteria as in the previous studies, our final sample size was 197 control participants and 198 trained participants (M_{age} = 40.39, SD_{age} = 13.67; 224 female, 171 male; 267 White, 46 Hispanic, 41 Black, 18 Asian,17 Multiracial, 1 Native American, 1 Pacific Islander, 3 other, 1 unidentified). Exclusions did not differ by training condition (see Supplemental Analysis S4).

Stimuli. Stimuli for the learning phase were identical to those of the previous studies. The stimuli for the test phase consisted of case files for ten fictitious small-claims-court cases used by Jaeger and colleagues (2020). Each case file contained a photograph and personal details of the plaintiff and the defendant. Both parties were White American men with their names hidden. Targets were 20 images of White male individuals from the Chicago Face Database (Ma et al., 2015). On the basis of the database's normed ratings, ten individuals were rated the lowest on perceived trustworthiness, and the other ten individuals were rated the highest on perceived trustworthiness. All targets were then manipulated to look even more trustworthy-looking versus less trustworthy-looking by morphing them with a corresponding face prototype. That is, trustworthy-looking faces were morphed with a trustworthy-looking face prototype (3 *SD*s above the mean of trustworthiness), whereas untrustworthy-looking faces were morphed with an untrustworthy-looking face prototype (3 *SD*s below the mean of trustworthiness). Each case included a plaintiff and a defendant, and one individual looked trustworthy whereas the other looked untrustworthy (e.g., trustworthy-looking defendant vs. untrustworthy-looking plaintiff, and vice versa; Fig. 5). We used four stimuli sets, each comprising ten case files and 20 face images (one defendant and one plaintiff in each case). Within each set, the face images were randomly assigned to a specific case and role (plaintiff or defendant). Half of the cases had a trustworthy-looking plaintiff and an untrustworthy-looking defendant, whereas the remaining half had the roles reversed.

Procedure. As in the previous three studies, participants first completed a learning phase. Next, the test phase followed the procedures of Jaeger et al. (2020). Participants were randomly assigned to one of the four stimulus sets. We instructed participants to read each case carefully and indicate a sentence either in favor of



Plaintiff

First name: [redacted] Last name: [redacted]

Sex: Male Citizen: U.S. Race: Caucasian

Defendant

First name: [redacted] Last name: [redacted]

Sex: Male Citizen: U.S. Race: Caucasian

Claim: **$700**

Case summary:

The plaintiff is seeking reimbursement for damages made to his car. An independent assessment presented by the plaintiff showed that the window on the passenger side was smashed in by a hockey puck. The puck also caused a dent in the door, causing a total damage of $600. The plaintiff claimed that the defendant carelessly shot around the puck while playing in the driveway. When the plaintiff asked the defendant to be more careful, the defendant shot the puck again in the same direction and hit the car. The defendant claimed that it was actually the plaintiff who had shot the puck. The wife of the plaintiff testified that her husband was playing in goal and that only a shot on goal by the defendant could have hit the car.

Fig. 5. An example of a case file with a trustworthy-looking plaintiff and an untrustworthy-looking defendant in Study 4. The figure was adapted from Jaeger et al. (2020).

the plaintiff or the defendant. After making sentencing decisions, participants indicated their confidence in their verdict on a 9-point scale (1 = *not confident at all*, 9 = *extremely confident*) after each case. If participants decided in favor of the plaintiff, they also specified the amount of compensation the plaintiff should receive on a scale that varied from 50% to 100% of the original claim, with increments of 10%. Although the primary aim with these latter ratings was to replicate the procedures of Jaeger et al. (2020), the confidence ratings also provide an opportunity to shed additional light on the mechanism behind the training. Because the training attempts to affect automatic associations at an implicit level, whereas more deliberate processing of the targets is presumably unaffected, we would not expect the training to diminish participants' confidence in their decisions.

Results

Sentencing decision. We used a logistic mixed-effects model to predict sentencing decisions (0 = *defendant is not guilty*, 1 = *defendant is guilty*) on the basis of facial trustworthiness (–0.5 = untrustworthy-looking defendant, 0.5 = trustworthy-looking defendant), training condition (control = –0.5, training = 0.5), and the interaction. The model allowed for random intercepts for participants and case, random slopes of facial trustworthiness for participants, and random slopes of training condition for case stimuli. Replicating Jaeger et al. (2020), the main effect of facial trustworthiness was significant, $b = -0.27$, $SE = .07$, $z = 3.92$, $p < .001$, $OR = 0.76$, 95% CI = [.66, .87], indicating that untrustworthy-looking defendants were more likely to be found guilty than trustworthy-looking defendants. The main effect of training was not significant, $b = 0.01$, $SE = .09$, $z = 0.13$, $p = .90$, $OR = 1.01$, 95% CI = [.85, 1.20]. Critically, the predicted interaction was significant, $b = 0.30$, $SE = .14$, $z = 2.16$, $p = .03$, $OR = 1.35$, 95% CI = [1.03, 1.77]. Among control participants the effect of facial trustworthiness was strong, $b = -0.42$, $SE = .10$, $z = 4.28$, $p < .001$, $OR = 0.66$, 95% CI = [.54, .80]. However, among trained participants facial trustworthiness no longer affected participants' decisions of whether defendants were guilty or innocent, $b = -0.12$, $SE = .10$, $z = 1.26$, $p = .21$, $OR = 0.88$, 95% CI = [.73, 1.07]; see Figure 6.

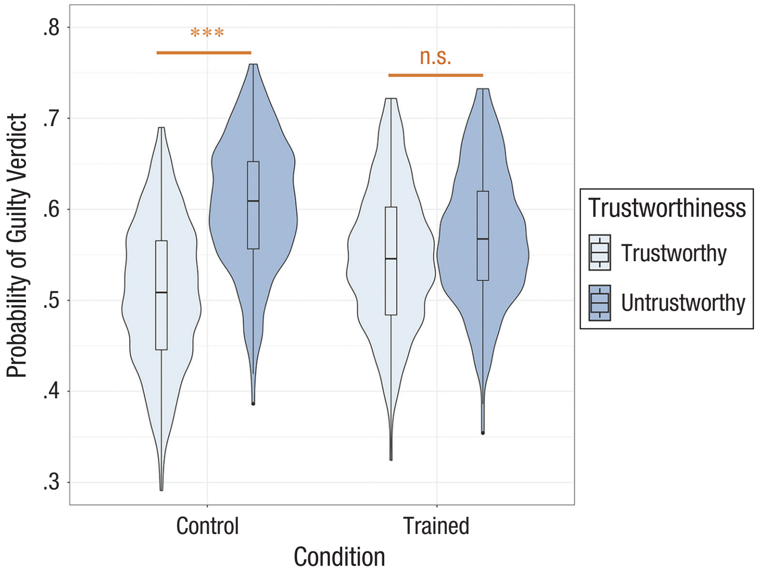


Fig. 6. The probability of a guilty (vs. an innocent) verdict as a function of facial trustworthiness for control and trained participants in Study 4. The whiskers represent 1.5 times the interquartile range.
n.s >.05. *** < .001.

Confidence in verdict. We used a linear mixed-effects model to predict confidence ratings on the basis of facial trustworthiness (–0.5 = untrustworthy-looking defendant, 0.5 = trustworthy-looking defendant), training condition (control = –0.5, training = 0.5), guilty verdict (innocent = –0.5, guilty = 0.5), and their interactions. The model allowed for random intercepts for participants and case, random slopes of facial trustworthiness for participants, and random slopes of training condition for case stimuli. We found a main effect of guilty verdict, *b* = 0.30, *SE* = .04, *z* = 6.27, *p* < .001, 95% CI = [.20, .39]. Replicating prior work (Jaeger et al., 2020), participants were more confident in their verdicts when they found the defendant guilty rather than innocent. We also found a significant interaction between guilty verdict and training condition, *b* = 0.20, β = 0.03, *SE* = .09, *p* = .04, 95% CI = [.01, .38]. Although control participants were more confident of their guilty (*M* = 6.18, *SE* = .05) versus innocent (*M* = 5.94, *SE* = .06) decisions, simple *b* = 0.20, *SE* = .07, *z* = 2.95, *p* = .003, 95% CI = [.07, .33], this effect nearly doubled in size among trained participants, with trained participants even more confident of their guilty (*M* = 6.36, *SE* = .05) versus innocent (*M* = 5.92, *SE* = .07) decisions, simple *b* = .40, *SE* = .07, *z* = 5.93, *p* < .001, 95% CI = [.26, .53]. No other effects were significant (*p* > .05).

Compensation awarded to the plaintiff. Lastly, we examined the amount of money that was awarded to the plaintiff after a guilty verdict. Again, we used a linear mixed-effects model to predict the money awarded to the plaintiff for a guilty verdict based on facial trustworthiness (–0.5 = untrustworthy-looking defendant, 0.5 = trustworthy-looking defendant), training condition (control = –0.5, training = 0.5), and the interaction between the two. No effects were significant (*p* > .05).

These results show that even in the presence of realistic, decision-relevant information, control participants rely on facial trustworthiness to make crucial sentencing decisions. Critically, our training intervention was effective in eliminating participants' reliance on facial trustworthiness, allowing them to exclusively use the decision-relevant information. Trained participants were also no less confident in their decisions than control participants; in fact, when making guilty verdicts they were slightly more confident than control participants.

General Discussion

Across four studies (*N* = 1,400), we have demonstrated that a counterstereotype training intervention was effective in reducing reliance on facial trustworthiness in consequential social judgments. In Studies 1 and 2, the training successfully mitigated the relationship between explicit judgments of prison inmates based on their facial photographs and their actual sentencing outcomes. Among control participants, we observed that faces judged as less trustworthy (or recommended for harsher sentences) were more likely to belong to the inmates who received the death penalty, replicating the findings from Wilson and Rule (2015, 2016). However, after training, these relationships were reduced or eliminated. In Study 3, we demonstrated that the training abolished even the relationship between automatically and implicitly perceived trustworthiness and real-world sentencing outcomes. Finally, in Study 4, we showed that facial trustworthiness influenced mock sentencing decisions in the presence of realistic, decision-relevant information among control participants; this reliance on facial trustworthiness was eliminated among trained participants. Together, these results show that dismantling the associations between specific facial appearances and perceived traits through this kind of associative learning paradigm has promise in mitigating harmful biases due to facial appearance that arise in consequential social judgments. Interestingly, Kramer and Gardner (2020) reported an inability to replicate the relationship between perceived trustworthiness and real-world sentencing outcomes originally found by Wilson and Rule (2015), although here we consistently replicated it among control participants. This likely relates to the smaller samples of faces they used and issues of restricted range that are not applicable to the current studies (see the full discussion in the Supplemental Material).

The findings expand on previous studies that have used statistical learning to remap the associations between specific facial features and traits (Chua & Freeman, 2021, 2022; Lick et al., 2017) by applying this training to real-world contexts. Previous studies employing this training have used faces that were artificially manipulated on trustworthiness-related features, raising questions of generalizability. In Studies 1 through 3, we used real, unaltered facial photographs of Florida prisoners convicted of homicide and provide evidence that the training can reduce the predictive power of facial trustworthiness in real-world sentencing outcomes. Such results demonstrate the general applicability of the training to natural, unconstrained faces. Our findings also show that the training does not lead to task artifacts, such as participants diverting attention away from the face stimuli, as trained participants were equally responsive to facial attractiveness (Study 1b) and equally confident, if not more confident, in their ultimate sentencing decisions as control participants (Study 4). The results suggest, instead, that the training changes specific associations between facial appearances and the targeted trait (trustworthiness). Finally, in contrast to other interventions that rely on more conscious, deliberate processes (e.g., nudging or educating), our findings suggest that operating on more automatic, implicit associations may have greater promise in mitigating facial stereotyping.

A critical question is the long-term persistence of the training. The duration of the training is short, and if we consider counterstereotype interventions in the context of racial bias we would not expect such a brief training to necessarily result in long-term changes on its own (Lai et al., 2016). Incorporating the counterstereotype training used here as part of more comprehensive, multiweek habit-breaking interventions may be worthwhile, although their long-term effects on racial bias are mixed (Devine et al., 2012; Forscher et al., 2017). However, even demonstrating initial malleability of facial stereotyping in real-world contexts is impressive, as previous interventions have failed, and reliance on facial appearance is regarded as having a strong evolutionary and functional basis, making it difficult to change (Jaeger et al., 2019; Oosterhof & Todorov, 2008; Zebrowitz & Montepare, 2008). This contrasts with racial bias, which is learned through the sociocultural environment.

Our studies are limited in several ways. The participants were recruited from online platforms, limiting the generalizability of our controlled experimental setting to the level of complexity and amount of information available to a real juror. Moreover, our studies focused on White male faces, further limiting generalizability, because we wished to focus exclusively on the trustworthiness trait dimension without interference from gender or racial bias. Although we would expect that the ability to mitigate facial stereotyping should generalize across diverse genders and races, undertaking such work may present several challenges. The facial features that drive trait impressions differ across gender and race, and gender and racial biases affect the structuring of these impressions (Xie et al., 2021). Making gender or race salient in the task may activate social-desirability concerns that reduce the intuitiveness of face judgments and lead to deliberately altered responses. From a statistical-learning perspective, it is not clear whether participants would process the counterstereotype trait information independently from gender and race or instead associate the information with specific social categories; this raises questions of stimulus attention and category generalization during the training. These challenges are hardly insurmountable, and future researchers should work toward developing an integrated paradigm to understand how facial stereotyping is mitigated in more diverse social contexts. More generally, applying associative learning principles to understand the malleability of simultaneous social group–based biases and facial appearance–based biases, and how these interact, may also lead to enhanced bias intervention success in both domains.

In summary, trait impressions from facial appearance have a significant impact on criminal sentencing, hiring, politics, and other areas of life (Olivola et al., 2014). Here we show that real-world facial stereotyping is malleable to the information we receive about faces and their statistical co-occurrence with specific traits. If there are consequential judgments that are biased by facial stereotypes, our findings suggest that they have the potential to be flexibly remapped and dismantled.

Transparency

Action Editor: Mark Brandt
Editor: Patricia J. Bauer
Author Contribution(s)

Youngki Hong: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Validation; Visualization; Writing – original draft; Writing – review & editing.

Kao-Wei Chua: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Writing – original draft; Writing – review & editing.

Jonathan B. Freeman: Conceptualization; Funding acquisition; Investigation; Methodology; Project administration; Validation; Writing – original draft; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported in part by Research Grant No. BCS-2235130 to J. B. Freeman.

Open Practices

This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iD

Jonathan B. Freeman https://orcid.org/0000-0002-2061-8460

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976231215238

References

Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). *OpenFace: A general-purpose face recognition library with mobile applications* (Technical report). CMU-CS-16-118. CMU School of Computer Science.

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*(2), 269–278. https://doi.org/10.1037/1528-3542.6.2.269

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of afrocentric facial features in criminal sentencing. *Psychological Science*, *15*(10), 674–679. https://doi.org/10.1111/j.0956-7976.2004.00739.x

Brants, T., & Franz, A. (2006). *Web 1T 5-gram Version 1* [data set]. Linguistic Data Consortium. https://doi.org/10.35111/CQPA-A498

Chua, K.-W., & Freeman, J. B. (2021). Facial stereotype bias is mitigated by training. *Social Psychological and Personality Science*. Advance online publication. https://doi.org/10.1177/1948550620972550

Chua, K.-W., & Freeman, J. B. (2022). Learning to judge a book by its cover: Rapid acquisition of facial stereotypes. *Journal of Experimental Social Psychology*, *98*, Article 104225.

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*, *48*(6), 1267–1278.

Forscher, P. S., Mitamura, C., Dix, E. L., Cox, W. T. L., & Devine, P. G. (2017). Breaking the prejudice habit: Mechanisms, timecourse, and longevity. *Journal of Experimental Social Psychology*, *72*, 133–146.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731.

Gawronski, B., & Hahn, A. (2019). Implicit measures: Procedures, use, and interpretation. In H. Blanton, J. M. LaCroix, & G. D. Webster (Eds.), *Measurement in social psychology* (pp. 29–55). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9780429452925-2

Hong, Y., & Freeman, J. B. (2023). Shifts in facial impression structures across group boundaries. *Social Psychological and Personality Science*. Advance online publication. https://doi.org/10.1177/19485506231193180

Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2019). Explaining the persistent influence of facial cues in social decision-making. *Journal of Experimental Psychology: General*, *148*(6), 1008–1021.

Jaeger, B., Todorov, A. T., Evans, A. M., & van Beest, I. (2020). Can we reduce facial biases? Persistent effects of facial trustworthiness on sentencing decisions. *Journal of Experimental Social Psychology*, *90*, Article 104004.

Kramer, R. S. S., & Gardner, E. M. (2020). Facial trustworthiness and criminal sentencing: A comment on Wilson and Rule (2015). *Psychological Reports*, *123*(5), 1854–1868.

Krendl, A. C., Rule, N. O., & Ambady, N. (2014). Does aging impair first impression accuracy? Differentiating emotion recognition from complex social inferences. *Psychology and Aging*, *29*(3), 482–490. https://doi.org/10.1037/a0037146

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., . . . Nosek, B. A. (2016). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*, 1765–1785.

Lick, D. J., Alter, A. L., & Freeman, J. B. (2017). Superior pattern detectors efficiently learn, activate, apply, and update social stereotypes. *Journal of Experimental Psychology: General*, *147*(2), 209–227.

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*(4), 1122–1135.

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, *18*(11), 566–570. https://doi.org/10.1016/j.tics.2014.09.007

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, USA*, *105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105

Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PLOS ONE*, *7*(3), Article e3429310. https://doi.org/10.1371/journal.pone.0034293

Todorov, A., & Olson, I. R. (2008). Robust learning of affective trait associations with faces when the hippocampus is damaged, but not when the amygdala and temporal pole are damaged. *Social Cognitive and Affective Neuroscience*, *3*(3), 195–203.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. https://doi.org/10.1016/j.tics.2008.10.001

Walker, M., Schönborn, S., Greifeneder, R., & Vetter, T. (2018). The Basel Face Database: A validated set of photographs reflecting systematic differences in Big Two and Big Five personality dimensions. *PLOS ONE*, *13*(3), Article e0193190.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science*, *26*(8), 1325–1331. https://doi.org/10.1177/0956797615590992

Wilson, J. P., & Rule, N. O. (2016). Hypothetical sentencing decisions are associated with actual capital punishment outcomes: The role of facial trustworthiness. *Social Psychological and Personality Science*, *7*(4), 331–338. https://doi.org/10.1177/1948550615624142

Xie, S. Y., Flake, J. K., Stolier, R. M., Freeman, J. B., & Hehman, E. (2021). Facial impressions are predicted by the structure of group stereotypes. *Psychological Science*, *32*(12), 1979–1993.

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social Personality Psychology Compass*, *2*(3), 1497.