Journal of Experimental Psychology: General

© 2023 American Psychological Association
ISSN: 0096-3445

2023, Vol. 152, No. 6, 1622–1638
https://doi.org/10.1037/xge0001335

Stereotypes Disrupt Probabilistic Category Learning

Yrian Derreumaux[1], Jacob Elder[1], Gaurav Suri[2], Avi Ben-Zeev[2], Thelonious Quimby[3], and Brent L. Hughes[1]
[1] Department of Psychology, University of California, Riverside
[2] Department of Psychology, San Francisco State University
[3] Independent Researcher, Oakland, California

Racial stereotypes exert pernicious effects on decision-making and behavior, yet little is known about how stereotypes disrupt people's ability to learn new associations. The current research interrogates a fundamental question about the boundary conditions of probabilistic learning by examining whether and how learning is influenced by preexisting associations. Across three experiments, participants learned the probabilistic outcomes of different card combinations based on feedback in either a social (e.g., forecasting crime) or nonsocial (e.g., forecasting weather) learning context. During learning, participants were presented with either task-irrelevant social (i.e., Black or White faces) or nonsocial (i.e., darker or lighter clouds) stimuli that were stereotypically congruent or incongruent with the learning context. Participants exhibited learning disruptions in the social compared to nonsocial learning context, despite repeated instructions that the stimuli were unrelated to the outcome (Studies 1 and 2). We also found no differences in learning disruptions when participants learned in the presence of negatively (Black and criminal) or positively valenced stereotypes (Black and athletic; Study 3). Finally, we tested whether learning decrements were due to "first-order" stereotype application or inhibition at the trial level, or due to "second-order" cognitive load disruptions that accumulate across trials due to fears of appearing prejudiced (aggregated analysis). We found no evidence of first-order disruptions and instead found evidence for second-order disruptions: participants who were more internally motivated to respond without prejudice, and thus more likely to self-monitor their responses, learned less accurately over time. We discuss the implications of the influence of stereotypes on learning and memory.

Keywords: stereotyping and prejudice, learning and memory, social cognition, internal motivation to respond without prejudice

Supplemental materials: https://doi.org/10.1037/xge0001335.supp

Humans' ability to learn probabilistic associations through trial and error is key to survival. Consider the proverbial hot stove effect: A child might be told countless times to be careful playing around a stove without learning, but getting burned once can produce a lifelong association. Indeed, there exist numerous models of learning that account for how efficiently humans learn and update probabilities from subjective "hands-on" experiences (e.g., Cleeremans & McClelland, 1991; Gobet et al., 2001; Kumaran et al., 2016; Lee et al., 2012). One important unanswered question is how the cognitive system handles different salient contextual cues during learning, such as preexisting associations that are irrelevant to the learning context but may nonetheless impede new learning. For instance, police officers that deploy search and stop practices are tasked to learn objective indicators of suspicious behavior. However, research finds that officers often rely on preexisting negative stereotypes to inform suspicions when conducting stops and searches (Minhas & Walsh, 2021). Thus, negative preexisting associations between Black men as criminal seem to interfere with police officers' ability to learn objective indicators of suspicion. Actively ignoring such disruptive preexisting associations is an important challenge to the learning process (Niv et al., 2015). Here, we interrogate a fundamental question about the boundary conditions of probabilistic category learning: When are we able to ignore preexisting associations that may interfere with learning and when can we not?

There are countless different kinds of preexisting associations that people use daily that could influence learning, including both social and nonsocial features of their environments. For instance, many people have strong associations between cloudy skies and rain and thus carry an umbrella or put on a raincoat when they see cloudy

This article was published Online First March 6, 2023.
Yrian Derreumaux https://orcid.org/0000-0002-8445-7635
Brent L. Hughes https://orcid.org/0000-0001-8732-8727
Avi Ben-Zeev is now at London Metropolitan University.
This research benefited greatly from the input and feedback we received at several conferences and seminars, including posters that were presented at the Society for Personality and Social Psychology (2020), and the Social Affective Neuroscience Society (2019). We also thank Nick Camp (University of Michigan) and Jamil Bhanji (Rutgers University) for helpful discussions. Data and materials are publicly available at: https://osf.io/fxkh7/?view_only=a24caf36566b418480560925a9e65f5c.
Yrian Derreumaux served as lead for all stages of the research. Jacob Elder contributed to data curation and formal analysis. Thelonious Quimby contributed to programming and data curation. Gaurav Suri and Avi Ben-Zeev contributed to conceptualization, methodology, project administration, and writing. Brent Hughes contributed to funding acquisition, conceptualization, methodology, project administration, and writing.

Correspondence concerning this article should be addressed to Brent L. Hughes, Department of Psychology, University of California, Riverside, 900 University Avenue, Riverside, CA, United States. Email: bhughes@ucr.edu

skies. As highlighted in the example above, people also have strong preexisting associations about other people based on their demographics (e.g., race and gender), which may likewise influence learning and behavior (e.g., Allidina & Cunningham, 2021; Brewer, 2001; Brewer & Pierce, 2005; Hogg et al., 1995; Onorato & Turner, 2004; Willer et al., 1989). For example, past work finds that individuals are faster to learn the association between a cue and a face when the face is associated with outgroup threat (Lindström et al., 2014), and more readily associate aversive experiences (e.g., electric shocks) with Black compared to White individuals (Olsson et al., 2005). Moreover, recent work finds that exposure to social stereotypes shapes how people learn about new group members and can also shape people's own, personal group-based preferences (Stillerman et al., 2020). Together, this body of work suggests that social stereotypes may play a critical role in probabilistic learning. Importantly, this past work has not yet examined the influence of preexisting associations on new categorical learning, or whether social and nonsocial associations differentially influence new learning. One factor that may influence people's ability to ignore task-irrelevant associations is how "sticky" the association is. We use the term "stickiness" to describe associations that are resistant to change, difficult to inhibit, and may thus be more difficult to ignore (e.g., Arrow 1998; Blind & Lottani von Mandash, 2021).

There is good reason to believe that one class of social associations that people hold, namely racial stereotypes, may be particularly sticky and difficult to ignore, which may disrupt learning. Decades of research on stereotyping and prejudice show that racial associations powerfully influence decisions and behavior (e.g., Devine, 1989; Eberhardt et al., 2004; Simon, 1956; Welch, 2007; also see, Kirsch et al., 2004; Staddon & Cerutti, 2003), such that the mere presence of an individual can instantaneously conjure up rich and affectively laden mental representations based on their social category (Barnett et al., 2021; Fiske & Neuberg, 1990; Schiller et al., 2009; Stephan & Stephan, 1985). One such association that has garnered a great deal of attention in the United States is the stereotype that Black men are aggressive and dangerous (Bodenhausen & Wyer, 1985; Eberhardt et al., 2004; Welch, 2007). For instance, research finds that people are more likely to misjudge a tool as a weapon when they are held by Black compared to White men (Correll et al., 2015), are quicker to shoot an armed target if they are Black compared to White (Correll et al., 2002; Payne, 2001), and assign harsher sentences to inmates with more Afrocentric features (I. Blair et al., 2004). Together, this body of work suggests that this class of social associations may be particularly sticky and may thus impinge on learning.

At the same time, changes in the social milieu have reduced the social desirability and acceptability of expressing stereotypes, which has reduced explicit racial prejudice (Blinder et al., 2013). Consequently, people face personal pressures—due to internal egalitarian values and external social pressures to comply with social norms—to self-monitor and regulate the expression of stereotypes and prejudice (Plant & Devine, 1998). This act of self-regulation to avoid appearing prejudiced has been shown to deplete subsequent cognitive functioning (Richeson et al., 2003; Richeson & Shelton, 2003; Richeson & Trawalter, 2005; Rubien-Thomas et al., 2021; Shelton & Richeson, 2005). For instance, one study found that after an interaction with a Black confederate, White participants performed worse on a subsequent decision-making task (i.e., Stroop task), presumably due to depleted cognitive resources from self-regulation during the preceding interracial interaction (Richeson & Shelton, 2003). Notably, this work focused on the effects of cognitive depletion from interracial interactions on subsequent task performance, rather than examining whether the preexisting racial associations that people hold (e.g., stereotypes) disrupt new learning, in real time. As these two processes—learning on the one hand and decision-making on the other hand—are categorically distinct, we do not yet know how such preexisting associations disrupt the cognitive processes involved in learning.

The current work aims to fill this important gap by testing *whether* and *how* task-irrelevant social associations, such as racial stereotypes about Black men as threatening, disrupt probabilistic category learning and directly compare whether the effects of social associations are stronger than task-irrelevant nonsocial associations, such as associations between clouds and rain. Consistent with a host of research demonstrating that stereotypes influence cognition and behavior, we predict that racial associations will disrupt the acquisition of new associations more than nonsocial associations. Regarding *how* social associations may prove disruptive, past research suggests at least two possible mechanisms.

First, stereotypes may exert "first-order" effects that disrupt learning on a trial-by-trial basis due to stereotype application or inhibition. For instance, stereotypes that are irrelevant to a learning context may disrupt learning by increasing people's propensity to learn stereotypes congruent over incongruent outcomes (i.e., stereotype application). That is, if people apply a stereotype (e.g., a person predicts that an individual has a characteristic ["is a criminal"] simply because they belong to a racial group [e.g., "is Black"]), then this will lead to suboptimal learning, as race is orthogonal to predictions about crime. Likewise, stereotypes may disrupt learning on a trial-by-trial basis by increasing people's propensity to inhibit stereotypically congruent responses. That is, if people inhibit stereotype-congruent responses (e.g., a person is reluctant to predict that an individual has a characteristic ["is a criminal"] simply because they belong to a racial group [e.g., "is Black"]), then this will also lead to suboptimal learning, as race is orthogonal to predictions about crime. An optimal learner should instead ignore task-irrelevant features of the learning environment.

In contrast to first-order stereotype application or inhibition, the presence of stereotypes may also disrupt learning via "second-order" effects due to taxed cognitive functioning. Support for this hypothesis comes from past work showing that self-monitoring during interracial interactions can be cognitively taxing and thus diminish performance on subsequent cognitive control tasks (Richeson et al., 2003; Richeson & Shelton, 2003). This hypothesis is also consistent with work demonstrating that extraneous cognitive load manipulations (e.g., divided attention) disrupt performance on cognitive tasks (Sweller, 2011). Thus, the act of self-monitoring one's behavior while also performing a learning task may disrupt learning. Moreover, if taxed cognitive functioning underpins the learning disruptions, then participants who are more internally motivated to respond without prejudice should perform worse than participants with lower internal motivations (IMS), as these individuals are more burdened by the additional cognitive effort of monitoring their responses (Johns et al., 2008). Individual differences are important predictors of a variety of real-world outcomes (Ozer & Benet-Martínez, 2006), and may thus help to elucidate key mechanisms underlying probabilistic learning and their disruptions.

The Current Studies: Experimental Paradigm and Predictions

To examine whether and how preexisting task-irrelevant associations disrupt probabilistic learning, we created a modified version of the weather prediction task (WPT; Knowlton et al., 1994). The WPT is a probabilistic learning task where participants learn to classify 14 different combinations of 4 cues (cards that depict geometric symbols) into categories (sun or rain). Each card or card combination maintains a different probability of predicting sun or rain, which participants learn over time based on feedback. Our motivation for using the WPT is two-fold. First, the WPT emulates multidimensional statistical features commonly found in real-world learning, where people update learned associations over time based on discrete cues that are partially valid indicators of categorical outcomes (Kruschke & Johansen, 1999). Second, while learning during the WPT has traditionally been embedded within a nonsocial context (e.g., predicting rain or sun absent any social cues), the learning context can be adapted to include social cues (e.g., predicting crime in the presence of human faces). In doing so, we can directly test the extent to which contexts that elicit task-irrelevant social and nonsocial associations interfere with the learning process. Importantly, while there are many other kinds of social and nonsocial associations, here we focus on specific race-based social associations and weather-based nonsocial associations.

Across experiments, task-irrelevant social stimuli consisting of Black and White images of male faces, and task-irrelevant nonsocial stimuli consisting of darker and lighter images of clouds, were introduced into a social learning context (predicting “steal” vs. “no steal”) and nonsocial learning context (predicting “sun” vs. “rain”). By randomly presenting these images on a trial-by-trial basis, we introduced task-irrelevant stimuli that are either congruent or incongruent with one of the two possible predicted outcomes. For instance, the presence of a Black male face is stereotypically congruent with the prediction of “steal,” just as the presence of darker rain clouds is congruent with the prediction of “rain.” Conversely, the presence of a White male face is stereotypically incongruent with the prediction of “steal,” just as the presence of lighter clouds is incongruent with the prediction of “rain.” Thus, by introducing stimuli that are either congruent or incongruent with prior associations, we can examine the degree to which social and nonsocial associations interfere with learning. Critically, participants were repeatedly told that the social and nonsocial stimuli were distractors and should therefore be ignored. That is, the design created an expectation for participants to ignore these irrelevant distractors, and the empirical question is whether social associations are more difficult to ignore than nonsocial associations, and therefore disrupt learning.

Study 1 tests whether people learn differently in a social (i.e., forecasting crime) versus a nonsocial context (i.e., forecasting weather). Study 2 further tests whether learning disruptions in the social contexts is merely due to attention directed to distracting human faces or due to the unique combination of faces that were embedded within stereotype-eliciting contexts. Study 3 manipulates the valence of the learning context to examine whether stereotypes interfere with learning for both negatively valenced and positively valenced stereotypes. Next, we conduct an aggregated analysis that examines *how* social contexts disrupt learning. The aggregated analysis tests competing hypotheses regarding whether learning decrements are attributable to first-order disruptions at the trial level due to stereotype application or inhibition, or to second-order disruptions that accumulate over time due to taxed cognitive functioning. Study 2 was preregistered (see https://aspredicted.org/blind.php?x=n22299).

Study 1: Do Racial Versus Weather Associations Differentially Disrupt Learning?

Study 1 was designed to test whether one class of social associations, namely stereotypes about race and crime, disrupts learning more than one class of nonsocial associations, namely associations between clouds and rain.

Method

Participants

Participants ($N = 114$) were recruited from Amazon’s Mechanical Turk (MTurk). They received a $2.00 payment for completing the roughly 20-min study. Only MTurk workers with a history of providing good-quality responses were allowed to participate in the study (i.e., an acceptance rate of >98%). The sample size was based on prior research using the same task design (Gluck, 2002).[1] Participants were excluded if they failed to achieve an average accuracy of 52%, which indicated that they were randomly guessing.[2] After applying these exclusion criteria, $N = 100$ participants remained.[3] The sample had a mean age of 37.30, $SD = 11.07$ (56% self-identified male and 44% self-identified female), and was 73% White, 10% Black, 10% Asian, 4% Latino, and 3% Other.

Task Design

The task was built using JavaScript and was hosted on biz.nf—a third-party website (see https://osf.io/fxkh7/?view_only=a24caf36566b418480560925a9e65f5c for script and page 2 of the online supplemental materials for a description of how the task was built). Across all studies, participants were asked to learn the probabilities of four cards (i.e., square, diamond, circle, and triangle), which were independently associated with each possible outcome (weather: sun or rain; crime: steal or no steal) at a fixed probability (Knowlton et al., 1994). In each trial, participants were presented with a particular combination of one, two, or three of the four cards (never all four or none). There was a total of 14 possible card combinations used to generate 300 trials with different frequencies, in which the two outcomes occurred equally often (see Supplemental Table 1 in the online supplemental materials). The card or card patterns were pseudo-randomized across trials to ensure they did not appear twice in succession.

[1] Gluck (2002) employed a repeated measures analysis of variance (ANOVA) to examine learning across four blocks of 50 trials, using a sample size of 30 participants. Given the large differences in our design (between vs. within subjects) and analyses (logistic mixed model vs. repeated measures ANOVA), we were unable to use the effect size observed in Gluck (2002). We instead relied on a judgement call (see e.g., Lakens, 2022) ultimately deciding to roughly double the sample size for our initial study.

[2] The exclusion criterion for accuracy was chosen during a group discussion with the idea that 52% was just enough above 50% to constitute some learning. This value refers to the mean accuracy across all trials, rather than a running average. Importantly, the results do not meaningfully change when we apply a 50% accuracy threshold (see Supplemental Table 9 in the online supplemental materials).

[3] See Supplemental Tables 6–8 in the online supplemental materials for frequency tables of participants excluded by condition for Studies 1–3

The task-irrelevant race or sky images were displayed below each card or card combination. These images were also pseudo-randomized to ensure the same image did not appear more than twice in succession, but that each image had an equal probability of being displayed across all trials. The nonsocial stimuli consisted of four images of clouds that were gathered from free internet sources. These images were piloted and chosen based on the following criteria: (a) they could not be rendered in any way, (b) they could not contain rain or sun (i.e., only darker or lighter clouds), (c) they could not contain any land elements (i.e., only sky), and (d) they were roughly matched on how cloudy the sky images were. The social stimuli consisted of two White and two Black male faces selected from the Chicago Face Database (Ma et al., 2015), and were equated across various features (e.g., perceived attractiveness, aggressiveness, and age; see Supplemental Tables 2 and 3 in the online supplemental materials).

Training

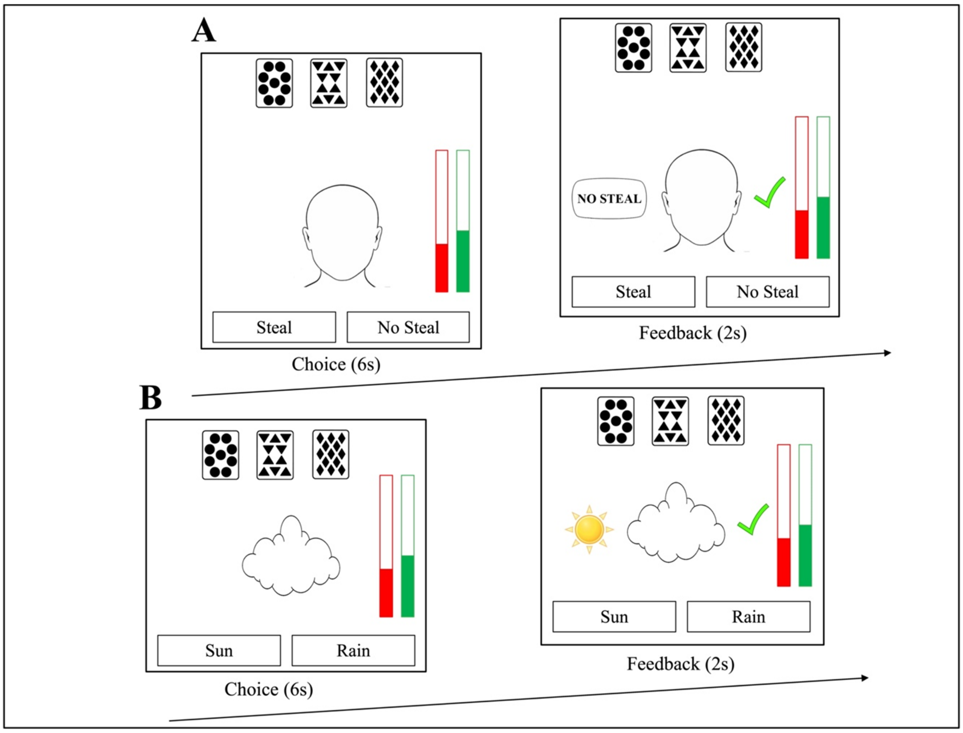
Participants received extensive instructions prior to the task (see p. 3 in the online supplemental materials for the cover story). The cover story included an explanation as to why the stimuli were present during each trial, and that the social and nonsocial images presented below the card combinations were completely orthogonal to the task and nondiagnostic of the outcome, and therefore should not influence their choices. Importantly, participants were repeatedly told that the stimuli should be ignored and that they should instead focus on the cards. Participants were also informed that their choices early in the task would feel like random guesses, but that their performance through trial-by-trial feedback would increase over time. In addition, participants completed a total of six practice trials (Figure 1). Thus, participants did have some exposure to card probabilities prior to the start of the task, which helps to explain why it appears that they are above chance at trial 1. Participants were also required to correctly answer multiple choice comprehension check questions that repeated until a correct response was made (see pp. 3–4 in the online supplemental materials for a list of questions). These questions covered important features of the task to ensure validity and comprehension of the nature of the feedback (i.e., optimal choice vs. actual outcome [see below for more detail]), as well as the stimuli and learning context.

Procedure

After accepting the HIT on mTurk, participants were provided a link to the task. Once entered, they were given informed consent detailing the risks and benefits of the study. Following the training phase, participants were instructed to predict whether the card or card combination would forecast "sun" or "rain" the next day (weather condition) or "steal" or "no steal" the next day (Crime Face condition). Participants made their prediction by using their cursor to click on one of the two options (participants were ineligible to participate in the study if they were using a tablet or phone), which were displayed below the stimuli (the side of the screen where the

Figure 1

*Schematic of the Learning Task in the Crime Face (**A**) and Weather Cloud (**B**) Conditions*



Note. Participants learned the probabilities of the card patterns by selecting their choice and receiving feedback as to if they were correct or incorrect (right side of stimuli) as well as what the actual outcome was on that trial (left side of stimuli). *That the stimuli in this diagram were the same stimuli shown to participants during the practice trials. See the online article for the color version of this figure.

buttons were presented were counterbalanced across participants; see Figure 1). In each trial, one card or card combination was presented on the top of the screen and participants had up to 7 s to make their prediction. Failing to respond within this 7 s time window would automatically start the next trial, and participants were not provided any feedback. Three failed responses in a row would trigger a notification that asked participants if they were still paying attention. Clicking "yes" would resume the task where it left off. If participants did respond within 7 s, they were provided feedback regarding whether they made the optimal choice or not (based on the cumulative probability of that card or card pattern) and what the actual outcome of that trial happened to be (based on the predetermined frequency of that pattern and the cumulative probability; see Supplemental Table 1 in the online supplemental materials). An optimal choice represents any prediction where the card or card patterns are $> 50\%$ of the predicted outcome. If participants made the optimal choice, they were shown a green check mark (correct) on the right side of the stimuli, whereas if participants made the suboptimal choice, they were shown a red × (incorrect).

To monitor overall performance on the task, each correct response corresponded with a one-unit increase in a green bar on the right side of the screen, whereas each incorrect response corresponded with a one-unit increase in a red bar. In addition to seeing feedback about whether or not the optimal choice was made, participants were also provided feedback about the actual outcome of that trial. This feedback was presented in text (weather: sun vs. rain; crime: steal vs. no steal) on the left of the stimuli. Finally, participants were also provided with a progress bar at the bottom of the screen.

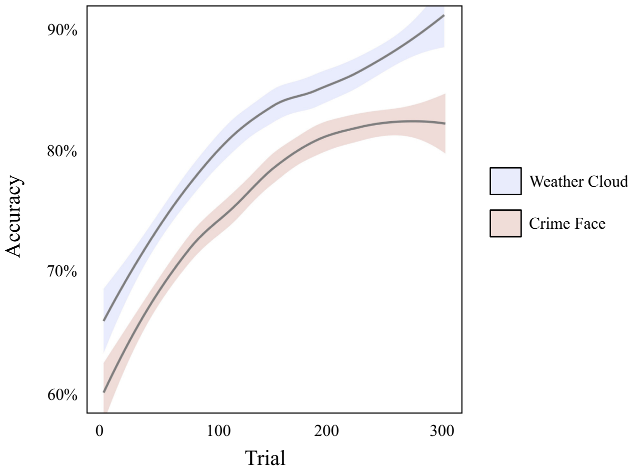
Analysis Plan

To test whether racial associations disrupt learning more than weather associations, we fit a logistic mixed model regressing accuracy (0 = *incorrect*; 1 = *correct*) onto a dummy coded factor representing the fixed effect of condition (0 = *Crime Face [reference group]*; 1 = *weather cloud*), and a variable representing time (i.e., trial). The time variable was scaled and estimated with random slopes for trials within subjects. To examine whether learning trajectories changed over time as a function of the learning environment, we included a time by condition interaction. In addition, to ensure the generalizability of our stimuli, we modeled stimuli and participants as random factors. The full model included four fixed effects (intercept, main effect of trial, main effect of condition, and trial by condition interaction) and three random effects (stimuli, participant, and trial). This mixed model was estimated using the *lme4* package in R (Bates et al., 2015), with *p* values generated with Satterthwaite approximation using the lmerTest package (Kuznetsova et al., 2017).

Results

This model revealed two significant main effects. First, we observed a main effect of the trial ($b = 0.45$, 95% CI [0.35, 0.55], $SE = 0.05$, $z = 8.63$, $OR = 1.5$, $p < .0001$), demonstrating that participants became more accurate as the task progressed. Second, we observed a significant main effect of condition ($b = 0.35$, [0.038, 0.67], $SE = 0.13$, $z = 2.21$, $OR = 1.42$, $p = .026$), demonstrating that participants in the Crime Face conditioned performed worse relative to participants in the weather condition (see Figure 2). We did not observe a significant Trial × Condition interaction ($b = 0.09$, $SE = 0.08$, $z = 1.18$, $p = .23$).

Figure 2
Learning Rates (Loess) as a Function of Condition



Note. The light blue line (light) denotes participants in the Weather Cloud condition, whereas the brown line (dark) denotes participants in the Crime Face condition. *That the error bars denote the standard error of the mean. See the online article for the color version of this figure.

Discussion

Study 1 provides preliminary evidence that social associations about race and crime disrupt learning more than nonsocial associations about the weather. This is an intriguing finding considering that people maintain associations between clouds and weather outcomes just as they maintain racial associations. These findings also raise the question of whether learning decrements are attributable to the mere presence of faces, or whether it was the combination of faces within a context that elicits stereotypical associations that influences learning. People are highly attuned to human faces from the first days of life (Barnett et al., 2021; Frank et al., 2009), and racial outgroup faces in particular have been shown to inadvertently garner greater attention and eye-gaze relative to ingroup faces (Trawalter et al., 2008). Across time, divided attention may contribute to suboptimal learning.

The interpretation of Study 1 results was also limited by a potential confound in the weather stimuli, due to the images of clouds being more strongly associated with the sun than rain (see Supplemental Table 4 in the online supplemental materials). That is, the weather stimuli were not balanced in terms of their associations with the outcomes of sun and rain. In contrast, the race stimuli were balanced in that half the images depicted a Black male face and the other half a White male face. In addition, it is unclear whether learning decrements are attributable to the forecasting context, whereby one kind of prediction (e.g., crime) is simply stranger than another (e.g., weather). Thus, we designed Study 2 to address these confounds and alternative explanations.

Study 2: Separating the Influence of Human Faces From Stereotypes on Learning

Prior research shows that people are highly attuned to human faces and that their presence inadvertently captures attention (Theeuwes &

Van der Stigchel, 2006). Thus, the mere presence of faces may inadvertently garner attention and thereby disrupt optimal learning. This effect is stronger when people are presented with faces of racial outgroups compared to racial ingroups (Trawalter et al., 2008). Study 2 was designed to replicate and extend Study 1 by dissociating learning decrements that may be due to the presence of human faces from stereotypical associations, in addition to addressing potential confounds in the cloud stimuli used in Study 1. To that end, Study 2 included two additional conditions that crossed the stimuli and learning context. If learning decrements are due to the presence of faces alone, then we would expect no difference in learning for participants presented with faces in the context of crime compared to participants presented with faces in the context of weather.

Method

Participants

Participants ($N = 393$) were recruited from the University of California SONA pool and received course credit for completing the study. Participants were once again excluded based on a priori exclusion criterion of 52% accuracy. After applying these exclusion criteria, $N = 373$ participants remained. The sample had a mean age of 20.01 ($SD = 3.32$; 63% self-identified female and 37% self-identified male) and was 43% Asian, 28% Latino, 11% White, 8% Other, and 6% Black.

We conducted a post hoc sensitivity analysis using the *Simr* package (Green & Macleod, 2016) in R which revealed an observed power of 0.91, 95% CI [0.87, 0.93] to detect the main effect of Condition in Study 2, demonstrating sufficient power.

Task Design

The configuration of the paradigm, including the placement of the card or card patterns, the race and weather stimuli, feedback, and the progress bar were all identical to Study 1. The script was once again hosted on biz.nf. One notable change is that Study 2 reduced the total number of trials from 300 to 200 to reduce participant fatigue given the length of the task.

One limitation of Study 1 was that the weather stimuli were, on average, more strongly associated with sun. Study 2 included a new and larger set of weather stimuli that were balanced in terms of their associations with sun and rain. Specifically, 30 images of clouds were piloted in a separate sample ($N = 22$), and eight images were selected for Study 2, four of which were more strongly associated with rain and four that were more strongly associated with the sun (see Supplemental Tables 4 and 5 in the online supplemental materials). The race stimuli were once again selected from the Chicago Face Database, but included the addition of four new faces images from the database to increase our sample of stimuli (see Supplemental Table 3 in the online supplemental materials).

Training

Participants once again received extensive instructions prior to the task to ensure validity and received the same comprehension checks as participants in Study 1.

Procedure

The procedure was similar to Study 1 but included two additional conditions that were designed to control for context and stimulus-driven effects. Specifically, participants were also randomly assigned into a condition where they were asked to (a) predict "sun" or "rain," while seeing images of Black and White faces (Weather Face), or (b) predict "steal" or "no steal," while seeing images of clouds (Crime Clouds). In all, participants were subject to a 2 (context: crime vs. weather) by 2 (stimuli: faces vs. clouds) design.

Analysis Plan

To examine differences in learning across the four conditions, we estimated a logistic mixed model regressing accuracy onto a dummy coded factor for condition (Weather Face condition = 0 [as reference group]), as well as a continuous variable for time. The time variable was scaled and estimated as a random slope for trial within subjects. To further test whether learning decrements varied as a function of time and if the effect of time depended on the learning environment (i.e., condition), the full model included a Time × Condition interaction term. Both stimuli and participants were once again estimated as random factors to increase generalizability.

To test the unique influence of faces from faces embedded within contexts that elicit stereotypical associations, we further conducted pairwise simple contrasts between the crime condition and all other conditions.

Finally, to test whether participant race moderated learning, we also report mixed models that test the interaction of Race × Condition. Given that our sample in Study 2 contained a relatively small proportion of self-identified "White", "Black," and "Other" participants, we report one model where each race category is coded as a separate factor (effects coded with "Other" coded as -1), as well as a second model where "White", "Black," and "Other" are collapsed into a single "Other" factor.[4]

Results

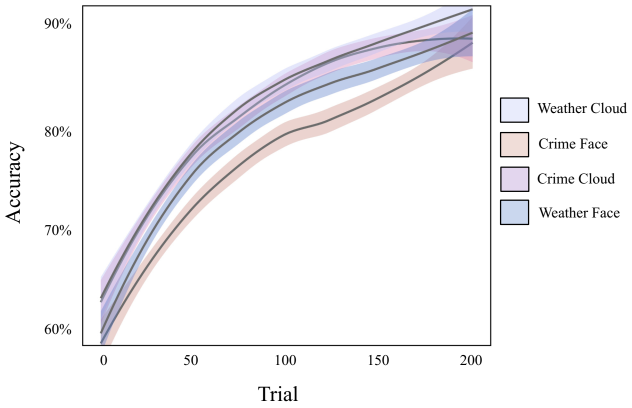
Replicating Study 1 and in line with our preregistered hypothesis, we observed one significant main effect of the condition (Figure 3), demonstrating that participants in the Crime Face condition ($M = 0.77$, $SD = 0.42$) were significantly less accurate relative to participants in the Weather Cloud condition ($M = 0.82$, $SD = 0.38$; $b = -0.36$, 95% CI [-0.57, -0.15], $SE = 0.11$, $z = -3.36$, $OR = 69$, $p = .0007$). Notably, we did not observe a significant difference between participants in the Weather Face and Weather Cloud condition ($b = -0.10$, $SE = 0.107$, $z = -0.94$, $p = .34$), or between participants in the Crime Cloud and Weather Cloud condition ($b = -0.15$, $SE = 0.105$, $z = -1.46$, $p = .14$).

Simple contrasts revealed that participants who made a prediction about crime in the presence of faces learned significantly worse than participants who made predictions about crime in the presence of clouds ($b = -0.25$, $OR = 0.77$, $SE = 0.108$, $z = -2.40$, $p = .01$), and learned marginally worse than participants who made predictions

[4] Individual differences (e.g., motivation to respond without prejudice, social dominance orientation) were also collected in Studies 2 and 3, and these data will be reported in an aggregated analysis later in the current article.

Figure 3

Learning Rates (Loess) as a Function of Condition



Note. The light blue line (light) denotes participants in the Weather Cloud condition, the brown line (dark) denotes participants in the Crime Face condition, the purple line (gray) denotes participants in the Crime Cloud condition, and the dark blue line (dark grey) denotes participants in the Weather Face condition. Error bars denote the standard error of the mean. See the online article for the color version of this figure.

about weather in the presence of faces ($b = -0.20$, $OR = 0.81$, $SE = 0.10$, $z = 1.93$, $p = .053$). To quantify the practical implication of the latter effect size, we calculated the *percent accuracy change* between participants in the Crime Face and Weather Face conditions. To do so, we set the Crime Face condition as the base rate for learning performance (4,286 [incorrect]/(14,028 [correct] + 4,286 [incorrect] = 0.23), and then used the odds ratio representing the difference in learning between the Crime Face and Weather Face (0.81) to calculate the relative difference in error rates between these two conditions (see page 13 in the online supplemental materials for formula). Doing so revealed an incorrect rate of 14.69% in the Weather Face condition compared to 23.40% in the Crime Face condition. In practical terms, the error rate increases more than 50% when participants learn in an environment where faces are stereotypically associated with predicted outcomes relative to when they are not. These results suggest that while the presence of faces does impede new learning, the combination of faces in a context that elicits racial associations with crime leads to larger learning disruptions.

Finally, we tested whether participant race moderated learning and found no significant difference in accuracy across conditions as a function of self-identified race (all $ps > .18$; see Supplemental Table 10 in the online supplemental materials for full model output). These effects did not meaningfully change when race was collapsed into a three-level factor compared to a five-level factor (all $ps > .185$; see Supplemental Table 11 in the online supplemental materials for full model output).

Discussion

Study 2 replicated results from Study 1 and found that learning in the presence of racial associations disrupted learning more than learning in the presence of weather associations. By crossing the stimuli and the learning context, Study 2 also dissociated the unique effect of faces on learning. Doing so revealed that error rates in the Crime Face condition increased by over 50% compared to error rates in the Weather Face condition, suggesting that the presence of racial stereotypes in the learning context may additionally disrupt learning over the mere presence of faces. These results highlight the impact of stereotypes on probabilistic category learning and updating and suggest that even when people are explicitly told that the racial stimuli have no association with the outcome, they still negatively impact learning.

One notable difference between Studies 1 and 2 pertains to the sample demographics. Study 1 was conducted with a majority White sample (73% White), whilst Study 2 was conducted with a majority minority sample (11% White). Importantly, participant race did not moderate learning in Study 2, which is consistent with past work showing that stereotypes are culturally shared and influence cognition across myriad racial and ethnic groups (Axt et al., 2014).

Study 3: Do Learning Decrements Extend to Positively Valenced Stereotypes?

Study 2 provided converging evidence that social contexts that elicit negative racial associations disrupt learning more than nonsocial contexts that elicit weather associations. Notably, learning in the Crime Face condition involves one of the most salient and harmful stereotypes in the United States of Black men as criminal (Welch, 2007). Thus, learning may be uniquely disrupted in this and other contexts related to intergroup threat (Chang et al., 2016), relative to contexts that elicit stereotypes that people perceive to be more innocuous. This begs the question of whether similar learning decrements would arise in contexts that elicit positively valenced stereotypical associations, for instance, stereotypes about Black men as athletic.

It is important to note that we refer to positively valenced stereotypes as those that are subjectively perceived to be favorable characteristics of a group, rather than those that are experienced as positive by members of the stereotyped group, or that lead to positive outcomes for the stereotyped group. Indeed, research finds that targets of positive stereotypes experience similar emotional responses—dislike, resentment, and negativity—as targets of negative stereotypes (Czopp, 2008; Siy & Cheryan, 2016). Moreover, people who tend to endorse negative stereotypes about Black Americans also tend to exhibit stronger stereotypes endorsing Black as athletic (Kurdi et al, 2019; see also Kay et al., 2013), suggesting that these two classes of stereotypes may similarly influence intrapersonal and interpersonal psychological processes. Due to their "complementary" nature, positive stereotypes are often treated as innocuous or even flattering (Bergsieker et al., 2012), but they have been shown to be as harmful as negative stereotypes, especially due to their pervasiveness and general acceptance in the social milieu (Czopp et al., 2015; Devine & Elliot, 1995). As such, people are more likely to endorse positive relative to negative stereotypes, and people who publicly endorse positive stereotypes are seen as less prejudiced than people who publicly endorse negative stereotypes (Mae & Carlston, 2005).

Given that people may be less concerned about appearing prejudiced in the context of positive stereotypes, one possibility is that they may exert less of an influence on learning. On the other hand, any stereotypical association may continue to disrupt learning because of its salience, and because people may want to avoid endorsing any kind of race-based stereotype. Study 3 thus sought to extend Study 2 and test whether learning decrements extend to contexts that elicit positive stereotypes about Black men as athletic.

Method

Participants

Participants ($N = 220$) were recruited from the University of California SONA pool and received course credit for completing the study. Participants were once again excluded based on a priori exclusion criterion of 52% accuracy. After applying these exclusion criteria, $N = 206$ participants remained. The sample had a mean age of 19.43 ($SD = 1.86$; 62% self-identified female and 38% self-identified male) and was 45% Asian, 32% Latino, 10% White, 7% Other, and 3% Black.

Task Design

The structure of the paradigm, including the placement of the card or card patterns, stimuli, feedback, and progress bar were all identical to Study 2, and the script was once again hosted on biz.nf. The social stimuli were also identical to Study 2 in both conditions and participants received extensive instructions prior to the task and received the same comprehension checks.

Procedure

The procedure was similar to Study 2, however, this time participants were either randomly assigned to a condition that elicited a positive stereotypical association (i.e., Black and athletic) or a context that elicited negative stereotypical associations (i.e., Black and criminal). Specifically, participants were either randomly assigned to a condition where they were asked to predict a "steal" or "no steal" outcome from card combinations, or a condition where they were asked to predict a "touchdown" or "no touchdown" outcome from card combinations. In the positive stereotype condition, optimal feedback was once again presented via a green check mark for correct prediction and a red × for incorrect predictions. The actual outcome was presented in text ("touchdown" vs. "no touchdown"). The negative stereotype condition was identical to Study 2. After the learning task, participants responded to the same individual differences measures as participants in Study 2. The stimuli (Black and White male faces) were identical to that in Study 2.

Analysis Plan

To test whether positively or negatively valenced stereotypes differentially disrupt learning, we fit a logistic mixed model similar to that in Study 1. Specifically, we regressed accuracy (0 = *incorrect*; 1 = *correct*) onto a dummy coded factor representing the fixed effect of condition (0 = *Crime Face [as reference group]*; 1 = *Athletic*), and a variable representing time (i.e., trial). The time variable was once again scaled and estimated with random slopes for trials within subjects. The full model included four fixed effects (intercept, main effect of trial, condition, and a Trial × Condition interaction term) and three random factors (stimuli, participants, and trial within subjects). To test whether participant race moderated learning, we also report a model that includes an interaction between self-identified race and condition. Each race category was coded as a separate factor (effects coded with "Other" coded as −1).

Finally, to test the likelihood that the observed null effect of the condition was true, we estimated a multilevel Bayesian logistic model using the *brms* package in R (Bürkner, 2017) with uninformative priors. One benefit of this approach is that it provides a more appropriate test for the probability of obtaining a null effect based on the proportion of the posterior distribution of the parameter estimate that falls within a range that would be considered negligible (i.e., the region of practical equivalence; see Kruschke, 2010, 2018). This model contained the exact same fixed and random effects as before.

Results

In the first model, we examined differences in learning across conditions and found no significant difference in learning between participants in the Crime Face and Athletic condition ($b = 0.048$, $SE = 0.11$, $z = 0.44$, $p = .66$; see Figure 4). Likewise, we found no evidence that accuracy differed across the condition as a function of participant race (all $ps > .14$; see Supplemental Table 12 in the online supplemental materials for full model output).

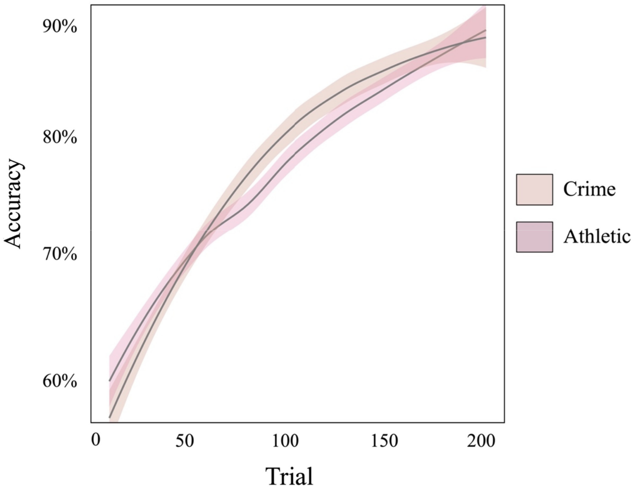
In the second model, we tested the likelihood that the null was true and demonstrated that participants in the Athletic condition had a 54.57% probability of being less accurate than participants in the Crime Face condition (median of the parameter estimate = −0.01, 89% CI [−0.23, 0.19]). More importantly, 100% of the posterior distribution fell inside the region of practical equivalence (see Supplemental Figure 2 in the online supplemental materials for credible intervals), demonstrating a high likelihood of the null effect of the condition. This null effect is intriguing because it suggests that although people are often less concerned about expressing positively valenced stereotypes, their presence may still impact learning similar to negatively valenced stereotypes.

Discussion

Study 3 examined whether learning decrements extend to social contexts that elicit positively valenced stereotypes about Black men as athletic. Interestingly, we found no difference in learning rates between the positive and negative stereotype conditions.

Figure 4

Learning Rates (Loess) as a Function of Condition



Note. The brown line (dark) denotes participants in the Crime condition, whereas the pink line (light) denostes participants in the Athletic condition. Error bars denote the standard error of the mean. See the online article for the color version of this figure.

One possible explanation for why we do not find a difference in learning between the Athletic and Crime condition may be that the saliency of both stereotypes leads people to inadvertently incorporate them into their predictions, thereby disrupting learning on a trial-by-trial basis by weighting predictions toward stereotypically congruent outcomes. Alternatively, the mere presence of a stereotype, and the added cognitive effort of monitoring one's responses to avoid appearing prejudiced, may increase cognitive load and disrupt learning on a more cumulative basis.

Aggregated Analysis: Testing Competing Mechanisms Underlying Learning Decrements

We find converging evidence that task-irrelevant racial associations disrupt probabilistic learning above and beyond weather associations (Studies 1 and 2). Notably, we found no difference in learning between the positively (Black and athletic) and negatively (Black and criminal) valenced conditions (Study 3). Next, we interrogate two plausible underlying mechanisms based on past research.

First, stereotypes may disrupt learning via first-order stereotype application or inhibition, which occurs at the trial level. Consistent with research demonstrating that people often automatically apply stereotypes during decision-making tasks (Eberhardt et al., 2004), stereotypes may similarly disrupt learning by increasing people's propensity to respond in a stereotype-congruent manner (i.e., stereotype application). For instance, participants may be more likely to predict "steal" when a Black face is present compared to when a White face is present. Conversely, stereotypes may disrupt learning if people actively try to suppress their influence over predictions (i.e., stereotype inhibition). For instance, in an attempt to appear unprejudiced, participants may try to control their responses and actively suppress "steal" predictions when a Black face is present. Given that the faces are orthogonal to the cumulative probability of the card patterns, any choice that is influenced by faces should lead to suboptimal predictions on a trial-by-trial basis.

Second, stereotypes may disrupt learning via second-order effects that accumulate over time due to the cognitive effort elicited by the presence of a salient stereotype. This hypothesis is motivated by research suggesting that regulating the expression of prejudice during interracial interactions (Richeson et al., 2003) and processing task-irrelevant race information is cognitively taxing (Rubien-Thomas et al., 2021). Moreover, given that the stereotypes presented in the current experiments relate to racial inequities, we speculate that the influence of stereotypes on learning may be greatest for individuals who are more motivated by egalitarian values (e.g., racial equity). Racial stereotypes may be particularly salient for individuals who are more conscious of their negative effects (Johns et al., 2008), and who actively work to reduce their own racial biases (Devine et al., 2002). If stereotypes disrupt probabilistic learning by increasing the cognitive effort needed to monitor responses, then individuals higher (vs. lower) in internal motivations to respond without prejudice should perform worse on the current task. Such individuals are more burdened by the cognitive effort required to monitor their responses.[5]

Motivations to respond without prejudice are composed of separable and largely independent underlying components: internal (IMS) and external motivations (EMS; Devine et al., 2002; Plant & Devine, 1998). IMS reflect the implications of appearing prejudiced to one's sense of self (e.g., "I attempt to act in nonprejudiced ways toward Black people because it is personally important to me"), and thus assess personal motivations that are not encumbered by external environmental factors. EMS focus on external social pressures that arise from normative pressures to not appear prejudiced (e.g., "I attempt to appear not prejudiced toward Black people in order to avoid disapproval from others"). Importantly, both IMS and EMS may moderate learning via different mechanisms, yet research has yet to examine if and how IMS and EMS influence learning. To ensure sufficient power and generalizability, we aggregated data across conditions and conducted an aggregated analysis (Eisenhauer, 2021; Goh et al., 2016).

Method

Participants

We selected all participants in the Crime Face ($N = 204$) and athletic conditions ($N = 93$) across Studies 2 and 3.[6] Study 1 was not included in this analysis because we did not collect individual differences for participants in that sample. This subsample had a mean age of 19.43 ($SD = 1.86$; 62% self-identified female and 38% self-identified male), and was 45% Asian, 32% Latino, 10% White, 7% Other, and 3% Black.

Self-Report Measures[7]

IMS to Respond Without Prejudice

A five-item questionnaire that asses a person's IMS to respond without prejudice (Plant & Devine, 1998). The scale demonstrated strong reliability in the current sample ($\omega = 0.85$).

EMS to Respond Without Prejudice

A five-item questionnaire that asses a person's EMS to respond without prejudice (Plant & Devine, 1998). The scale demonstrated moderate to strong reliability in the current sample ($\omega = 0.79$).

Social Dominance Orientation

A 16-item questionnaire that asses a person's preference for inequality among social groups (Pratto et al., 1994). The scale demonstrated strong reliability in the current sample ($\omega = 0.81$).

[5] We preregistered a series of reinforcement learning (RL) models to test competing mechanisms underlying learning. These RL models were consistent with our behavioral findings, such that we found no evidence for "first-order" effects on learning, and instead found evidence for "second-order" effects in the RL models. We felt that these analyses do not provide greater nuance to the behavioral evidence and therefore they are reported in the online supplemental materials (pp. 19–25).

[6] Note that the Athletic condition only applies to Study 3.

[7] Several other individual difference measures were collected for exploratory purposes. These include the extraversion facet from the Big Five Inventory (Soto & John, 2017), the honesty-humility facet from the HEXACO model of personality (Ashton & Lee, 2007; Ashton et al., 2014), intergroup anxiety (Stephan & Stephan, 1985), and the General Intergroup Contact and Quantity and Quality scale (Islam & Hewstone, 1993). See pages 7 and 8 as well as Supplemental Figure 1 in the online supplemental materials for more details.

Analysis Plan

First-Order Effects: Stereotype Application Versus Stereotype Inhibition

To test whether learning decrements are due to stereotype application or inhibition, we fit a series of mixed models interrogating the influence of various trial-level features on responses separately for participants in the Crime Face ($N = 204$) and Touchdown ($N = 93$) conditions. This included regressing accuracy onto a fixed effect of stimuli representing the race of the face present on a given trial (dummy coded with Black faces as the reference group), which would help to determine whether accuracy was systematically influenced by the race of the face present (see Model 1 in Supplemental Tables 13 and 14 in the online supplemental materials). We also regressed a congruency variable onto reaction time and accuracy (see Models 2 and 3 in Supplemental Tables 13 and 14 in the online supplemental materials). The congruency variable was set to 1 when a participant's prediction was stereotypically congruent (e.g., Black and steal), and 0 (reference group) when a participant's prediction was stereotypically incongruent (e.g., White and steal). This model would determine whether participants' predictions were systematically influenced by stereotype congruent compared to stereotype incongruent trials, for instance via less accurate or more rapid responses on stereotypically congruent trials. We also regressed the congruency variable onto the fixed effect of the predictive weight of the card patterns. Predictive weight was modeled both as a categorical factor (effects coded with 0.5 as reference), which would indicate differences based on high vs. low predictive trials (see Model 4 in Supplemental Tables 13 and 14 in the online supplemental materials), and as an absolute deviation from 0.5, providing a continuous measure of ambiguity (see Model 5 in Supplemental Tables 13 and 14 in the online supplemental materials). This model would determine whether participants were more likely to rely on the stereotype when the cards were less determinant of the outcome. Finally, we regressed the congruency variable onto individual differences that may moderate stereotype application or inhibition (e.g., Social Dominance Orientation; Pratto et al., 1994; IMS; see Models 6 and 7 in Supplemental Tables 13 and 14 in the online supplemental materials). All models included a random factor for stimuli, as well as a time (trial) variable that was scaled and estimated as a random slope for trial within subjects. This time variable was first estimated as a covariate, and then subsequently as a moderator. Each model was tested both for participants in the Crime Face (Supplemental Table 13 in the online supplemental materials) and Athletic conditions (Supplemental Table 14 in the online supplemental materials). All models examining behavior in the Crime condition included a dummy coded covariate to control for the study.

Second-Order Effects: Cumulative Learning Decrements Over Time

If learning is not disrupted on a trial-by-trial basis, but instead accumulates over the course of the task due to tax cognitive functioning, then participants who are higher on IMS to respond without prejudice should perform worse relative to those lower. Moreover, this added cognitive effort may result in longer response times for those who score the highest on motivations to respond without prejudice.[8] To test these questions, we first estimated separate mixed models regressing accuracy onto a measure of IMS and EMS to respond without prejudice (see Models 1 and 2 in Supplemental Tables 15 and 16 in the online supplemental materials). We also tested a model that included both IMS and EMS to examine the unique effects of IMS above and beyond EMS, as well as the interactive effects of the two (see Model 3 in Supplemental Tables 15 and 16 in the online supplemental materials). Further, to test whether IMS to respond without prejudice were associated with longer response times, we estimated a model that regressed log(reaction time) onto IMS, while controlling for EMS (see Model 4 in Supplemental Tables 15 and 16 in the online supplemental materials).[9] Finally, we also tested whether response times differed as a function of the race of the stimuli present on each trial by regressing log (reaction time) onto a dummy coded factor for stimuli (Black = 0 [reference group]; White = 1: see Model 5 in Supplemental Tables 15 and 16 in the online supplemental materials). All continuous variables were scaled, and each model tested the interaction of the fixed effects with time. Once again, all models included random factors for stimuli, participants, and trials within participants. Each model was tested both in the Crime Face condition (Supplemental Table 15 in the online supplemental materials) and Athletic condition (Supplemental Table 16 in the online supplemental materials).

Results

First-Order Effects in the Crime Face Condition

Across all models, we found no evidence that learning was disrupted on a trial-by-trial basis for participants in the Crime Face condition. For instance, there was no difference in accuracy as a function of the race of the stimuli ($b = -0.002$, $SE = 0.03$, $z = -0.071$, $p = .94$), no difference in accuracy or reaction time for stereotype-congruent versus incongruent trials ($b = 0.037$, $SE = 0.02$, $z = 1.13$, $p = .16$, and $b = -0.003$, $SE = 0.004$, $t = 0.064$, $p = .94$, respectively), nor were there any differences in the predictive weights of the card patterns for stereotype-congruent versus incongruent trials, regardless of whether ($ps > .11$) or as an absolute difference from 0.5 ($b = 0.013$, $SE = 0.01$, $z = 0.1.27$, $p = .20$). Moreover, individual differences associated with increased stereotype application (i.e., Social Dominance Orientation [SDO]) and inhibition (i.e., IMS) did not predict stereotype-congruent responding ($\beta = -0.011$, $SE = 0.01$, $z = -1.02$, $p = .30$, and $\beta = -0.006$, $SE = 0.011$, $z = -0.55$, $p = .54$, respectively). Taken together, these results suggest that learning decrements are not attributable to first-order stereotype application or inhibition occurring at the trial level (see Supplemental Table 13 in the online supplemental materials for all model outputs).

First-Order Effects in the Athletic Condition

Across all models, we likewise found no evidence that learning was disrupted on a trial-by-trial basis for participants in the

[8] We attempted to collect a measure of stereotype misperception (Krieglmeyer & Sherman, 2012) but due to a coding error this data was sadly unusable and therefore not analyzed. As such, our preregistered hypothesis for IMS (H2) could not be directly tested. Instead, we test the effects of IMS on learning over time, which is consistent with the preregistered hypothesis.

[9] We also estimated a model for participants in Study 2 to examine whether the effect of IMS was specific to the Crime Face condition, or whether it extended to any context with faces (e.g., Weather Face). This model demonstrated that the effect of IMS on learning was unique to the Crime Face condition (Supplemental Figure 4 in the online supplemental materials).

Athletic condition. For instance, there was no difference in accuracy as a function of the race of the stimuli ($b = 0.01$, $SE = 0.04$, $z = 0.33$, $p = .73$), no difference in accuracy or reaction time for stereotype-congruent versus incongruent trials ($b = 0.02$, $SE = 0.04$, $z = 0.51$, $p = .607$, and $b = -0.001$, $SE = 0.006$, $t = -0.24$, $p = .80$, respectively), nor were there any differences in the predictive weights of the card patterns for stereotype-congruent versus incongruent trials, regardless of whether card pattern weight was modeled as categorical (all $ps > .25$) or as an absolute difference from 0.5 ($b = -0.01$, $SE = 0.01$, $z = -0.709$, $p = .47$). Moreover, individual differences associated with increased stereotype application (i.e., SDO) and inhibition (i.e., IMS) did not predict stereotype-congruent responding ($\beta = -0.02$, $SE = 0.01$, $z = 1.47$, $p = .14$, and $\beta = 0.01$, $SE = 0.01$, $z = 0.78$, $p = .43$, respectively). Taken together, these results suggest that learning decrements are also not attributable to first-order stereotype application or inhibition occurring at the trial level in the Athletic condition (see Supplemental Table 14 in the online supplemental materials for all model outputs).

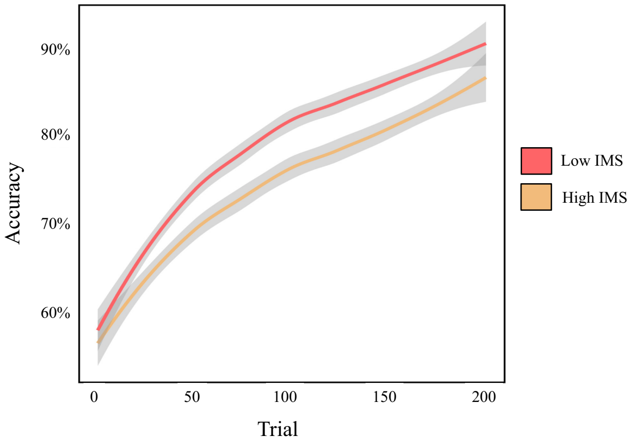
Second-Order Effects in the Crime Face Condition

If the presence of stereotypes does not disrupt learning on a trial-by-trial basis, an alternative explanation is that the presence of stereotypes is cognitively demanding which accumulates to disrupt learning. To test this, we examined whether IMS or EMS respond without prejudice moderated (a) task performance and (b) response times. Regarding performance, the first model revealed a significant main effect of IMS, and a significant Trial × IMS interaction ($\beta = -0.13$, 95% *CI* [-0.24, -0.01], $SE = 0.06$, $z = -2.23$, $OR = 0.87$, $p = .024$, and $\beta = -0.06$, [-0.12, -0.01], $SE = 0.027$, $z = -2.40$, $OR = 0.93$, $p = .016$, respectively), demonstrating that participants higher on IMS performed worse compared to those lower on IMS over time (see Figure 5).

Regarding EMS, we observed a marginal Trial × EMS interaction on accuracy ($\beta = -0.05$, [-0.10, 0.002], $SE = 0.02$, $z = -1.87$, $OR = 0.95$, $p = .061$), note however, the 95% CI for the interaction overlaps with 0, and therefore this effect should be interpreted with caution.

Figure 5

Learning Rates (Loess) as a Function of Internal Motivation to Respond Without Prejudice



Note. IMS were median split for visual purposes only. Red line (dark) denotes low IMS, whereas yellow line (light) denotes high IMS. Error bars denote the standard error of the mean. See the online article for the color version of this figure.

Importantly, both the main effect of IMS and the IMS × Trial interaction remained significant after controlling for EMS ($\beta = -0.14$, 95% CI [-0.25, -0.027], $SE = 0.05$, $z = -2.44$, $OR = 0.86$, $p = .014$, and $\beta = -0.06$, [-0.122, -0.01], $SE = 0.02$, $z = -2.49$, $OR = 0.93$, $p = .012$, respectively), whereas EMS was no longer significant after controlling for IMS ($\beta = 0.001$, $SE = 0.009$, $z = 0.121$, $p = .90$). We did not observe a significant IMS × EMS interaction ($\beta = -0.02$, $SE = 0.048$, $z = -0.62$, $p = .53$), nor a significant three-way interaction between trial, IMS, and EMS ($\beta = -0.009$, $SE = 0.023$, $z = -0.41$, $p = .68$).

Finally, examining the relationship between IMS and response times, as well as stimuli and response times for participants in the Crime Face condition revealed: (a) a significant Trial × IMS interaction ($\beta = 0.02$, 95% CI [0.006, 0.0034], $SE = 0.007$, $t = 2.86$, $p = .004$), suggesting that participants higher on IMS were slower to respond over time, and (b) a significant main effect of stimuli ($\beta = -0.009$, [-0.018, -0.0001], $SE = 0.004$, $t = -1.99$, $p = .045$), suggesting that participants tended to respond slower on trials when Black faces were present compared to when White faces were present. Together, these findings help to bolster the claim that increased cognitive load for participants higher on IMS accumulates over the course of the task to disrupt learning (Supplemental Table 15 in the online supplemental materials).

Second-Order Effects in the Athletic Condition

We did not find any evidence for second-order effects in the Athletic condition. For instance, both IMS and EMS were not associated with accuracy ($\beta = 0.03$, $SE = 0.08$, $z = 0.43$, $p = .662$, and $\beta = 0.03$, $SE = 0.08$, $z = 0.43$, $p = .66$, respectively). Likewise, both IMS and the race of the stimuli were not associated with response times ($\beta = 0.02$, $SE = 0.02$, $t = 1.02$, $p = .31$, and $\beta = -0.01$, $SE = 0.006$, $t = -0.227$, $p = .11$, respectively: Supplemental Table 16 in the online supplemental materials).

Taken together, these findings suggest that individuals who are more *internally* motivated to respond without prejudice learn significantly worse when learning in the presence of negatively valenced stereotypes, even when controlling for EMS, whereas we found no such evidence for participants in the positively valenced stereotype condition.

Discussion

These aggregated analyses tested competing hypotheses regarding whether learning decrements in the social conditions were attributable to first-order effects due to stereotype application or inhibition, or whether learning was disrupted by second-order effects due to increased cognitive load. Notably, although disruptions may occur via first- or second-order effects, these are not mutually exclusive. For instance, individuals high on IMS to respond without prejudice may suffer learning decrements due to increased cognitive load, but may also suffer learning decrements due to stereotype inhibition, as these motivations may manifest in increased cognitive load due to self-monitoring of responses and/or increase stereotype inhibition.

A series of analyses that regressed both accuracy and choices onto different trial-level features (e.g., race of the face present, weights of the cards, etc.) failed to provide evidence for first-order stereotype application or inhibition in either the Crime Face or Athletic conditions. That is, participants were no more or less likely to make trial-by-trial predictions that aligned or misaligned with the present stereotype. Instead, we find that individuals higher on IMS to respond without prejudice suffered significantly worse learning decrements relative to those lower in the Crime Face condition. These results suggest that the presence of negative stereotypes is more salient for individuals with higher IMS to respond without prejudice as they activate egalitarian values and associated thoughts (Johns et al., 2008). As such, the presence of negative stereotypes may disrupt learning due to second-order effects via increased cognitive load extraneous to the learning task. This is consistent with the assumption that the bottleneck for acquiring new knowledge is limited by working memory, and that cognitive effort extraneous to a learning task can disrupt optimal learning (Sweller et al., 2019).

Notably, IMS only moderated task performance for participants in the Crime Face condition but had no influence on learning in the Athletic condition, despite both environments leading to similar learning rates. This raises new questions about how preexisting stereotypes disrupt learning, and whether positive and negative stereotypes disrupt learning via unique or parallel mechanisms. For instance, learning disruptions in the Athletic condition may also be attributable to cognitive load due to the presence of a sticky association. However, in this context where people are often less concerned about appearing prejudiced, the disruption may instead be due to the attention required to follow task instructions and ignore the presence of a salient stereotype.

General Discussion

The current work extends decades of research on the destructive role of stereotypes on decision-making by examining whether and how they disrupt processes that precede decision-making, such as how people learn probabilistic associations in the first place. Across experiments, participants who were presented with task-irrelevant *social* associations (i.e., predicting crime in the presence of Black and White male faces) learned card pattern probabilities worse than participants who were presented with task-irrelevant *nonsocial* associations (i.e., predicting weather in the presence of darker and lighter cloud formations), despite the fact that participants in both contexts were repeatedly told that the stimuli were irrelevant to the learning context and should therefore be ignored. This is remarkable considering that people maintain strong nonsocial associations throughout their life, such as associations between clouds and rain, just as they maintain strong social associations, such as stereotypes that connect race and crime.

These findings bear several important theoretical implications. First, probabilistic category learning is integral to how people learn about the world through trial and error (Dayan & Niv, 2008; Gobet et al., 2001; Lee et al., 2012), yet extant research has not yet examined how preexisting associations interfere with learning new associations. The current research interrogates a fundamental question about the boundary conditions of probabilistic category learning, demonstrating that preexisting associations can interfere with new learning. As such, the current work expands models of human learning by beginning to describe the cognitive processes by which preexisting associations get in the way of new learning. These findings open up new avenues for future research on the different classes of preexisting associations and their associated mechanisms that constrain learning, which we describe in the discussion below.

Second, different classes of preexisting associations that people hold may impact new learning. For instance, people maintain both social (e.g., Black and criminal) and nonsocial (e.g., dark clouds with rain) preexisting associations, but do they exert a similar influence over how people learn new associations? Our results suggest that there is something unique about preexisting social associations about race that disrupt learning, as these contexts led to significantly worse learning rates compared to a nonsocial context. Throughout daily life, people encounter many situations in which they must learn to predict discrete outcomes that are imperfectly correlated with predictive cues. Moreover, these situations are often embedded within contexts where task-irrelevant social and nonsocial cues compete for attention. These findings are the first to compare the effects of social versus nonsocial associations on new learning and highlight the insidious nature of racial stereotypes by demonstrating that they disrupt the basic processes that people use to understand and make predictions about the world.

A third important implication of this work pertains to *how* stereotypes disrupt new category learning. The influence of stereotypes may be underpinned by at least two plausible mechanisms: first-order stereotype application or inhibition at the trial level, and second-order cognitive load disruptions that accumulate across trials. We tested these competing hypotheses in a number of ways and failed to find any evidence for any first-order disruption effects. In other words, we found no evidence that task-irrelevant stereotypes increased or decreased the propensity for any particular stereotypically congruent or incongruent response in either the Crime Face or Athletic conditions. Instead, we find evidence that learning decrements in the Crime Face context were attributable to second-order effects, as people higher in IMS performed significantly worse than those lower in IMS. These results extend past research that finds that interracial interactions deplete cognitive resources that then disrupt performance on subsequent cognitive control tasks (Richeson et al., 2003). Specifically, we demonstrate that the cognitive demands of regulating responses that may appear prejudiced can disrupt learning as it unfolds in real time.

One potential explanation for why social associations disrupt learning via second-order effects is that they are generally more affect-laden (Gawronski & Bodenhausen, 2006; Stephan & Stephan, 1985; Zelazo & Cunningham, 2007), and are therefore more salient during decision-making (Finucane et al., 2000; Todd et al., 2012). For instance, the presence of racial associations may elicit anticipated negative affect based on expected outcomes, such as people's fear of appearing prejudiced (Mellers et al., 1997). Notably, it is plausible that fear of appearing prejudiced may disrupt learning even in the absence of racial stereotypes. For instance, people who are led to believe that their behavior on a probabilistic learning task will reveal racial bias may still suffer learning decrements as they are burdened by the excess cognitive load of monitoring their responses. One way to interrogate a purely affective mechanism is to examine whether learning decrements extend to nonsocial affective associations. For instance, people have prepared fear responses to certain classes of nonsocial stimuli (e.g., snakes, spiders; Öhman & Soares, 1994), much like aversive learning that is facilitated by certain kinds of social stimuli that are perceived as threatening (e.g., Black faces; Olsson et al., 2005). If affect contributes to learning decrements, then people should also learn cue-outcome associations more poorly when they are paired with irrelevant affectively laden nonsocial associations (e.g., predicting "bite" when spiders are present). Another way to test affect as a mechanism may be to measure arousal via indirect measures (e.g., eye-tracking and galvanic skin response; Proudfoot et al., 2016; Shi et al., 2007) during learning. In contrast to behavioral responses that people can monitor and control, people rarely have access to or the ability to control their own physiological responses (Goff et al., 2008). Thus, indirect methods that measure arousal on a moment-to-moment basis may help to elucidate the conditions under which affective associations disrupt learning.

Consistent with the hypothesis that affect may moderate learning via second-order effects due to anticipated negative affective responses, we find that participants with higher internal motivations to respond without prejudice (IMS) suffered learning decrements more than those with lower internal motivations. Prior research has found that individuals higher on IMS are often more successful at regulating prejudicial responses (Plant & Devine, 1998; Schlauch et al., 2009), and are less likely to show implicit forms of racial bias (e.g., associating negative words with Black faces and positive words with White faces; Devine et al., 2002). Conversely, people higher on external motivations to respond without prejudice (EMS) can successfully disguise prejudice on self-report measures (Plant & Devine, 1998), but often fail to inhibit more difficult to control racial bias (Butz & Plant, 2009). One explanation for this discrepancy is that, whereas EMS tracks societal norms and obligations, IMS tracks personal values and goals which facilitate responses that are consistent with that values (Devine et al., 2002; Ryan & Connell, 1989). The success of IMS in regulating bias is often driven by the negative feelings elicited when people violate their personal values (Plant & Devine, 1998). Thus, despite explicit instructions that faces are not predictive of outcomes, participants higher on IMS may nevertheless experience negative arousal in anticipation that their responses may signal prejudice. As a result, individuals higher on IMS may monitor their responses, which increases extraneous cognitive effort and impedes learning.

The current findings also raise new questions about the influence of IMS on active versus passive learning contexts. The current findings suggest that participants higher on IMS are more sensitive to stereotypical cues when actively learning new probabilistic associations. At first blush, these findings diverge from research showing that participants higher on IMS are less sensitive to race during impression formation (Li et al., 2016). One reason that our findings may diverge is that the tasks used in these studies are asking different things of participants. Specifically, the current work employs an active learning task in which participants make predictions and receive feedback in the presence of stereotype-eliciting faces that likely elicit conflict monitoring between task demands and internal goals to respond without prejudice (Dignath et al., 2020). Impression formation tasks differ from active learning tasks in that participants passively learn face/word pairs in the absence of predictions and feedback, and as such do not typically evoke the same responses (Mende-Siedlecki et al., 2013). Greater conflict monitoring for higher IMS individuals may tax cognitive functioning and disrupt learning in active versus passive learning tasks. These findings highlight a unique outcome of IMS previously not reported in the literature by demonstrating that IMS may inadvertently "get in the way" of optimal active learning as people attempt to monitor behaviors that may signal prejudice even when they are unrelated to the learning context.

A parallel explanation for learning decrements may be that social associations simply capture more attention and are therefore stickier and harder to ignore than nonsocial associations. For instance, we find no difference in learning between participants presented with positive stereotypes (e.g., Black and athletic) relative to participants presented with negative stereotypes (e.g., Black and criminal). Although motivations to respond without prejudice did not moderate learning rates in the positive stereotype condition, both positive and negative learning contexts led to similar task performance. This suggests that stereotypes might simply be stickier in general, regardless of affect. Indeed, research demonstrates that people are more likely to endorse stereotypes that they experience as less affectively laden, such as stereotypes about gender and age (Czopp et al., 2015; Devine & Elliot, 1995; Mae & Carlston, 2005). This raises new questions about the role of affect and attention elicited by preexisting associations in probabilistic learning, and whether they are context-dependent and dissociable based on the nature of the preexisting associations. For instance, learning decrements caused by task-irrelevant positive associations may likewise be attributable to the greater cognitive effort required to follow task instructions and ignore a salient association, regardless of whether it is affect-laden. One potential route to test whether such disruptions are due to the stickiness of social associations is to test whether this phenomenon generalizes to less affectively charged social associations (e.g., status, gender, and age). Doing so may shed light on whether specific learning mechanisms are context-dependent, such as anticipated negative affect for fear of social repercussion in one context versus the cognitive demand needed to ignore stereotypical associations in another.

The current findings also contribute to a growing body of research on the role of multiple memory systems in social cognition (Amodio, 2019). Converging evidence shows that humans have several types of memory systems (e.g., procedural, habitual, semantic, and declarative) that are independently mediated by different brain systems, but that often work in parallel (Cabeza & Moscovitch, 2013). The extent to which one system (e.g., declarative) is privileged over another (e.g., procedural) during learning can fluctuate and compete across time based on cognitive load (Foerde et al., 2006). For instance, prior research finds that participants employ different learning strategies when performing the WPT under stress (cold pressor) compared to control (Schwabe & Wolf, 2013). The stress manipulation modulated the engagement of different memory systems, such that stressed participants used less declarative (i.e., hippocampus-based) and more procedural (i.e., striatum-based) learning systems compared to control. This shift toward procedural learning under stress also rescued task performance, but disrupted task performance when participants tried to engage in declarative learning (i.e., explicitly learning the card probabilities). The notion that multiple memory systems simultaneously mediate learning may help to explain how second-order effects disrupt learning. For instance, analogous to the influence of stress on learning, the presence of stereotypes may engage more procedural learning mechanisms. However, the deliberate act of self-monitoring responses may engage declarative learning, thereby suppressing optimal learning (aggregated analysis). While participants may overcome this competition later in the task, the suppressed learning rates early on may accumulate to disrupt overall learning. Future research should explicitly test whether social versus nonsocial learning contexts engage different learning mechanisms and associated memory systems.

Limitations

The current findings and their implications should be considered within the limitations of each study. For instance, we did not include a pure control condition where participants learned in the absence of any additional stimuli. Future research that includes a pure control condition would help to rule out alternative explanations, such as whether the presence of clouds improved performance or the presence of faces reduced it. However, it is highly unlikely that performance would be improved in the presence of clouds given that the stimuli were orthogonal to the learning task and thus could not improve performance. Moreover, Study 2 was designed to rule out a number of alternative explanations by employing a fully crossed design that controlled for context and stimulus-driven effects. For instance, if the effects are driven by the forecasting context, whereby one kind of prediction (e.g., crime) is stranger than the other (e.g., weather), then we would expect a significant difference between participants predicting crime in the presence of clouds from participants predicting weather in the presence of faces, which we did not find. Likewise, if the effects are driven by the stimuli (i.e., the presence of faces or clouds) then we would expect a significant difference between the Weather Cloud and Weather Face conditions, which we also did not find. Our goal in the current research was to directly compare the influence of social to nonsocial preexisting associations on new learning to determine which exerts a stronger influence on the learning context. A pure control condition would help to shed further light on whether learning in the presence of nonsocial preexisting associations disrupts performance compared to learning in the absence of any preexisting associations.

Relatedly, the current findings are limited to contexts involving preexisting associations about weather and race, and as such, we cannot generalize across all social and nonsocial learning contexts. An alternative explanation is that learning decrements are attributable to social learning environments where faces are present, even when they are devoid of stereotypical associations. For instance, it may be that social associations are simply more sticky than nonsocial associations about weather and that these learning decrements would persist in a context where participants make social forecasts (e.g., “fired” or “not fired” from a job) in the presence of White and Black faces. The goal of the current work was to test whether one class of distractors—such as racial associations—are more disruptive than other distractors, rather than to demonstrate all different contexts under which these effects occur. We believe that the current findings are generative and invite future research to examine the interplay between different preexisting associations and new learning contexts.

Another potential limitation of the current research is that it is unclear whether preexisting associations disrupt the acquisition of new learning or rather people’s performance on the task. In other words, because learning takes place gradually over time, it is unclear if taxed cognitive functioning impairs people’s ability to learn the new associations or merely their ability to produce the learned skill and make the correct response. Past work suggests that performance—rather than learning—may be disrupted during a probabilistic task when accompanied by a secondary cognitive task (Foerde et al., 2007). However, in this past work, participants were asked to optimize performance on both cognitive tasks simultaneously, whereas participants in the current work were explicitly told to ignore any distractors and instead focus on the card or card patterns. More research is needed to test the extent to which task-irrelevant associations interfere with the acquisition or expression of newly learned probabilistic associations.

Conclusions

People constantly learn new associations by making iterative predictions and observing feedback in order to navigate their environments successfully. An important challenge to the learning process is learning when to ignore task-irrelevant prior associations that may “get in the way” of optimal learning. The current findings highlight the insidious and sticky nature of stereotypes on learning: Even when stereotypic associations are explicitly task-irrelevant, they influence how people learn, think, and behave. In order to achieve sustainable social progress, we need to better understand how stereotypes affect the basic processes we use for understanding and making predictions about the world, to pinpoint the exact mechanisms and contexts in which they operate. The current work is a step in this direction.

Context of the Research

Our research group with members from San Francisco State University and the University of California, Riverside investigated whether associations of a social nature may be particularly difficult to ignore, and thus impinge on learning more than nonsocial associations. Building on decades of research on stereotyping and prejudice, we provide evidence that the mere presence of stereotypes disrupts probabilistic learning and updating even when they are orthogonal to task demands. In future work, we aim to generalize these findings across different learning environments.

References

Allidina, S., & Cunningham, W. A. (2021). Avoidance begets avoidance: A computational account of negative stereotype persistence. *Journal of Experimental Psychology: General*, *150*(10), 2078–2099. https://doi.org/10.1037/xge0001037

Amodio, D. M. (2019). Social cognition 2.0: An interactive memory systems account. *Trends in Cognitive Sciences*, *23*(1), 21–33. https://doi.org/10.1016/j.tics.2018.10.002

Arrow, K. J. (1998). What has economics to say about racial discrimination? *Journal of Economic Perspectives*, *12*(2), 91–100. https://doi.org/10.1257/jep.12.2.91

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*(2), 150–166. https://doi.org/10.1177/1088868306294907

Ashton, M. C., Lee, K., & de Vries, R. E. (2014). The HEXACO honesty-humility, agreeableness, and emotionality factors: A review of research and theory. *Personality and Social Psychology Review*, *18*(2), 139–152. https://doi.org/10.1177/1088868314523838

Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2014). The rules of implicit evaluation by race, religion, and age. *Psychological Science*, *25*(9), 1804–1815. https://doi.org/10.1177/0956797614543801

Barnett, B. O., Brooks, J. A., & Freeman, J. B. (2021). Stereotypes bias face perception via orbitofrontal–fusiform cortical interaction. *Social Cognitive* *and Affective Neuroscience*, *16*(3), 302–314. https://doi.org/10.1093/scan/nsaa165

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bergsieker, H. B., Leslie, L. M., Constantine, V. S., & Fiske, S. T. (2012). Stereotyping by omission: Eliminate the negative, accentuate the positive. *Journal of Personality and Social Psychology*, *102*(6), 1214–1238. https://doi.org/10.1037/a0027717

Blair, I., 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

Blind, G. D., & Lottanti von Mandach, S. (2021). Of pride and prejudice: Agent learning under sticky and persistent stereotype. *Journal of Economic Interaction and Coordination*, *16*(2), 381–410. https://doi.org/10.1007/s11403-020-00307-0

Blinder, S., Ford, R., & Ivarsflaten, E. (2013). The better angels of our nature: How the antiprejudice norm affects policy and party preferences in Great Britain and Germany. *American Journal of Political Science*, *57*(4), 841–857. https://doi.org/10.1111/ajps.12030

Bodenhausen, G. V., & Wyer, R. S. (1985). Effects of stereotypes on decision making and information-processing strategies. *Journal of Personality and Social Psychology*, *48*(2), 267–282. https://doi.org/10.1037/0022-3514.48.2.267

Brewer, M. B. (2001). The many faces of social identity: Implications for political psychology. *Political Psychology*, *22*(1), 115–125. https://doi.org/10.1111/0162-895x.00229

Brewer, M. B., & Pierce, K. P. (2005). Social identity complexity and outgroup tolerance. *Personality and Social Psychology Bulletin*, *31*(3), 428–437. https://doi.org/10.1177/0146167204271710

Bürkner, P. C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Butz, D. A., & Plant, E. A. (2009). Prejudice control and interracial relations: The role of motivation to respond without prejudice. *Journal of Personality*, *77*(5), 1311–1341. https://doi.org/10.1111/j.1467-6494.2009.00583.x

Cabeza, R., & Moscovitch, M. (2013). Memory systems, processing modes, and components: Functional neuroimaging evidence. *Perspectives on Psychological Science*, *8*(1), 49–55. https://doi.org/10.1177/1745691612469033

Chang, L. W., Krosch, A. R., & Cikara, M. (2016). Effects of intergroup threat on mind, brain, and behavior. *Current Opinion in Psychology*, *11*, 69–73. https://doi.org/10.1016/j.copsyc.2016.06.004

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*(3), 235–253. https://doi.org/10.1037/0096-3445.120.3.235

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*(6), 1314–1329. https://doi.org/10.1037/0022-3514.83.6.1314

Correll, J., Wittenbrink, B., Crawford, M. T., & Sadler, M. S. (2015). Stereotypic vision: How stereotypes disambiguate visual stimuli. *Journal of Personality and Social Psychology*, *108*(2), 219–233. https://doi.org/10.1037/pspa0000015

Czopp, A. M. (2008). When is a compliment not a compliment? Evaluating expressions of positive stereotypes. *Journal of Experimental Social Psychology*, *44*(2), 413–420. https://doi.org/10.1016/j.jesp.2006.12.007

Czopp, A. M., Kay, A. C., & Cheryan, S. (2015). Positive stereotypes are pervasive and powerful. *Perspectives on Psychological Science*, *10*(4), 451–463. https://doi.org/10.1177/1745691615588091

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*(2), 185–96. https://doi.org/10.1016/j.conb.2008.08.003

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5–18. https://doi.org/10.1037/0022-3514.56.1.5

Devine, P. G., & Elliot, A. J. (1995). Are racial stereotypes really fading? The Princeton trilogy revisited. *Personality and Social Psychology Bulletin*, *21*(11), 1139–1150. https://doi.org/10.1177/01461672952111002

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology*, *82*(5), 835–848. https://doi.org/10.1037/0022-3514.82.5.835

Dignath, D., Eder, A. B., Steinhauser, M., & Kiesel, A. (2020). Conflict monitoring and the affective-signaling hypothesis—An integrative review. *Psychonomic Bulletin and Review*, *27*(2), 193–216. https://doi.org/10.3758/s13423-019-01668-9

Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing black: Race, crime, and visual processing. *Journal of Personality and Social Psychology*, *87*(6), 876–893. https://doi.org/10.1037/0022-3514.87.6.876

Eisenhauer, J. (2021). Meta-analysis and mega-analysis: A simple introduction. *Teaching Statistics*, *43*(1), 21–27. https://doi.org/10.1111/test.12242

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making*, *13*(1), 1–17. https://doi.org/10.1002/(SICI)1099-0771(200001/03)13:1<1::AID-BDM333>3.0.CO;2-S

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). Elsevier. https://doi.org/10.1016/S0065-2601(08)60317-2

Foerde, K., Knowlton, B. J., & Poldrack, R. A. (2006). Modulation of competing memory systems by distraction. *Proceedings of the National Academy of Sciences*, *103*(31), 11778–11783. https://doi.org/10.1073/pnas.0602659103

Foerde, K., Poldrack, R. A., & Knowlton, B. J. (2007). Secondary-task effects on classification learning. *Memory and Cognition*, *35*(5), 864–874. https://doi.org/10.3758/BF03193461

Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition*, *110*(2), 160–170. https://doi.org/10.1016/j.cognition.2008.11.010

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. https://doi.org/10.1037/0033-2909.132.5.692

Gluck, M. A. (2002). How do people solve the "weather prediction" task?: Individual variability in strategies for probabilistic category learning. *Learning and Memory*, *9*(6), 408–418. https://doi.org/10.1101/lm.45202

Gobet, F., Lane, P., Croker, S., Cheng, P., Jones, G., Oliver, I., & Pine, J. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*(6), 236–243. https://doi.org/10.1016/S1364-6613(00)01662-4

Goff, P. A., Eberhardt, J. L., Williams, M. J., & Jackson, M. C. (2008). Not yet human: Implicit knowledge, historical dehumanization, and contemporary consequences. *Journal of Personality and Social Psychology*, *94*(2), 292–306. https://doi.org/10.1037/0022-3514.94.2.292

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*(10), 535–549. https://doi.org/10.1111/spc3.12267

Green, P., & Macleod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Hogg, M., Terry, D., & White, K. (1995). A tale of two theories: A critical comparison of identity theory with social identity theory. *Social Psychology Quarterly*, *58*(4), 255–269. https://doi.org/10.2307/2787127

Islam, M. R., & Hewstone, M. (1993). Dimensions of contact as predictors of intergroup anxiety, perceived out-group variability, and out-group attitude: An integrative model. *Personality and Social Psychology Bulletin*, *19*(6), 700–710. https://doi.org/10.1177/0146167293196005

Johns, M., Cullum, J., Smith, T., & Freng, S. (2008). Internal motivation to respond without prejudice and automatic egalitarian goal activation. *Journal of Experimental Social Psychology*, *44*(6), 1514–1519. https://doi.org/10.1016/j.jesp.2008.07.003

Kay, A. C., Day, M. V., Zanna, M. P., & Nussbaum, A. D. (2013). The insidious (and ironic) effects of positive stereotypes. *Journal of Experimental Social Psychology*, *49*(2), 287–291. https://doi.org/10.1016/j.jesp.2012.11.003

Kirsch, I., Lynn, S. J., Vigorito, M., & Miller, R. R. (2004). The role of cognition in classical and operant conditioning. *Journal of Clinical Psychology*, *60*(4), 369–392. https://doi.org/10.1002/jclp.10251

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning and Memory*, *1*(2), 106–120. https://pubmed.ncbi.nlm.nih.gov/10467589/

Krieglmeyer, R., & Sherman, J. W. (2012). Disentangling stereotype activation and stereotype application in the stereotype misperception task. *Journal of Personality and Social Psychology*, *103*(2), 205–224. https://doi.org/10.1037/a0028764

Kruschke, J. K. (2010). Bayesian data analysis. *WIRES Cognitive Science*, *1*(5), 658–676. https://doi.org/10.1002/wcs.72

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1083–1119. https://doi.org/10.1037/0278-7393.25.5.1083

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, *20*(7), 512–534. https://doi.org/10.1016/j.tics.2016.05.004

Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences*, *116*(13), 5862–5871. https://doi.org/10.1073/pnas.1820240116

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience*, *35*(1), 287–308. https://doi.org/10.1146/annurev-neuro-062111-150512

Li, T., Cardenas-Iniguez, C., Correll, J., & Cloutier, J. (2016). The impact of motivation on race-based impression formation. *NeuroImage*, *124*(Pt A), 1–7. https://doi.org/10.1016/j.neuroimage.2015.08.035

Lindström, B., Selbing, I., Molapour, T., & Olsson, A. (2014). Racial bias shapes social reinforcement learning. *Psychological Science*, *25*(3), 711–719. https://doi.org/10.1177/0956797613514093

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. https://doi.org/10.3758/s13428-014-0532-5

Mae, L., & Carlston, D. E. (2005). Hoist on your own petard: When prejudiced remarks are recognized and backfire on speakers. *Journal of Experimental Social Psychology*, *41*(3), 240–255. https://doi.org/10.1016/j.jesp.2004.06.011

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science*, *8*(6), 423–429. https://doi.org/10.1111/j.1467-9280.1997.tb00455.x

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience*, *8*(6), 623–631. https://doi.org/10.1093/scan/nss040

Minhas, R., & Walsh, D. (2021). The role of prejudicial stereotypes in the formation of suspicion: An examination of operational procedures in stop and search practices. *International Journal of Police Science and Management*, *23*(3), 293–305. https://doi.org/10.1177/14613557211016499

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015

Öhman, A., & Soares, J. J. F. (1994). "Unconscious anxiety": Phobic responses to masked stimuli. *Journal of Abnormal Psychology*, *103*(2), 231–240. https://doi.org/10.1037/0021-843X.103.2.231

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), 785–787. https://doi.org/10.1126/science.1113551

Onorato, R., & Turner, J. (2004). Fluidity in the self-concept: The shift from personal to social identity. *European Journal of Social Psychology*, *34*(3), 257–278. https://doi.org/10.1002/ejsp.195

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, *57*, 401–421. https://doi.org/10.1146/annurev.psych.57.102904.190127

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, *81*(2), 181–192. https://doi.org/10.1037/0022-3514.81.2.181

Plant, A., & Devine, P. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811–832. https://doi.org/10.1037/0022-3514.75.3.811

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, *67*(4), 741–763. https://doi.org/10.1037/0022-3514.67.4.741

Proudfoot, J. G., Jenkins, J. L., Burgoon, J. K., & Nunamaker, J. F. (2016). More than meets the eye: How oculometric behaviors evolve over the course of automated deception detection interactions. *Journal of Management Information Systems*, *33*(2), 332–360. https://doi.org/10.1080/07421222.2016.1205929

Richeson, J. A., Baird, A. A., Gordon, H. L., Heatherton, T. F., Wyland, C. L., Trawalter, S., & Shelton, J. N. (2003). An fMRI investigation of the impact of interracial contact on executive function. *Nature Neuroscience*, *6*(12), 1323–1328. https://doi.org/10.1038/nn1156

Richeson, J. A., & Shelton, J. N. (2003). When prejudice does not pay: Effects of interracial contact on executive function. *Psychological Science*, *14*(3), 287–290. https://doi.org/10.1111/1467-9280.03437

Richeson, J. A., & Trawalter, S. (2005). Why do interracial interactions impair executive function? A resource depletion account. *Journal of Personality and Social Psychology*, *88*(6), 934–947. https://doi.org/10.1037/0022-3514.88.6.934

Rubien-Thomas, E., Berrian, N., Cervera, A., Nardos, B., Cohen, A. O., Lowrey, A., Daumeyer, N. M., Camp, N. P., Hughes, B. L., Eberhardt, J. L., Taylor-Thompson, K. A., Fair, D. A., Richeson, J. A., & Casey, B. J. (2021). Processing of task-irrelevant race information is associated with diminished cognitive control in Black and White individuals. *Cognitive, Affective, and Behavioral Neuroscience*, *21*(3), 625–638. https://doi.org/10.3758/s13415-021-00896-8

Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology*, *57*(5), 749–761. https://doi.org/10.1037//0022-3514.57.5.749

Schiller, D., Freeman, J. B., Mitchell, J. P., Uleman, J. S., & Phelps, E. A. (2009). A neural mechanism of first impressions. *Nature Neuroscience*, *12*(4), 508–514. https://doi.org/10.1038/nn.2278

Schlauch, R. C., Lang, A. R., Plant, E. A., Christensen, R., & Donohue, K. F. (2009). Effect of alcohol on race-biased responding: The moderating role of internal and external motivations to respond without prejudice. *Journal of Studies on Alcohol and Drugs*, *70*(3), 328–336. https://doi.org/10.15288/jsad.2009.70.328

Schwabe, L., & Wolf, O. T. (2013). Stress and multiple memory systems: From 'thinking' to 'doing'. *Trends in Cognitive Sciences*, *17*(2), 60–68. https://doi.org/10.1016/j.tics.2012.12.001

Shelton, J. N., & Richeson, J. A. (2005). Intergroup contact and pluralistic ignorance. *Journal of Personality and Social Psychology*, *88*(1), 91–107. https://doi.org/10.1037/0022-3514.88.1.91

Shi, Y., Ruiz, N., Taib, R., Choi, E., & Chen, F. (2007). Galvanic skin response (GSR) as an index of cognitive load. In *CHI EA'07: CHI'07 extended abstracts on human factors in computing systems* (pp. 2651–2656). Association for Computing Machinery. https://doi.org/10.1145/1240866.1241057

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. https://doi.org/10.1037/h0042769

Siy, J. O., & Cheryan, S. (2016). Prejudice masquerading as praise: The negative echo of positive stereotypes. *Personality and Social Psychology Bulletin*, *42*(7), 941–954. https://doi.org/10.1177/0146167216649605

Soto, C. J., & John, O. P. (2017). The next big five inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Staddon, J. E. R., & Cerutti, D. T. (2003). Operant conditioning. *Annual Review of Psychology*, *54*(1), 115–144. https://doi.org/10.1146/annurev.psych.54.101601.145124

Stephan, W. G., & Stephan, C. W. (1985). Intergroup anxiety. *Journal of Social Issues*, *41*(3), 157–175. https://doi.org/10.1111/j.1540-4560.1985.tb01134.x

Stillerman, B., Lindström, B., Schultner, D., Hackel, L., Hagen, D., Jostmann, N., & Amodio, D. (2020). *Internalization of societal stereotypes as individual prejudice*. PsyArXiv. https://psyarxiv.com/mwztc/

Sweller, J. (2011). Cognitive Load Theory. In J. P. Mestre, & B. H. Ross (Eds.), *Psychology of learning and motivation—advances in research and theory* (Vol. 55, pp. 37–76). Elsevier. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Theeuwes, J., & Van der Stigchel, S. (2006). Faces capture attention: Evidence from inhibition of return. *Visual Cognition*, *13*(6), 657–665. https://doi.org/10.1080/13506280500410949

Todd, R. M., Cunningham, W. A., Anderson, A. K., & Thompson, E. (2012). Affect-biased attention as emotion regulation. *Trends in Cognitive Sciences*, *16*(7), 365–372. https://doi.org/10.1016/j.tics.2012.06.003

Trawalter, S., Todd, A. R., Baird, A. A., & Richeson, J. A. (2008). Attending to threat: Race based patterns of selective attention. *Journal of Experimental Social Psychology*, *44*(5), 1322–1327. https://doi.org/10.1016/j.jesp.2008.03.006

Welch, K. (2007). Black criminal stereotypes and racial profiling. *Journal of Contemporary Criminal Justice*, *23*(3), 276–288. https://doi.org/10.1177/1043986207306870

Willer, D., Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1989). Rediscovering the social group: A self-categorization theory. *Contemporary Sociology*, *18*(4), 645–646. https://doi.org/10.2307/2073157

Zelazo, P. D., & Cunningham, W. A. (2007). Executive function: Mechanisms underlying emotion regulation. In J. J. Gross (Ed.), *Handbook of emotion regulation* (pp. 135–158). The Guilford Press.

Received January 26, 2022
Revision received October 13, 2022
Accepted October 27, 2022 ■