# Effects of moral stereotypes on the formation and persistence of group preferences

Inga K. Rösler[1], Isabel Kerber[2], and David M. Amodio[1]

[1]University of Amsterdam, [2]Humboldt University of Berlin

Funding: This research was supported by a grant from the Netherlands Organisation for Scientific Research to DMA (VICI 016.185.058).

Preregistration: Hypotheses, sample sizes, exclusion criteria, and analysis plans for both studies were preregistered (Study 1: https://aspredicted.org/FYB_WPX; Study 2: https://aspredicted.org/PRG_VEX). Materials, data, and analysis scripts are publicly available at OSF: https://osf.io/7kpn4/

Correspondence regarding this article should be addressed to Inga Rösler (i.k.rosler@uva.nl) or David Amodio (d.m.amodio@uva.nl), Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129, REC G, 1001 NK Amsterdam, NL.

## Abstract

Do stereotypes have a stronger and more persistent effect on impressions when they are moral in tone? In two experiments (*N* = 187), participants interacted with members of two groups in an interactive social reward learning task. Prior to the task, participants were exposed to positive or negative group stereotypes that were moral or nonmoral in content. Although players from each group were, on average, equally likely to provide reward feedback, participants formed choice preferences for players from positively-stereotyped over negatively-stereotyped groups. Importantly, this effect was stronger and more resistant to change when stereotypes contained moral content. Computational modeling indicated that moral stereotypes induced more extreme initial reward expectancies and influenced how reward associations were updated over time. Additionally, moral stereotypes generalized more strongly to subsequent evaluations of novel group members, suggesting that the biasing effect of moral stereotypes on learning contributed to group-level prejudice.

(145 words)

*Keywords*: morality, stereotypes, reinforcement, learning, computational, prejudice

## Effects of moral stereotypes on the formation and persistence of group preferences

Stereotypes of minority ethnic groups are often moral in tone (e.g., Abele-Brehm et al., 2020; Fiske et al., 2002): in the United States, Latinos are often associated with crime (Welch et al., 2011), African Americans are stereotyped as hostile (Devine & Elliot, 1995), and Muslim men as ‘violent terrorists’ (e.g., Jackson, 2010). While ethnic stereotypes also often include nonmoral traits, such as unintelligence or laziness, moralized characteristics, which convey a sense of right and wrong, are especially likely to fuel intergroup intolerance, conflict, and harm (Cuddy et al., 2008; Mooijman & Hoover, 2018; Papakyriakopoulos & Zuckerman, 2021; Skitka, 2010). In this research, we examined whether moral stereotypes, as compared with stereotypes lacking a moral component, have stronger effects on group-based impression formation and whether such impressions are more resistant to updating.

### Moral stereotypes and impression formation

Stereotypes are culturally-held descriptions of social groups that typically refer to their members’ traits (e.g., criminal) or life circumstances (e.g., poor). A major function of stereotypes is to guide impressions of individual group members (Amodio & Cikara, 2021; Fiske, 1998). These impressions, in turn, can help a perceiver characterize a group member’s behavior and predict their future actions. However, because stereotypes are generalizations that often misrepresent or exaggerate a group's attributes, stereotypes can bias impressions and thus contribute to prejudice and discrimination (Allport, 1954; Hilton & Von Hippel, 1996).

Although the content and functions of moral stereotypes have been examined in much prior research (Graham et al., 2012; Nicolas & Fiske, 2023; Phalet & Poppe, 1997), the effects of moral stereotypes on impression formation have not been investigated. However, research in the impression formation literature shows that moral information profoundly affects a perceiver's trait inferences (Abele-Brehm et al., 2020; Brambilla et al., 2011, 2021; Brambilla & Leach, 2014). Moral information is perceived to reveal one's 'true,' essential character and thus used to understand people's intentions and behaviors (Cuddy et al., 2008; Goodwin et al., 2014; Heiphetz, 2019; Strohminger & Nichols, 2014). For instance, information about a person's honesty was found to weigh most heavily in perceivers' expectations of their cooperative behavior in social dilemmas (Van Lange & Kuhlman, 1994), and information about a job candidate's morality more strongly influenced perceivers' impressions and decisions than more relevant information about the candidate's competence (Luttrell et al., 2022). Across a variety of contexts, moral information has been shown to have an especially potent effect on impressions, perceived intentions, and expectations (Luttrell et al., 2016; Day et al.,2014; Van Bavel et al., 2012; Wojciszke et al., 1998).

How might moral information affect impression formation in the context of social stereotypes? To the extent moral content in stereotypes guides impressions of group members in the same way that it guides individual impressions (Abele-Brehm et al., 2020; Brambilla et al., 2011, 2021; Brambilla & Leach, 2014), moral stereotypes should also more strongly influence first impressions of group members relatively to nonmoral stereotypes.

## Moral stereotypes and impression updating

In addition to guiding initial impressions of group members, stereotypes impede the degree to which impressions are updated in response to new information about a group member (Allport, 1954; Fiske, 1998). That is, stereotypes shape the construal of new individuating information about a group member to fit stereotypic expectancies (Darley &

Gross, 1983; Kunda & Sherman-Williams, 1993). Because moral information is seen as especially diagnostic (Brambilla et al., 2019; Goodwin, 2015; Mende-Siedlecki et al., 2013; Wojciszke et al., 1998), moral impressions are especially resistant to change (Luttrell et al., 2020, 2022; Reeder & Coovert, 1986; Skowronski & Carlston, 1992). These findings suggest that in the context of stereotypes, a moralized stereotype may impede the updating of an initial stereotype-based impressions even after repeated stereotype-disconfirming experiences with a group member.

Recent computational modeling research suggests that stereotypes impede the updating of impressions by altering the construal of new information (Schultner et al., 2024; Traast et al., in press). Using social instrumental learning tasks, adapted from prior reinforcement learning paradigms (e.g., Frank et al., 2004; Hackel et al., 2022), these studies investigated how stereotypical information is continuously updated across repeated interactions with group members. In Schultner et al. (2024), participants were presented with stereotype descriptions of two groups—one positive and one negative—and then interacted with individual members of each group in a social reward task. Although members of each group behaved identically in the task, on average, thereby disconfirming the stereotypes, participants' choices reflected a persistent preference for members of the positively-stereotyped group. Computational modeling revealed that stereotypes influenced these preferences by setting initial expectancies of group members and then biasing how they learned from new information. Given the amplified effect of moral traits on impressions shown in prior research (e.g., Abele-Brehm et al., 2020; Brambilla et al., 2021), it is possible that moral content may enhance the constraining effect of stereotypes on updating relative to nonmoral stereotypes.

## Research Overview

We propose that moral stereotypes have a stronger and more persistent effect on impressions of group members, compared with nonmoral stereotypes (e.g., concerning competence; Wojcizke et al., 1998). In two experiments, we tested the hypothesis that stereotypes with moral (vs. nonmoral) content would induce more extreme initial expectancies and impede impression updating in response to stereotype-inconsistent information. We tested these hypotheses using a social reinforcement learning paradigm, in combination with behavioral analysis and computational modeling (Schultner et al., 2024, Traast et al., in press), which allowed us to assess impressions across repeated interaction and examined underlying learning mechanisms. Hypotheses, exclusion criteria for participants, outliers, and sample size were preregistered for both studies before data collection and can be found together with data, code, and materials at [redacted]; deviations and any analyses not included in preregistration are noted. Approval was obtained from the local Ethics Review Board. All studies, measures, manipulations, and data/participant exclusions are reported in the manuscript or its Supplementary Material.

## Study 1

In Study 1, participants interacted with members of two groups: one described with positive stereotypes and the other with negative stereotypes. In one between-subjects condition, stereotype descriptions included moral content; in the other condition, stereotypes did not include moral content. We predicted that exposure to positive and negative moral stereotypes, relative to nonmoral stereotypes, would induce (a) more extreme initial reward expectancies and (b) attenuated updating of preferences for group members.

## Method

### *Participants*

Eighty US-based participants completed the study on CloudResearch in exchange for \$5.00 and an additional performance-based bonus (\$0.00-\$2.50). This sample size (N=80) was preregistered and based on prior research using a similar task (Schultner et al., 2024). The self-identified race/ethnicity of the sample was 78.30% White/Caucasian, 7.25% Asian, 5.80% African American, and 1.45% Hispanic, with 1.45% indicating 'Other' and 5.80% who did not indicate their race/ethnicity. Following preregistered exclusion criteria, we excluded one participant who showed non-compliant behavior (i.e., responding too fast on all but one trial in the test phase) and ten participants who failed to reach a 50% learning criterion for extreme reward rates (30%, 70%) from analysis (Schultner et al., 2024).[1] The final sample included 69 participants ($M_{\text{age}}$ = 42.35 years, $SD_{\text{age}}$ = 13.67, 32 females, 37 males). Sensitivity power analysis in G*Power indicated that for N = 69 the minimum detectable effect size is $d$ = 0.11 ($\alpha$ = 0.05, power = 0.80).

### *Design and Procedure*

To investigate whether moral (vs. nonmoral) stereotypes bias and impair reward learning from group members, we used an adapted version of a validated probabilistic reward reinforcement learning task (Frank et al., 2004; Schultner et al., 2024; Hackel et al., 2022). The experimental design included mixed factors: 2 (stereotype valence: positive vs. negative;

[1] Although exclusion based on below-50% accuracy was preregistered, the preregistration omitted that this applied to extreme (30 vs. 70%) reward pairs, which provide a clear learning signal; 40 vs. 60% pairs, by contrast, represent near-chance probabilities, and are thus less diagnostic for this learning criterion.

within-subjects) x 2 (reward rate of group members: 70%, 60%, 40%, and 30%; within-subjects) x 2 (stereotype morality: morality vs. competence; between-subjects).

Following consent, participants learned they would engage in an interactive social reward learning task, which entailed interacting with eight players belonging to two distinct social groups. Participants were then presented with stereotype descriptions of each group which contained moral or nonmoral content that was either positive or negative in valence. Despite these group descriptions, participants were told that individual group members varied in their tendency to yield reward and thus they should attend to the behavior of each player.

Participants then completed the social reward reinforcement learning task, presented to them as an interactive social reward learning game. Participants were told that the other players had taken part in a previous experiment where they were asked to give or withhold money in a series of choices, and that in the present study, participants would play with these participants and receive feedback based on their past behavior.

Participants were told that previous players would be represented by avatars, ostensibly to protect their identities. To indicate players' group memberships, these avatars differed in hair color, eye color, and color of clothing (as in Schultner et al., 2024). The two groups were labeled "Group A" and "Group B," and one group was always positively stereotyped and the other negatively stereotyped. Avatar features and group labels were counterbalanced between participants. Participants played with either all-male or all-female avatars, randomized across participants, to control for potential gender effects.

Participants were instructed to learn individual players' reward rates in order to maximize their cash earnings. On each trial, participants chose the player with whom they wished to interact and then received feedback on whether the chosen player rewarded them with a point. Points were converted to a cash bonus at the end of the session. Importantly, although reward rates differed between individual players, reward rates were equated between the two groups. Thus, any group-level difference in choice preference would reflect the influence of the stereotype as opposed to group members' actual behaviors.

**Stereotype descriptions**. Stereotype descriptions were modeled after previous research (Kunst et al., 2017; Leach et al., 2007; Pratto et al., 2013; Wojciszke, 2005) and referred to both societal-level and stable individual group member-level attributes (see SI for descriptions). Moral stereotypes referred to morality-related attributes of a society (e.g., low or high governmental corruption) and described group members as immoral, untrustworthy, dishonest, and unfair (negative stereotype condition) or moral, trustworthy, honest, and fair (positive stereotype condition). Nonmoral stereotypes referred to competence-related attributes of society (e.g., a high- or low-performing educational system) and described group members as incompetent, unsuccessful, unintelligent, and unambitious (negative stereotype condition) or competent, successful, intelligent, and ambitious (positive stereotype condition).

It is notable that because of the heightened diagnosticity of moral traits, they may be perceived as more extreme in valence than nonmoral traits (Brambilla et al., 2019; Cone & Ferguson, 2015; Reeder & Coovert, 1986; Skowronski & Carlston, 1987; Wojciszke et al., 1998) and thus may have asymmetric effect on impression updating (Mende-Siedlecki et al., 2013). This issue has been addressed statistically in prior research, such that the enhanced effects of morally-based attitudes on judgments, relative to nonmoral attitudes, remained after adjusting for attitude strength (Luttrell et al., 2016; Skitka et al., 2005). To address this potential confound in the current work, we assessed valence extremity ratings of these stereotype descriptions in a separate sample (*N*=100; see SI).[2] Results indicated that negative moral (*M* = 1.66, *SD* = .89) and nonmoral (*M* = 1.90, *SD* = 1.05) stereotypes did not differ in valence, *t*(95.46) = 1.23, *p* = .223. However, positive moral (*M* = 6.56, *SD* = .64) and nonmoral (*M* = 6.18, *SD* = 1.04) stereotypes did differ slightly, t(81.58) = -2.19, *p* = .031. Thus, while valence extremity did not pose a confound for negative stereotype descriptions, the possibility remained that any effect of moral content in positive stereotypes could be due in part to a valence difference. To further rule out this potential confound, the valence ratings for these descriptions were used as covariates in tests of our main hypothesis to statistically adjust for any potential effect of valence extremity (e.g., for positive stereotypes).

**Categorization task**. Prior to completing the main learning task, participants completed a classification task where they categorized both players and stereotype attributes as belonging to either Group A or Group B. This task ensured that participants learned to associate both the individual players and the stereotype descriptions with the appropriate group labels. The task included 16 trials, and accuracy feedback was given following each response.

**Learning task.** The learning task included a training phase and a test phase. The *training phase* consisted of 160 trials in which participants could learn about each player through repeated interaction and feedback. Participants were instructed that on each round they would be presented with two players, one from each group, and choose one to interact with. The participant's goal was to choose the player that would give them a point. After each choice, the participant received immediate reward feedback from the chosen player (+1 or 0 points). It was

[2] This validation study was descriptive and not preregistered.

emphasized that despite belonging to different groups, individual players would vary in their tendency to give points (i.e., their reward rate). During this training phase, pairs of players presented on each trial had fixed complementary reward rates (i.e., 30%-70%, 40%-60%, 60%-40%, 70%-30%).

Upon completing the training phase, participants took a short break and then began the *test phase*. The test phase, which included 96 trials, was designed to assess the reward-based associations acquired during training. Participants again viewed pairs of players and were instructed to choose the player most likely to give points. However, choice pairs in the test phase included all possible combinations of Group A and Group B members, which permitted a fine-grained assessment of the reward-based associations participants formed for each player. Although no feedback was given during the test phase, to prevent new learning, participants received a cash bonus for choosing more rewarding players.

**Results**

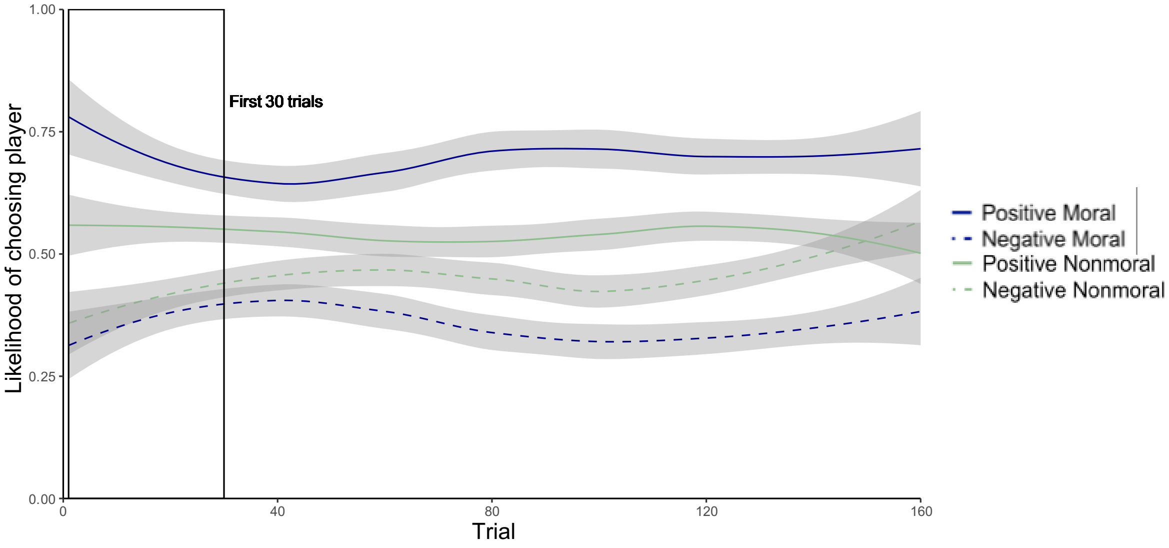
To ensure that valid responses were included in the analysis, data from trials in which participants responded very quickly (i.e., < 200ms) or very slowly (i.e., > 2000ms, also see Schultner et al., 2024) were excluded. All analyses were performed using the lme4 and lmerTest packages for R (Bates et al., 2015; Kuznetsova et al., 2017; Team, 2020). All effect sizes were calculated using the R package “EMAtools”. Additionally, we calculated effect sizes for fixed effects using semi-partial R, as recommended for generalized linear mixed models (Jaeger, Edwards, Das & Sen, 2017), using the R packages *r2glmm* (Jaeger et al., 2017) and glmmPQL (Venables & Ripley, 2002).

### *Effect of moral stereotypes on initial reward expectancies*

To test our first hypothesis—that moral stereotypes have a stronger effect on participants' group preferences than stereotypes without a moral tone—we examined participants' preferences during the first set of trials within the learning phase. That is, we tested whether participants' reward expectancies were initially more biased towards negatively-stereotyped group members when stereotypes contained moral content than nonmoral content. Following Traast et al. (in press), we selected the first 30 trials for this analysis based on visual inspection to capture participants' initial preferences while including enough responses to provide a valid estimate (Figure 1; see unsmoothed choice behavior in SI).[3]

**Figure 1**

*Choice Behavior During Training Phase Over Time*



*Note.* Smoothed choice behavior during the training phase depicting the likelihood of choosing a player across trials as a function of stereotype valence (within subjects) and stereotype

[3] Although this hypothesis was preregistered, this analysis was not. We originally planned to test this hypothesis only using computational modeling, as stated in the preregistration. However, following Traast et al. (in press), this analysis was added to provide a more direct behavioral test of the hypothesis.

morality (between subjects). The x-axis displays the trial number. The box indicates the first 30 trials of the training phase and grey lines show confidence intervals.

A mixed effects logistic regression predicting whether a participant would choose to interact with a certain player (0 = not chosen, 1 = chosen) was fitted to the first 30 trials of the *training phase*. Predictors included 1) relative reward rate (standardized and centered) of the target player compared to the second player shown in each trial, 2) stereotype valence, 3) stereotype morality, and 4) the interaction of stereotype valence and stereotype morality to this model. Random intercepts were included for subjects and random slopes for the within-subjects factors relative reward rate and stereotype valence. Because of singular fit indicating model overfit and the random effect structure being too complex, random slopes for reward rate were excluded. For the simple effects of nonmoral stereotypes, excluding all random slopes did not eliminate the singular fit, so we opted to employ the same model used in our other analyses (analysis with a repeated measures ANOVA produced similar results).

This analysis produced main effects of relative reward rate, *B* = 0.14, *SE* = 0.05, *z* = 2.86, *p* = .004, d = .13, indicating a choice preference for players with higher reward rates, and for stereotype valence, B = 1.49, SE = .29, *z* = 5.10, *p* < .001, d = 1.25, indicating a preference for players from positively-stereotyped than negatively-stereotyped groups. Although there was no main effect of stereotype morality, *B* = .22, *SE* = .19, z = 1.14, *p* = .255, d = .27, the interaction between stereotype valence and stereotype morality was significant, *B* = -0.84, *SE* = 0.38, *z* = -2.21, *p* = .027, d = -.51 (see Figure 1). This interaction, decomposed with simple effects, indicated that effect of stereotype valence on preferences was larger in the moral condition, *B* = 1.50, *SE* = 0.27, *z* = 5.63, *p* < .001, d = 1.72, than in the nonmoral condition, *B* = .65, *SE* = 0.26, *z* = 2.45, *p* = .014, d = 1.27. This interaction effect remained significant when stereotype valence ratings from the validation study were covaried, *B* = -.44, *SE* = .2, z = -2.16, *p* = .03, d = -.48 (See SI)[4].

### ***Effect of moral stereotypes on updating of reward expectancies***

Next, we asked whether moral stereotypes impair the updating of preferences for group members, such that they are more likely to persist despite stereotype-disconfirming feedback. To test this, we examined participants' preferences during the test phase. We expected that participants would continue to prefer interacting with positively over negatively stereotyped players during the test phase when the stereotypes were moral, whereas the effect of nonmoral stereotypes on reward expectancies would be diminished. To test this prediction, we fit a mixed effects logistic regression predicting whether a participant would choose to interact with a certain player in test phase trials (0 = not chosen, 1 = chosen). Predictors included 1) the relative reward rate (standardized and centered) of the target player compared to the second player shown in each trial, 2) stereotype valence, 3) stereotype morality, and 4) the interaction of stereotype valence and stereotype morality as fixed effects to this model. We added random intercepts for subjects and random slopes for the within-subjects factors reward rate and stereotype valence.

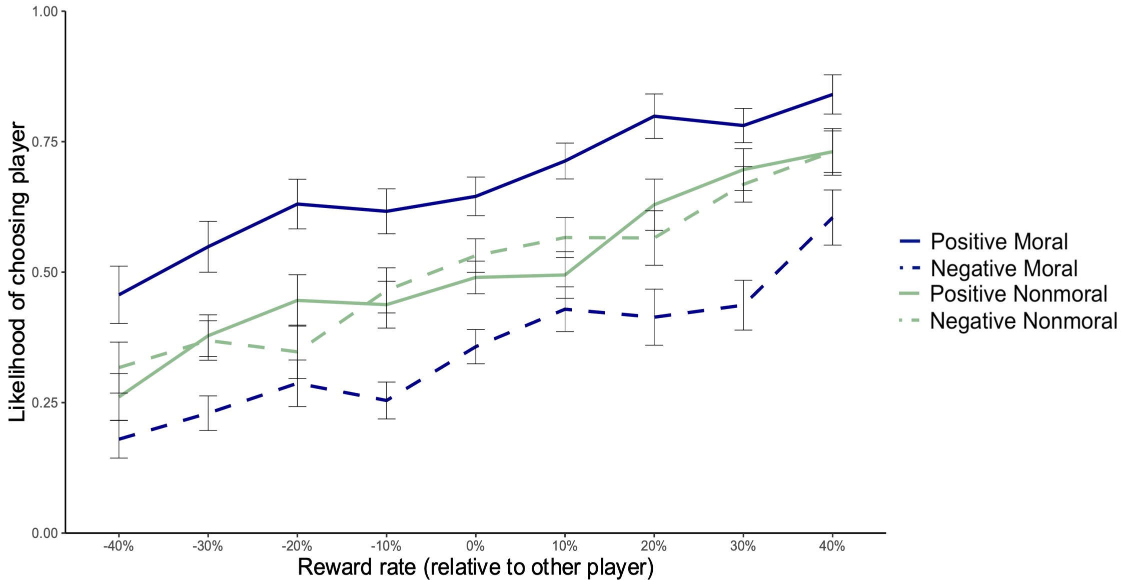
This analysis produced a main effect of relative reward rate on choice behavior, *B* = 0.67, *SE* = .09, *z* = 7.56, *p* < .001, d = 1.91, and a main effect of stereotype valence, *B* = 2.71, *SE* = 0.72, *z* = 3.74, *p* < .001, d = .84, such that participants preferred to interact with players who

[4] We did not preregister this analysis in which stereotype valence was covaried (here and subsequently); it was conceived following the preregistered analysis to address the potential effect of valence. However, the prediction follows from the preregistered hypothesis.

were more rewarding and who were stereotyped in positive terms, similar to the learning phase. A main effect of stereotype morality, *B* = 1.34, *SE* = 0.46, *z* = 2.90, *p* = .004, d = .67, additionally revealed an overall preference for players from groups stereotyped in nonmoral than moral terms. Importantly, and as predicted, these effects were qualified by a Stereotype Valence x Stereotype Morality interaction, *B* = -2.77, *SE* = 0.94, *z* = -2.94, *p* = .003, d = -.66 (Figure 2). Simple effects analyses indicated a valence-based effect of the stereotype only when the stereotype had moral content, *B* = 2.86, *SE* = 0.87, *z* = 3.27, *p* = .001, d = .50, but not when stereotypes did not have moral content, *B* = -0.05, *SE* = 0.50, *z* = -0.10, *p* = .923, d = .50. This interaction effect remained significant when stereotype valence ratings were covaried, B = -1.32, SE = .43, z = -3.05, p = .002, d = -.68 (see SI). Thus, the inclusion of moral content in group stereotypes appeared to impair the updating of impressions in response to individuating information.

**Figure 2**

*Choice Behavior During Test Phase*



*Note*. Choice behavior in the test phase depicting the relationship between the Likelihood of choosing a player, the Reward rate relative to the other player, the within-subjects factor Stereotype valence, and the between-subjects factor Stereotype morality. Error bars show standard error.

### *Computational modeling*

We used computational modeling to examine the mechanisms underlying the effect of moral stereotypes on impression formation and updating. Past research has proposed and found evidence for the *group-based learning model* (Schultner et al., 2024; Traast et al., in press), which states that stereotypes influence impression formation by (a) setting initial reward expectations for group members and then (b) biasing the updating of reward associations in response to feedback from interactions. We fitted this model to behavioral data to test whether either or both processes were amplified for moral stereotypes, relative to stereotypes lacking moral content. Following Schultner et al. (2024), initial reward expectancies were modeled as *priors* and the updating of reward associations was modeled using separate *learning rates* for each group. Using a Q-learning approach that uses training phase data to predict test phase behavior, we assessed model fit to participants' behavioral data and then compared model-derived parameter estimates of priors and learning rates between moral and nonmoral stereotype conditions.

As in Schultner et al. (2024; also Traast et al., in press), we compared the hypothesized group-based learning model with alternatives suggested by prior research. These included a) a baseline model assuming no effect of stereotypes on initial expectancies or updating (no prior and a single learning rate), b) a classic stereotype model that assumes that stereotypes bias initial expectancies but not updating (biased priors and a single learning rate; i.e., the bookkeeping model (Rothbart, 1981), and c) a model that specifies stereotype-biased learning but no difference in initial expectancies (no prior and separate group learning rates). We additionally included other plausible models suggested by reinforcement learning research (Schultner et al., 2024; Traast et al., in press, see SI for an overview).

**Model fit.** Replicating past findings (Schultner et al., 2024; Traast et al., in press), the group-based learning model, which includes biased group-based priors and separate group learning rates, provided the best fit to the data relative to alternative models (see model fits in Table S2 and predicted choices in Figure S3 in SI). Furthermore, fit to the group-based model was comparatively better for data in the moral stereotype condition than the nonmoral stereotype condition (see Tables S3-S4 in SI).

**Estimated parameter values.** To determine whether moral stereotyping effects were due to a specific component of learning, we compared parameter estimates derived from the group-based learning model for group-based priors and learning rates between conditions using a Wilcoxon signed rank test. Results revealed a stronger prior in the moral stereotype condition (*Mdn* = 0.26, *SD* = 39.65) than the nonmoral stereotype condition (*Mdn* = -0.01, *SD* = 30.25), *W* = 415, *p* = .045, r = .241, suggesting a stronger relative preference for members of the positively-stereotyped group moral stereotype condition. This effect replicated the behavioral result reported above. Parameter estimates of learning rates did not differ significantly between conditions, ($\alpha_{pos\_group}$: Moral condition *Mdn* = .13, *SD* = .36, nonmoral condition *Mdn* = .024, SD = .28; *W* = 429, *p* = .067, *r* = -.22; $\alpha_{neg\_group}$: Moral condition *Mdn* = .16, *SD* = .35, nonmoral condition *Mdn* = .02, *SD* = .34, *W* = 483, *p* = .24, *r* = -.14).suggesting that although group stereotypes influenced learning rates in both conditions, the magnitude of this effect did not differ between conditions. This combination of effects—stronger priors in the moral condition but no difference between conditions in learning rate—is consistent with the idea that stronger group-based expectancies in the moral stereotype condition, along with independent updating of reward value for each group, produced the enhanced bias in the moral stereotype condition.

**Discussion**

Study 1 supported the hypothesis that moral stereotypes have stronger and persistent effects on group member impressions than nonmoral stereotypes. Although both moral and nonmoral stereotypes influenced initial impressions, this effect was larger for moral stereotypes. Furthermore, the effect of moral stereotypes persisted as participants experienced group-equated reward feedback, whereas impressions in the nonmoral stereotype condition were updated to match feedback.

This pattern of results supports our hypothesis that moral content in stereotypes induces stronger and more persistent effects on group member impressions, relative to nonmoral content, while ruling out several alternative explanations. First, this pattern emerged after controlling for potential differences in valence extremity between moral and nonmoral stereotype attributes, through both experimental design and statistical adjustment, and thus valence extremity could not account for this effect. A second potential concern is that the task context—in which players share money—is more relevant to moral traits (e.g., generosity) than nonmoral traits; the implication would be that participants would be more responsive to moral stereotype-inconsistent feedback and thus show greater updating in the moral condition. However, our results showed the opposite pattern: updating was impaired in the moral condition. Finally, although moral traits tend to reflect stable trait characteristics, which could contribute to their persistence, we were careful to include similarly stable trait characteristics in both conditions; thus, trait stability cannot account for the observed effects. Rather, our findings are consistent with the idea that moral traits imply an essentialized goodness or badness about one's character that is difficult for a perceiver to revise (Brambilla et al., 2019; Cone & Ferguson, 2015; Strohminger & Nichols, 2014).

## Study 2

In Study 2, we replicate Study 1 while additionally testing whether the effect of moral stereotypes on impressions generalizes to novel group members in non-economic social decisions.

## Method

### Participants

This study was completed by 148 US-based participants via CloudResearch in exchange for $5.00 and a performance-based monetary bonus ($0.00-$2.50). The preregistered sample size (N=150) was doubled relative to Study 1 to increase power for post-task self-report measures; data collection stopped at 148 due to an error discovered after the study conclusion but prior to any analysis. Participants self-identified as 75.40% White/Caucasian, 9.32% African American, 3.39% Asian, 3.39% Hispanic, 0.85% Native American, and 1.69% Other, and 5.93% did not indicate their race/ethnicity. As preregistered, we excluded 30 participants who failed to reach a 50% learning criterion. The final sample size was $N$ = 118 ($M_{\text{age}}$ = 41.14 years, $SD_{\text{age}}$ = 11.76, 46 females, 65 males, 7 other). Sensitivity power analysis conducted in G*Power indicated that for N = 118, the minimum detectable effect size is d = 0.08 (α = 0.05, power = 0.80).

### ***Design and Procedure***

The design and procedure of the learning task were the same as in Study 1. Following the learning task, participants in this study additionally completed post-task ratings. The study was approved by the local Ethics Review Board.

*Post-task ratings*

To test whether preferences formed of group members during the learning task generalized to non-economic social decisions and to decisions regarding novel group members, participants were asked to rate both the players encountered during the task and new group members they had not encountered. They rated one new group member for each of the two groups encountered in the task. Participants rated their likeability (i.e., "*How much do you like [player x]?*"), willingness to work together (i.e., "*How much would you want to work together with [each player/ a new member of group A/B]?"*), hiring likelihood (i.e., "*How likely would you be to hire [each player/ a new member of group A/B] for a job?")*, and helping likelihood (i.e., "*How likely would you be to help[ each player/ a new member of group A/B with a work problem]?*"). All ratings were made on 7-point Likert scales ranging from 1 (extremely unlikely/not at all) to 7 (extremely likely/very much).

**Results**

***Effect of moral stereotypes on initial reward expectancies***

We first tested whether participants' behavior displayed a stronger initial group preference in response to moral than nonmoral stereotypes, as in Study 1. We used the same model as in Study 1, focusing on the first 30 trials (Figure 3; see unsmoothed choice behavior in SI).[5] To avoid overfitting, we excluded random slopes for reward rate for this model.

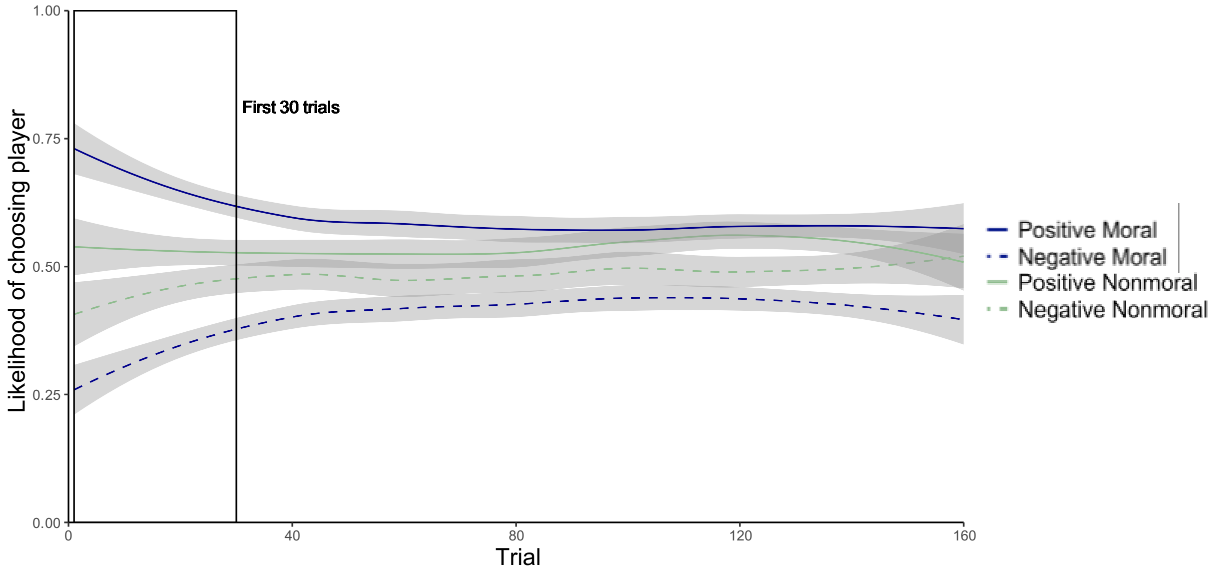
This analysis revealed main effects of relative reward rate, *B* = 0.19, *SE* = 0.05, *z* = 4.00, *p* < .001, d = .79, stereotype valence, *B* = 1.54, SE = .23, z = 6.70, p < .001, d = 1.27, and stereotype morality, *B* = .56, SE = .19, z = 3.04, p = .002, d = .58, on choice behavior. Importantly for our theoretical question, the Group Valence x Stereotype Morality interaction was significant, *B* = -1.06, *SE* = 0.34, *z* = -3.12, *p* = .002, d = -.60, indicating the valence effect of stereotypes was stronger for moral, *B* = 1.49, *SE* = 0.23, *z* = 6.44, *p* < .001, d = 1.72, than nonmoral stereotypes, *B* = .47, *SE* = 0.24, *z* = 2.00, *p* = .045, d = .52 This interaction effect remained significant when stereotyping valence ratings were covaried, B = -.53, SE = .16, z = -3.29, p < .001, d = -.62. These

[5] As in Study 1, this analysis was not preregistered, but it tests the preregistered hypothesis.

results replicated Study 1 findings and again suggest that moral stereotypes had a stronger influence on initial group preferences than nonmoral stereotypes.

**Figure 3**

*Choice Behavior During Training Phase Over Time*



*Note*. Smoothed choice behavior during the training phase depicting the likelihood of choosing a player across trials as a function of stereotype valence (within subjects) and stereotype morality (between subjects). The x-axis displays trial number. The bar indicates the first 30 trials of the training phase and grey lines show confidence intervals.

### *Effect of moral stereotypes on updating of reward expectancies*

Next, we tested whether moral stereotypes impaired the updating of preferences for group members relative to nonmoral stereotypes, using the same analysis of test phase data as in Study 1.[6] This analysis produced main effects of relative reward rate, *B* = 0.86, *SE* = 0.09, *z* =

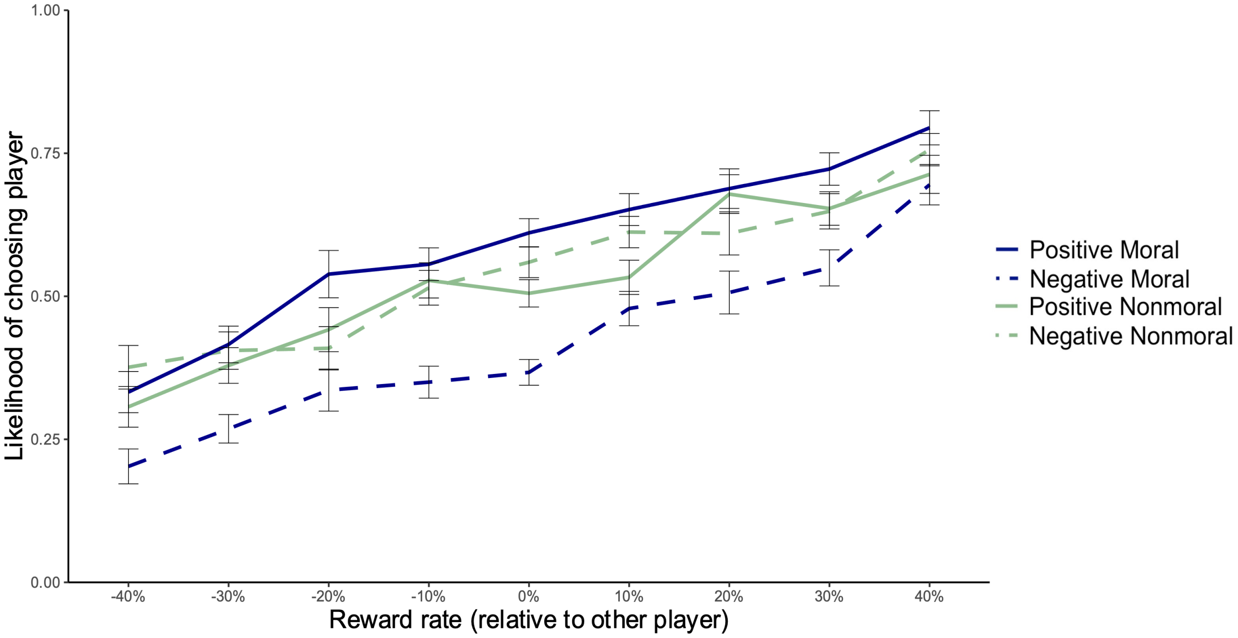
[6] The Study 2 preregistration includes an additional hypothesis that the predicted effect of moral stereotypes on impressions would be stronger for negative than positive stereotypes. This pattern was

9.79, *p* < .001, d = 1.95, stereotype morality, *B* = 1.01, *SE* = 0.29, *z* = 3.47, *p* < .001, d = .63, and stereotype valence, *B* = 1.33, *SE* = 0.39, *z* = 3.43, *p* < .001, d = .65, on choice behavior. Critically, the Stereotype Valence x Stereotype Morality interaction was significant, *B* = -1.62, *SE* = 0.57, *z* = -2.84, *p* = .004, d = -.48 (Figure 4). Simple effects analyses indicated a valence-based effect of the stereotype only when the stereotype had moral content, *B* = 1.29, *SE* = 0.37, *z* = 3.48, *p* < .001, d = .89 and not when stereotypes did not have moral content, *B* = -0.27, *SE* = 0.44, *z* = -0.61, *p* = .545, d = -.07. This critical Stereotype Valence x Stereotype Morality interaction remained significant when stereotype valence ratings were covaried, B = -.73, SE = .27, z = -2.7, p = .007, d = -.6). Thus, replicating Study 1, these results show that nonmoral stereotypes were updated in response to player feedback whereas moral stereotypes persisted.

not found and was then dropped from the preregistered replication of the main task (presented here as Study 1), and thus not discussed. These two studies were re-ordered because Study 2 includes additional measures of generalization; aside from these measure, the studies are identical.

**Figure 4**

*Choice Behavior During Test Phase*



*Note*. Choice behavior during the test phase as a function of training-phase reward rate relative to the alternative player on a given trial, stereotype valence, and stereotype morality. Error bars show standard error.

### ***Computational modeling***

As in Study 1, we used computational modeling to examine the cognitive mechanisms involved in the observed stereotype effects. Model fit comparisons again determined that model behavioral choice data were fit best by the group-based learning model, which includes a group-based biased prior and separate learning rates for each group, relative to all alternative models. We also observed better fit to this model in the moral stereotype condition than the nonmoral stereotype condition (Tables S2-4 in SI). Finally, parameter estimates for group-based priors derived from this model were larger for moral stereotypes (*Mdn* = 0.13, *SD* = 17.20) than for nonmoral stereotypes (*Mdn* = -0.01, *SD* = 14.40), *W* = 1227, *p* = .006, replicating Study 1.

### ***Post-task ratings***

Next, we asked whether the preferences formed during the learning task generalized to non-economic social decisions and decisions concerning novel group members. Because ratings across the four social decision measures were highly correlated and showed high internal consistency for both the previously-experienced players from the learning task (αs ≥ .81) and novel group members (αs ≥ .83), we chose to examine them in the same analyses for each of the experienced and novel players, respectively, and controlled for the type of rating by adding rating type as a factor to the model.

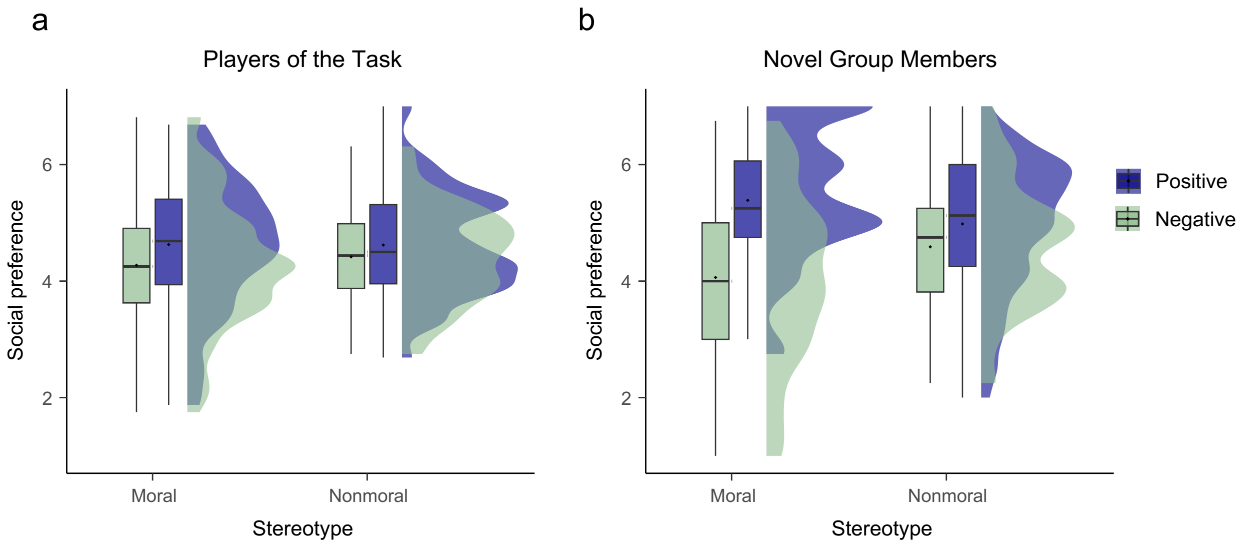
**Ratings of previously-experienced players.** To test whether stereotype-based group preferences generalized to non-economic social decision-making concerning players of the task, we fit a mixed effects regression predicting participants' ratings. Predictors included player reward rate (standardized and centered), 2) stereotype valence, 3) stereotype morality, 4) rating type, and 5) the interaction of stereotype valence and stereotype morality. We included random intercepts for subjects and random slopes for the within-subjects factors reward rate and stereotype valence (adding slopes for type of rating led to singular fit).

There was a main effect of reward rate on ratings, *B* = .38, *SE* = .04, *t* = 8.55, *p* < .001, d = 1.58, such that participants showed an increased social preference for players who were more rewarding, controlling for the type of rating. There was also a main effect of stereotype valence, *B* = 0.35, *SE* = 0.14, *t* = 2.59, *p* = .011, d = .47: participants rated players of the task more positively if they had been positively-stereotyped than negatively-stereotyped (Figure 5).

However, the interaction between stereotype valence and stereotype morality was not significant, *B* = -0.14, *SE* = 0.20, *t* = -.72, *p* = .474, d = -.13, indicating that effect of moral stereotypes on reward learning and choice did not generalize to explicit social preferences.

**Figure 5**

*Figure Displaying Choice Post-task Ratings of Players and Novel Group Members*



*Note*. Social preferences towards a) players of the task and b) novel group members depicting the relationship between the within-subjects factor Stereotype valence and the between-subjects factor Stereotype morality. We display a boxplot, mean scores, and half-density distributions.

**Ratings of novel group members.** Next, we tested whether choice biases generalized to novel group members, which would suggest a group-level generalization of learning, using the model described above (excluding reward rates). This analysis produced main effects of stereotype valence, *B* = 1.32, *SE* = .21, *t* =6.34, *p* < .001, d = 1.17, and stereotype morality, *B* = 0.53, *SE* = 0.24, *t* = 2.21, *p* = .029, d = .41, which were qualified by a Stereotype Valence x

Stereotype Morality interaction, *B* = -0.93, *SE* = 0.31, *t* = -3.01, *p* = .003, d = -.55 (Figure 5). Simple effects analyses showed that the stereotype valence effect generalized to novel members of the morally-stereotyped group, *B* = 1.32, *SE* = 0.22, *t* = 6.10, *p* < .001, d = 1.53, but not to novel members of the group stereotyped in nonmoral terms, *B* = 0.39, *SE* = 0.22, *t* = 1.82, *p* = .074, d = .50. In addition, social preferences toward novel group members were directly predicted by participants' choice preferences toward the respective group in the learning task, *B* = 1.38, *SE* = 0.48, *t* = 2.90, *p* = .004, d = .54. Together, these results show that preferences formed about morally-stereotyped group members during the learning task were more likely to generalize to novel group members.

**Discussion**

Study 2 replicated the results of Study 1: moral stereotypes had a stronger influence on initial impressions and were more resistant to change, relative to nonmoral stereotypes. Also, as in Study 1, stereotype effects were best explained by a computational model of learning whereby stereotypes induced divergent group expectancies and separate group-based updating, and this pattern was expressed more strongly for moral than nonmoral stereotypes.

Study 2 also demonstrated that moral stereotypes generalized to social decisions about novel members, whereas nonmoral stereotype did not. This pattern may reflect the fact that, in the nonmoral condition, impressions of group members were revised in response to player feedback, whereas in the moral condition, stereotype effects persisted. As a result, the stereotype continued to affect judgments of novel members in the moral condition but not the nonmoral condition, indicating the effect of a group-level representation.

By contrast, post-task judgments of existing group members, with whom participants interacted during the learning task, did not depend on the moral content of the group stereotype. Instead, post-task judgments reflected only players' actual reward feedback, replicating Hackel et al. (2015, 2020, 2022), and stereotype valence, a novel finding. It is unclear why the persistent effect of moral stereotypes seen in choice behavior was not evident in these judgments. One possibility is that participants updated their explicit impressions of players during the learning task while the s28 persistent stereotype influence on their choice behaviors reflected an implicit effect—a pattern of implicit-explicit dissociation found in past research (Schultner et al., 2024; Traast et al., in press).

## General Discussion

Social stereotypes are often moral in tone. We asked whether this moral content is what leads stereotypes to have extreme and persistent effects on impressions of group members. In two social-interactive reward learning studies, we found that moral stereotypes more strongly influenced initial impressions and were more resistant to change, relative to nonmoral stereotypes. Computational modeling indicated that moral and nonmoral stereotypes influenced impressions through the same mechanisms—by inducing biased expectancies and separate group-based updating rules—but these mechanisms were expressed more strongly for moral stereotypes. In Study 2, we found that even after repeated stereotype-disconfirming interactions with group members, moral stereotypes continued to influence social decisions about novel group members, whereas nonmoral stereotypes did not. These studies isolate the effect of moral content in stereotype-based impression formation and show that it drives the extreme and persistent effects typically associated with racial and ethnic stereotypes.

**Morality, stereotyping, and impression formation**

Our findings extend research on moral impression formation to the domain of stereotypes. Whereas prior research has demonstrated an enhanced effect of moral traits on impressions of individuals (Brambilla et al., 2019; Reeder & Coovert, 1986; Skowronski & Carlston, 1987, 1992; Wojciszke et al., 1998), we show that moral stereotypes have similarly enhanced effects on impressions of group members, which in turn may be expressed as prejudice. As with moral traits, we speculated that moral stereotypes may be considered more diagnostic of a group and thus more essential to the group's identity. Consequently, in a group context, moral content amplifies the stereotype's effect on perceivers' expectancies and interpretations of group members' behaviors (e.g., Darley & Gross, 1983; Heilman et al., 2019; Kunda & Sherman-Williams, 1993). This effect leads to more extreme initial impressions that are resistance to change. These findings may explain why real-life racial stereotypes, typically moral in tone, can be so persistent, while establishing a theoretical link between research on moral impression formation and intergroup bias.

**Dissociating effects of morality and valence extremity**

Given their heightened diagnosticity, moral traits may be perceived as more extreme in valence than nonmoral traits (Brambilla et al., 2019; Cone & Ferguson, 2015; Reeder & Coovert, 1986; Skowronski & Carlston, 1987; Wojciszke et al., 1998); thus is may be difficult to distinguish effects of moral content from valence on impressions. The present work addressed this issue in three ways. First, independent ratings of manipulated stereotype content showed that negative moral and nonmoral traits did not differ in extremity, although positive moral and nonmoral traits differed slightly. Second, we showed that moral stereotype effects on behavior remained significant after adjusting for any extremity differences in trait ratings (see SI), conceptually replicating prior research in which moral attitude effects remained after adjusting for attitude strength (e.g., Lutrell et al., 2022; Skitka et al., 2005). And third, we observed effects of both positive and negative moral stereotypes on impressions, relative to nonmoral stereotypes, further indicating that the effect of morality was not dependent on valence. Together, these findings add support to our conclusion that moral content enhances the impact and durability of stereotypes beyond any effect of stereotype valence extremity.

## Moral stereotyping in intergroup social interactions

A novel feature of this research is its focus on social interaction-based impression formation. By contrast, past research on moral impression effects has focused on conceptual trait learning and explicit judgments. Our approach, which used a social reinforcement learning paradigm involving action and feedback, allowed us to examine the effect of moral stereotypes on impression formation and updating across repeated interaction. This approach also permitted tests of specific underlying learning mechanisms using computational modeling. It is unclear whether the effects of moral stereotypes we observed in behavioral preferences would also be evident in conceptual judgments. The post-task measures in Study 2 suggest that self-reported trait judgments may be updated more readily than behavioral choice preferences, suggesting the possibility that moral stereotypes may have different effects on semantic and instrumental components of impression formation (Amodio, 2019; Amodio & Cikara, 2021).

## Conclusion

The current research examined whether moral stereotypes have stronger and more persistent effects on impression formation of group members than stereotypes without a moral component. Our findings suggest that by biasing both initial impressions and the updating of impressions over time, they contribute to group prejudices that are difficult to change.

**Open Practices**

Hypotheses, sample sizes, exclusion criteria, and analysis plans for both studies were preregistered (Study 1: https://aspredicted.org/FYB_WPX; Study 2: https://aspredicted.org/PRG_VEX). Materials, data, and analysis scripts are publicly available at OSF: https://osf.io/7kpn4/?view_only=dee015d400d545188c63f06eb33ce593. All studies, measures, manipulations, and data/participant exclusions are reported in the manuscript or its Supplementary Material, and any deviations from preregistrations or analyses not described in a preregistration are noted.

## References

Abele-Brehm, A., Ellemers, N., Fiske, S. T., Koch, A., & Yzerbyt, V. (2020). Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychological Review*, *128*(2), 290–314.

Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley. https://psycnet.apa.org/record/1954-07324-000

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

Amodio, D. M., & Cikara, M. (2021). The Social Neuroscience of Prejudice. *Annual Review of Psychology*, *72*, 439–469. https://doi.org/10.1146/annurev-psych-010419-050928

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/doi:10.18637/jss.v067.i01

Brambilla, M., Carraro, L., Castelli, L., & Sacchi, S. (2019). Changing impressions: Moral character dominates impression updating. *Journal of Experimental Social Psychology*, *82*(May 2018), 64–73. https://doi.org/10.1016/j.jesp.2019.01.003

Brambilla, M., & Leach, C. W. (2014). On the importance of being moral: The distinctive role of morality in social judgment. *Social Cognition*, *32*(4), 397–408. https://doi.org/10.1521/soco.2014.32.4.397

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology*, *41*(2), 135–143. https://doi.org/10.1002/ejsp.744

Brambilla, M., University, S. S., Rusconi, P., & Goodwin, G. P. (2021). The Primacy of Morality in Impression Development: Theory, Research, and Future Directions. *Advances in Experimental Social Psychology Brambilla, 64*.

Byron C. Jaeger, Lloyd J. Edwards, Kalyan Das & Pranab K. Sen (2017) An *R*2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics, 44(6)*, 1086-1105, DOI: 10.1080/02664763.2016.1193725

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, *108*(1), 37–57. https://doi.org/10.1037/pspa0000014

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map. *Advances in Experimental Social Psychology*, *40*(07), 61–149. https://doi.org/10.1016/S0065-2601(07)00002-0

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20–33. https://doi.org/10.1037/0022-3514.44.1.20

Day, M. V., Fiske, S. T., Downing, E. L., & Trail, T. E. (2014). Shifting liberal and conservative attitudes using moral foundations theory. *Personality and Social Psychology Bulletin, 40(12),* 1559–1573. http://dx.doi.org/10.1177/0146167214551152.

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

Fiske, S. T. (1998a). *Stereotyping, prejudice, and discrimination.* (D. T. Gilbert, S. T. Fiske, & G. Lindzey (eds.); pp. 357–411). McGraw-Hill. https://psycnet.apa.org/record/1998-07091-025

Fiske, S. T. (1998b). Stereotyping, prejudice, and discrimination. In & G. L. D. T. Gilbert, S. T. Fiske (Ed.), *The handbook of social psychology*. McGraw-Hill.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. https://doi.org/10.1037//0022-3514.82.6.878

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. *Advances in Experimental Social Psychology*, *23*(C), 1–74. https://doi.org/10.1016/S0065-2601(08)60317-2

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science*, *306*(5703), 1940–1943. https://doi.org/10.1126/science.1102941

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science, 24(1),* 38–44. https://doi.org/10.1177/0963721414550709

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. https://doi.org/10.1037/a0034726

Graham, J., Nosek, B. A., & Haidt, J. (2012). The Moral Stereotypes of Liberals and

Conservatives: Exaggeration of Differences across the Political Spectrum. *PLoS ONE*, *7*(12). https://doi.org/10.1371/journal.pone.0050092

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience*, *18*(9), 1233–1235. https://doi.org/10.1038/nn.4080

Hackel, L. M., Kogon, D., Amodio, D. M., & Wood, W. (2022). Group value learned through interactions with members: A reinforcement learning account. *Journal of Experimental Social Psychology*, *99*(January 2021), 104267. https://doi.org/10.1016/j.jesp.2021.104267

Hackel, L. M., Mende-Siedlecki, P., & Amodio, D. M. (2020). Reinforcement learning in social interaction: The distinguishing role of trait inference. *Journal of Experimental Social Psychology*, *88*(December 2019), 103948. https://doi.org/10.1016/j.jesp.2019.103948

Heilman, M. E., Manzi, F., & Caleo, S. (2019). Updating impressions: The differential effects of new performance information on evaluations of women and men. *Organizational Behavior and Human Decision Processes*, *152*, 105–121. https://doi.org/10.1016/J.OBHDP.2019.03.010

Heiphetz, L. (2019). Moral essentialism and generosity among children and adults. Journal of Experimental Psychology: General, 148, 2077-2090.

Hilton, J. L., & Von Hippel, W. (1996). Stereotypes Article in Annual Review of Psychology ·. *Annual Review of Psychology*, *47*, 237–71. https://doi.org/10.1146/annurev.psych.47.1.237

Jackson, L. (2010). Images of Islam in US Media and their Educational Implications. *Educational Studies*, *46*(1), 3–24. https://doi.org/10.1080/00131940903480217

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017) An R2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics, 44*, 1086-1105, DOI: 10.1080/02664763.2016.1193725

Kunda, Z., & Sherman-Williams, B. (1993). Stereotypes and the construal of individuating information. *Personality and Social Psychology Bulletin*, *19*(1), 90–99.

Kunst, J. R., Fischer, R., Sidanius, J., & Thomsen, L. (2017). Preferences for group dominance track and mediate the effects of macro-level social inequality and violence across societies. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(21), 5407–5412. https://doi.org/10.1073/pnas.1616572114

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models . *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Mooijman, M., & Hoover, J. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-018-0353-0

Nicolas, G., & Fiske, S. T. (2023). Valence Biases and Emergence in the Stereotype Content of Intersecting Social Categories. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001416

Papakyriakopoulos, O., & Zuckerman, E. (2021). The Media During the Rise of Trump: Identity Politics, Immigration,"Mexican" Demonization and Hate-Crime. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*, 467–478. https://doi.org/10.1609/icwsm.v15i1.18076

Phalet, K., & Poppe, E. (1997). Competence and morality dimensions of national and ethnic stereotypes: A study in six eastern-European countries. *European Journal of Social Psychology*, *27*(6), 703–723. https://doi.org/10.1002/(SICI)1099-0992(199711/12)27:6<703::AID-EJSP841>3.0.CO;2-K

Pratto, F., Çidam, A., Stewart, A. L., Zeineddine, F. B., Aranda, M., Aiello, A., Chryssochoou, X., Cichocka, A., Cohrs, J. C., Durrheim, K., Eicher, V., Foels, R., Górska, P., Lee, I. C., Licata, L., Liu, J. H., Li, L., Meyer, I., Morselli, D., … Henkel, K. E. (2013). Social Dominance in Context and in Individuals: Contextual Moderation of Robust Effects of Social Dominance Orientation in 15 Languages and 20 Countries. *Social Psychological and Personality Science*, *4*(5), 587–599. https://doi.org/10.1177/1948550612473663

Reeder, G. D., & Coovert, M. D. (1986). Revising an Impression of Morality. *Social Cognition*, *4*(1), 1–17. https://doi.org/10.1521/SOCO.1986.4.1.1

Rothbart, M. (1981). Memory processes and social beliefs. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping* (pp. 145–181).

Schultner, D. T., Stillerman, B. S., Lindström, B. R., Hackel, L. M., Hagen, D. R., Jostmann, N. B., & Amodio, D. (2024). *Societal stereotypes shape learning to produce group-based preferences*. 1–48. https://doi.org/10.31234/osf.io/mwztc

Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass, 4(4)*, 267–281. https://doi.org/10.1111/j.1751-9004.2010.00254.x

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology, 88(6), 895–917.* https://doi.org/10.1037/0022-3514.88.6.895

Skowronski, J. J., & Carlston, D. E. (1987). Social Judgment and Social Memory: The Role of Cue Diagnosticity in Negativity, Positivity, and Extremity Biases. *Journal of Personality and Social Psychology*, *52*(4), 689–699. https://doi.org/10.1037/0022-3514.52.4.689

Skowronski, J. J., & Carlston, D. E. (1992). Caught in the act: When impressions based on highly diagnostic behaviours are resistant to contradiction. *European Journal of Social Psychology*, *22*(5), 435–452. https://doi.org/10.1002/EJSP.2420220503

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171. https://doi.org/10.1016/j.cognition.2013.12.005

Team, R. C. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Traast, I. J., Schultner, D. T., Doosje, B., & Amodio, D. M. (in press). *Race effects on impression formation in social interaction: An instrumental learning account*. https://doi.org/10.31234/osf.io/3j2rm

Van Bavel, J. J., Packer, D. J., Haas, I. J., & Cunningham, W. A. (2012). The importance of moral construal: Moral versus non-moral construal elicits faster, more extreme, uni- versal evaluations of the same actions. PloS One, 7(11), e48693. http://dx.doi.org/10. 1371/journal.pone.0048693.

Van Lange, P. A. M., & Kuhlman, D. M. (1994). Social Value Orientations and Impressions of Partner's Honesty and Intelligence: A Test of the Might Versus Morality Effect. *Journal of Personality and Social Psychology*, *67*(1), 126–141. https://doi.org/10.1037/0022-3514.67.1.126

Venables, W. N., Ripley, B. D., Venables, W. N., & Ripley, B. D. (2002). Random and mixed effects. *Modern applied statistics with S*, 271-300.

Welch, K., Payne, A. A., Chiricos, T., & Gertz, M. (2011). The typification of Hispanics as criminals and support for punitive crime control policies. *Social Science Research Journal*, *40*, 822–840. https://doi.org/10.1016/j.ssresearch.2010.09.012

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *PSPB*, *24*(12), 1251–1263.