Journal of Experimental Social Psychology 119 (2025) 104750

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Effects of moral stereotypes on the formation and persistence of group preferences[☆]

Inga K. Rösler[a,*], Isabel Kerber[b,**], David M. Amodio[a,*]

[a] University of Amsterdam, Department of Psychology, PO Box 15900, 1001 NK Amsterdam, The Netherlands
[b] Humboldt University of Berlin, Faculty of Life Sciences, Institute of Psychology, Unter den Linden 6, 10099, Berlin

ARTICLE INFO

Keywords:
Moral
Stereotype
Learning
Computational
Prejudice

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 decision game, modeled on a reward reinforcement task, in which they formed impressions of players based on their feedback. 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 behavioral choice preferences for players from positively-stereotyped groups over negatively-stereotyped groups. Importantly, this effect was moderated by the moral content of the stereotypes: in the moral stereotype condition, participants formed more extreme initial expectancies for players' feedback and showed more resistance to updating in response to stereotype-disconfirming feedback, whereas in the nonmoral stereotype condition, initial expectancies were weaker and preferences were updated over time to match players' actual feedback. Study 2 replicated this effect and additionally showed that moral stereotypes generalize more strongly to impressions of novel group members compared with nonmoral stereotypes. Computational modeling suggests this moral stereotype effect is due to extreme initial expectancies combined with group-based updating of member impressions. Together, these studies demonstrate that moral stereotypes have a stronger influence on person impressions than nonmoral stereotypes, and that they do so by inducing stronger expectancies for a group member's behavior while impairing individuated updating.

Stereotypes of minority ethnic groups are often moral in tone (e.g., Abele-Brehm et al., 2020; Fiske et al., 2002). For example, in the United States, stereotypes commonly asscociate Latinos with crime (Welch et al., 2011), African Americans with hostility (Devine & Elliot, 1995), and Muslim men with terrorism (e.g., Jackson, 2010). Although ethnic stereotypes can also include nonmoral traits such as unintelligence or lack of skill, 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 impressions formed of group members and whether such impressions are more resistant to updating.

1. Moral stereotypes and impression formation

Stereotypes are societal-level beliefs regarding social groups and their members which can refer to members' traits (e.g., criminal) or life circumstances (e.g., being poor; Allport, 1954; Devine & Elliot, 1995). A major function of stereotypes is to guide impressions of individual group members by creating expectancies, which in turn help a perceiver characterize a group member's behavior and predict their future actions (Fiske, 1998; Hamilton et al., 1990). However, because stereotypes are generalizations that may 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

[☆] This paper has been recommended for acceptance by Paul Conway.

[*] Corresponding authors at: Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129, REC G, 1001 NK Amsterdam, Netherlands.

[**] Corresponding author at: Department of Psychology, Humboldt University of Berlin, Rudower Chaussee 18, 12489 Berlin, Germany.

E-mail addresses: ingakaroesler@gmail.com (I.K. Rösler), d.m.amodio@uva.nl (D.M. Amodio).

https://doi.org/10.1016/j.jesp.2025.104750
Received 12 March 2024; Received in revised form 10 March 2025; Accepted 17 March 2025

impression formation have not been investigated. In the impression formation literature, however, research shows that moral information profoundly affects a perceiver's trait inferences (Abele-Brehm et al., 2020; Brambilla et al., 2011, 2021; Brambilla & Leach, 2014). One explanation for this effect is that moral information is often perceived to be diagnostic, revealing a person's 'true' essential character and thus used to understand their intentions and behaviors (Cuddy et al., 2008; Goodwin et al., 2014; Heiphetz, 2019; Strohminger & Nichols, 2014). Alternatively, this effect has been explained in terms of the "other-profitability" of moral traits—that is, their potential to help or harm others, including the perceiver—in comparison to the "self-profitability" of competence-related traits that may primarily serve those who posses those traits (Peeters & Czapinski, 1990; Wojciszke, 2005). Consistent with these accounts, information about a person's honesty has been found to weigh more 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 a candidate's competence (Luttrell et al., 2022). Thus, across a variety of contexts, moral information has been shown to have an especially potent effect on impressions, perceived intentions, and expectations (Day et al., 2014; Luttrell et al., 2016; 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 that moral content in stereotypes guides impressions of group members in the same way that it guides impressions of unaffiliated individuals (Abele-Brehm et al., 2020; Brambilla et al., 2011, 2021; Brambilla & Leach, 2014), moral stereotypes should have a stronger influence on impressions of group members compared with nonmoral stereotypes.

2. 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 individuating information about a group member to fit stereotypic expectancies (Darley & Gross, 1983; Hamilton et al., 1990; Kunda & Sherman-Williams, 1993). This effect may be exacerbated when stereotypes include moral context: because moral information tends to be seen as diagnostic of a person's true character (Brambilla et al., 2019; Cone & Ferguson, 2015; Mende-Siedlecki et al., 2013), moral impressions are especially resistant to change (Luttrell et al., 2016, Luttrell et al., 2022; Reeder & Coovert, 1986; Skowronski & Carlston, 1992). Thus, if moral content operates similarly in the context of stereotypes, then a moralized stereotype may also impede the updating of impressions toward members of the stereotyped group.

Although research has not previously examined the effect of moral stereotypes on updating, Schultner et al. (2024) recently found that group stereotypes impede the updating of impressions by altering the construal of a group member's behavior. Using social instrumental learning tasks, adapted from prior reinforcement learning paradigms (e.g., Frank et al., 2004; Hackel et al., 2022), their studies investigated how stereotypical information is continuously updated across repeated interactions with group members. Participants in Schultner et al. (2024) 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 decision task. Although members of each group behaved identically in the task, on average, thereby disconfirming the stereotypes, participants' choice behaviors reflected a persistent preference for members of the positively-stereotyped group. Schultner et al. (2024) proposed that this stereotype effect was due to two concerted processes: stereotypes (a) created initial divergent expectancies for group members' behavior and then (b) led perceivers to update their impressions of individual group members at the group level, as opposed to the individual level, thereby undermining the potential for individuation. Using a computational modeling approach in five studies, they found that this mechanism explained the effect of stereotypes on participants' impression formation and updating, and that it did so better than several alternative theoretical accounts. 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 this biasing effect of stereotypes on impression updating.

3. Research overview

In the present research, 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; Wojciszke et al., 1998). Integrating prior research on moral impression formation and stereotyping, we hypothesized that stereotypes with moral (vs. nonmoral) content would induce more extreme initial expectancies and impede impression updating in response to stereotype-inconsistent information. In two experiments, we tested these hypotheses using a social reinforcement learning paradigm in combination with behavioral analysis and computational modeling (Schultner et al., 2024, Traast, Schultner, et al., 2024). This approach allowed us to test our main hypothesis experimentally with behavioral data from repeated interactions with group members over time, while also testing our proposed theoretical mechanism using computational modeling.

Hypotheses, exclusion criteria, and sample size were preregistered for both studies prior to data collection (Study 1: https://aspredicted.org/FYB_WPX;; Study 2: https://aspredicted.org/PRG_VEX); deviations and any analyses not included in preregistration are noted. All data, code, and materials are available at the Open Science Framework [https://osf.io/7kpn4/]. Approval for both studies 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.

4. Study 1

In Study 1, participants learned they would complete a social decision task in which they would interact with members of two groups. Before beginning the task, participants read group descriptions that described one group with positive stereotypes and the other with negative stereotypes. For participants in the moral condition, these descriptions included moral content; for participants in the nonmoral condition, descriptions related to competence and did not include moral content. Participants then completed the decision task in which, on each trial, they viewed a member from each group and had to choose one to interact with in exchange for a potential point. The participant's goal was to form an impression of players, through trial and error, based on players' likelihood of providing a reward (i.e., a point), and to then use this impression to guide future interaction choices. By including these essential elements of direct socio-interactive learning—that is, the formation of person preferences through action and feedback—it permitted an experimentally controlled test of interaction-based impression formation (Hackel et al., 2015; Schultner et al., 2024).

Importantly, despite group stereotypes, players from each group provided rewards at identical rates, on average. Therefore, over time, participants' choice preferences should be updated to reflect individual players' reward rates rather than the stereotype. However, we predicted that, compared with nonmoral stereotypes, moral stereotypes would induce stronger initial reward expectancies as well as group-based representations of individual players, which together would make participants' impressions more resistant to updating.

4.1. Method

4.1.1. 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$) was preregistered and based on prior research using a similar task (Schultner et al., 2024). The self-identified race/ethnicity of the sample was 78.30 % White/Caucasian, 7.25 % Asian, 5.80 % African American, and 1.45 % Hispanic, with 1.45 % indicating 'Other' and 5.80 % who did not indicate their race/ethnicity. Following preregistered exclusion criteria, we excluded one participant who showed non-compliant behavior (i.e., responding too fast on all but one trial in the test phase) and ten participants who failed to reach a 50 % learning criterion for extreme reward rates (30 %, 70 %) from the analysis (Schultner et al., 2024).[1] The final sample included 69 participants ($M_{\text{age}} = 42.35$ years, $SD_{\text{age}} = 13.67$, 32 females, 37 males). Sensitivity power analysis conducted in G*Power determined that the minimal detectable effect size for the mixed-factors Stereotype Valence x Stereotype Morality interaction ($N = 69$), our primary effect of interest, was $d = 0.11$ ($\alpha = 0.05$, power $= 0.80$).

4.1.2. Experimental design

To investigate whether moral (vs. nonmoral) stereotypes influence reward learning from group members, we used an adapted version of a probabilistic reward reinforcement learning task that has been used and validated in prior work (Frank et al., 2004; Hackel et al., 2022; Schultner et al., 2024). The experimental design included mixed factors: 2 (stereotype valence: positive vs. negative; within-subjects) x 4 (reward rate of group members: 70 %, 60 %, 40 %, and 30 %; within-subjects) x 2 (stereotype morality: morality vs. competence; between-subjects).

4.1.3. Procedure

Following consent, participants learned that the study examined impression formation, and that they would engage in an interactive decision task with players from two different groups who come from different places (see Fig. 1). The groups were referred to as "Group A" and "Group B," and players would be represented by avatars, presumably to maintain their anonymity. Participants were further told that these groups are known to differ in the cultures and traits and were then presented with 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 individuals within each group vary, too, and thus the participant should attend to the behavior of each individual player. This instruction, in combination with the group descriptions, represents the conventional construct of a stereotype: a group-level attribute description that generalizes to some, but not all, members of the group (Allport, 1954). Participants then completed a social categorization task to confirm that they learned the group membership of individual players and the attributes associated with each group.

Next, participants began the social decision task. They were told that the other players had taken part in a previous experiment in which they could give or withhold points across multiple rounds of the game, and that in the present study, participants would play with these participants, represented by avatars. This procedure, in which participants respond to and receive feedback based on players' previous behavior in a prior study, is commonly used in behavioral economics research to reduce deception and enhance believability while maintaining experimental control (e.g., Levine & Schweitzer, 2015). To indicate players' group memberships, player avatars differed in hair color, eye color, and color of clothing (as in Schultner et al., 2024). 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. Across studies, avatar gender did not interact with our main findings, $ps > 0.148$, and therefore gender effects are not further discussed.

Participants were instructed to learn which individual players were most likely to share points 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 responded with a point. Participants were told that some individuals provide more points than others and that no player always shares. 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.

4.2. Materials and tasks

Stereotype Descriptions. Stereotype descriptions were based on 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. 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). Moral and competence descriptions were designed to be equated in valence, respectively for positive and negative descriptions, and equated in their perceived trait stability. The same proportion of person- and societal-level attributes were used for each description.

Although the moral and competence stereotype descriptions were designed to be matched in valence, the greater diagnosticity of moral traits may nevertheless lead them to be perceived as more extreme in valence than nonmoral traits (Brambilla et al., 2019; Cone & Ferguson, 2015; Reeder & Coovert, 1986; Skowronski & Carlston, 1987), which can in turn create asymmetric effects on impression updating (Mende-Siedlecki et al., 2013). To address this possibility in our stimuli, we assessed valence extremity ratings of the full stereotype descriptions presented to participants in the task, including counterbalanced presentation, in a separate sample ($N = 100$; see SI)[2]. Results indicated that negative moral ($M = 1.66$, $SD = 0.89$) and nonmoral ($M = 1.90$, $SD = 1.05$) stereotypes did not differ in valence, $t(95.46) = 1.23$, $p = .223$. However, positive moral ($M = 6.56$, $SD = 0.64$) and nonmoral ($M = 6.18$, $SD = 1.04$) stereotypes did differ, $t(81.58) = -2.19$, $p = .031$. 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.

In prior research, such valence differences have been addressed statistically, such that effects of morally-based attitudes on judgments, relative to nonmoral attitudes, were tested and found after adjusting for attitude strength (Luttrell et al., 2016; Skitka et al., 2005). Thus, although valence effects in our stimuli were small and only observed for positive stereotypes, we included the valence ratings from the independent pilot sample as covariates in tests of our main hypothesis to rule out any potential confounding effect of valence extremity.

Categorization task. Prior to completing the main learning task, participants completed a classification task where they categorized both

[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.

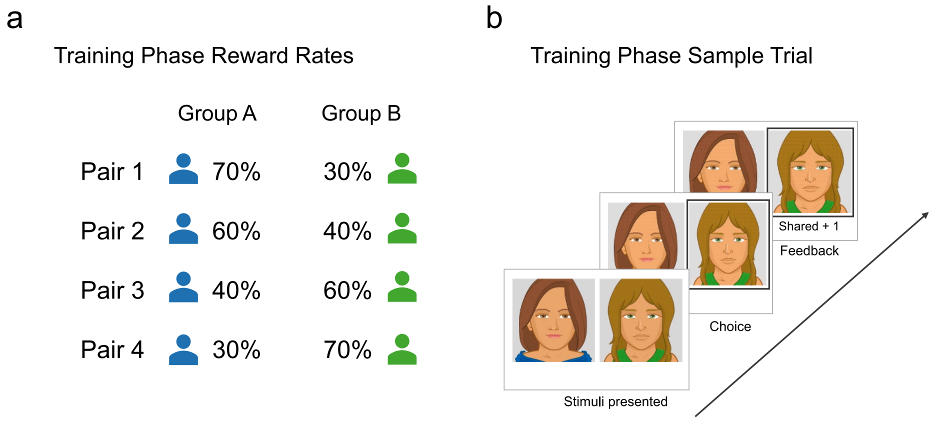

Fig. 1. Player reward rates and sample trial in training phase.
Note. Panel (a) displays reward rates for player pairs in the training phase. Panel (b) shows a sample trial sequence of the training phase. Participants viewed two players, chose one to interact with (right player in this sample trial), and then received feedback.

players and stereotype attributes as belonging to either Group A or Group B. This task ensured that participants learned to associate individual players and the stereotype descriptions with the appropriate group labels. On each trial, participants viewed either a player avatar or a stereotype attribute and classified it as belonging to either Group A or Group B, via keypress. Target stimuli alternated between avatar images and stereotype words. The task included 16 trials, and accuracy feedback was given following each response. Average classification accuracy was 86.3 % for Study 1 and 89.8 % for Study 2 (both above chance level, $t(68) = 20.84$, $p < .001$, $t(117) = 33.201$, $p < .001$, respectively), demonstrating learning of both group stereotypes and group memberships.

Social Decision Task. The social decision task, in which participants could form and update impressions, 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). Participants were told that, on each trial, only one player would offer a point. It was 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 (30 %–70 %, 40 %–60 %, 60 %–40 %, or 70 %–30 %, see Fig. 1).

Upon completing the training phase, participants took a short break and then began the *test phase*. The test phase included 96 trials and was designed to assess the reward-based associations formed during training. Participants again viewed pairs of players, always one Group A player and one Group B player, 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 (e.g., including pairs with the same reward rates during training), 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.

4.3. Results

To ensure that only valid responses were included in the analysis, data from trials in which participants responded very quickly (i.e., < 200 ms) or very slowly (i.e., > 2000 ms, 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; R Core Team, 2024). 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 et al., 2017a), using the R packages r2glmm (Jaeger et al., 2017b) and glmmPQL (Venables et al., 2002).

4.3.1. Effect of moral stereotypes on initial reward expectancies

To test our first hypothesis—that moral stereotypes have a stronger effect on participants' initial group preferences than stereotypes without a moral tone—we examined participants' preferences during the first set of trials within the training phase. That is, we tested whether participants' reward expectancies were initially more biased, favoring positively-stereotyped members over negatively-stereotyped members, when stereotypes contained moral content compared to nonmoral content. Following Traast et al. (2024), 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 (Fig. 2; see unsmoothed choice behavior in SI).[2]

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 (a) relative reward rate (standardized and centered) of the target player compared to the alternative player shown on each trial, (b) stereotype valence, (c) stereotype morality, and (d) the interaction of stereotype valence and stereotype morality. 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. However, tests of simple effects included random effects for all predictors.

This analysis produced main effects of relative reward rate, $B = 0.14$, $SE = 0.05$, $z = 2.86$, $p = .004$, $d = 0.13$, indicating a choice preference for players with higher reward rates, and for stereotype valence, $B = 1.49$, $SE = 0.29$, $z = 5.10$, $p < .001$, $d = 1.25$, indicating a preference for players from positively-stereotyped groups over those from negatively-stereotyped groups. Although the main effect of stereotype morality was nonsignificant, $B = 0.22$, $SE = 0.19$, $z = 1.14$, $p = .255$, $d = 0.27$, it was qualified by the Stereotype Valence X Stereotype Morality

[2] 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, Doosje, & Amodio, 2025, this analysis was added to provide a more direct behavioral test of the hypothesis Traast et al. (2025).

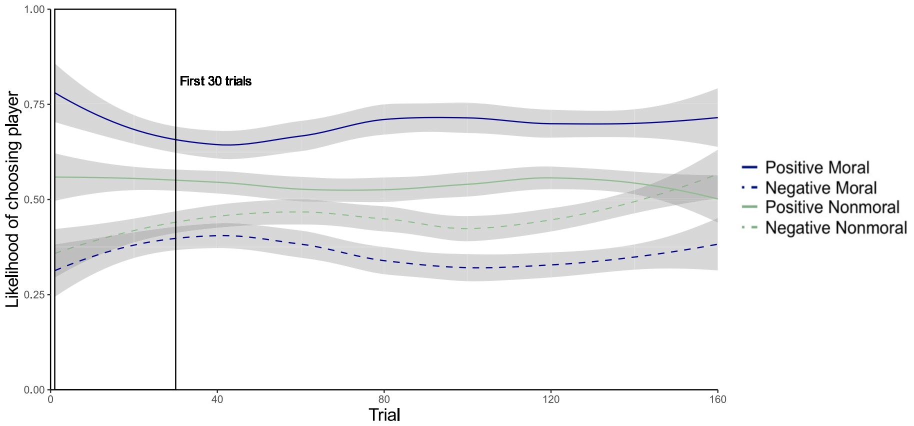

Fig. 2. Study 1 choice behavior during training phase over time.
Note. Smoothed choice behavior during the training phase depicting the likelihood of choosing a player across trials as a function of stereotype valence (within subjects) and stereotype morality (between subjects). The x-axis displays trial number. The box indicates the first 30 trials of the training phase and grey shading shows confidence intervals.

interaction, $B = -0.84$, $SE = 0.38$, $z = -2.21$, $p = .027$, $d = -0.51$ (see Fig. 2). This interaction, decomposed with simple effects, indicated that the 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, $B = 0.65$, $SE = 0.26$, $z = 2.45$, $p = .014$, $d = 1.27$. This interaction effect remained significant when stereotype valence ratings from the validation study were covaried, $B = -0.44$, $SE = 0.2$, z $= -2.16$, $p = .030$, $d = -0.48$ (See SI).[3] These results supported our first hypothesis that moral stereotypes would induce more extreme expectancies for group member behavior compared with nonmoral stereotypes.

4.3.2. *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 given player in test phase trials (0 = not chosen, 1 = chosen). Predictors included (a) the relative reward rate (standardized and centered) of the target player compared to the alternative player in each trial, (b) stereotype valence, (c) stereotype morality, and (d) the interaction of stereotype valence and stereotype morality as fixed effects. 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, $B = 0.67$, $SE = 0.09$, $z = 7.56$, $p < .001$, $d = 1.91$, and a main effect of stereotype valence, $B = 2.71$, $SE = 0.72$, $z = 3.74$, $p < .001$, $d = 0.84$, such that participants preferred to interact with players who were more rewarding and who were stereotyped in positive terms, similar to the training phase. A main effect of stereotype morality, $B = 1.34$, $SE = 0.46$, $z = 2.90$, $p = .004$, $d = 0.67$, 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, $B = -2.77$, $SE = 0.94$, $z = -2.94$, $p = .003$, $d = -0.66$ (Fig. 3). Simple effects analyses indicated a valence-based effect of the stereotype only when the stereotype had moral content, $B = 2.86$, $SE = 0.87$, $z = 3.27$, $p = .001$, $d = 0.50$, but not when stereotypes did not have moral content, $B = -0.05$, $SE = 0.50$, $z = -0.10$, $p = .923$, $d = 0.50$. This interaction effect remained significant when stereotype valence ratings were covaried, $B = -1.32$, $SE = 0.43$, $z = -3.05$, $p = .002$, $d = -0.68$ (see SI). Thus, the inclusion of moral content in group stereotypes appeared to impair the updating of impressions in response to individuating information, supporting our second hypothesis.

Finally, we considered a potential alternative explanation for the lack of updating in the moral stereotype condition: given the stronger expectancies created by moral stereotypes, participants might have paid less attention to player's individual-level feedback. If this alternative were true, then learning of players' relative reward rates would be weaker in the moral condition compared with the nonmoral condition. However, when the Stereotype Morality x Reward Rate interaction was includedd in the model tested above, this interaction was not significant, $B = -0.01$, $SE = 0.19$, $z = -0.06$, $p = .950$, indicating that the learning of players' relative reward rates within condition did not differ by condition (as can be seen by the similar slopes for reward rate across conditions in Fig. 3). Thus, the effect of moral stereotypes on updating is not explained by reduced attention to individual player feedback in the moral condition.

4.3.3. *Computational modeling*

We used computational modeling to examine the mechanisms underlying the effect of moral stereotypes on impression formation and updating. According to the *stereotype-learning model* (Schultner et al., 2024), stereotypes influence impression formation by (a) setting initial reward expectations for group members and then (b) updating reward associations for individual players according to a group-level value representation. We hypothesized that this model would explain the overall effect of stereotypes in our study and, to account for the stronger impact of moral stereotypes on choice preferences, that the fit of this model would be better in the moral stereotype condition than in the nonmoral stereotype condition.

Following Schultner et al. (2024), initial reward expectancies were modeled as opposing *priors*, where P denotes a prior for either the positively- or negatively-stereotyped group:

$$Q^{t=0}_{positive} = P, \text{and } Q^{t=0}_{positive} = -P$$

The updating of reward representations followed the Rescorla-Wagner learning rule:

[3] 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.

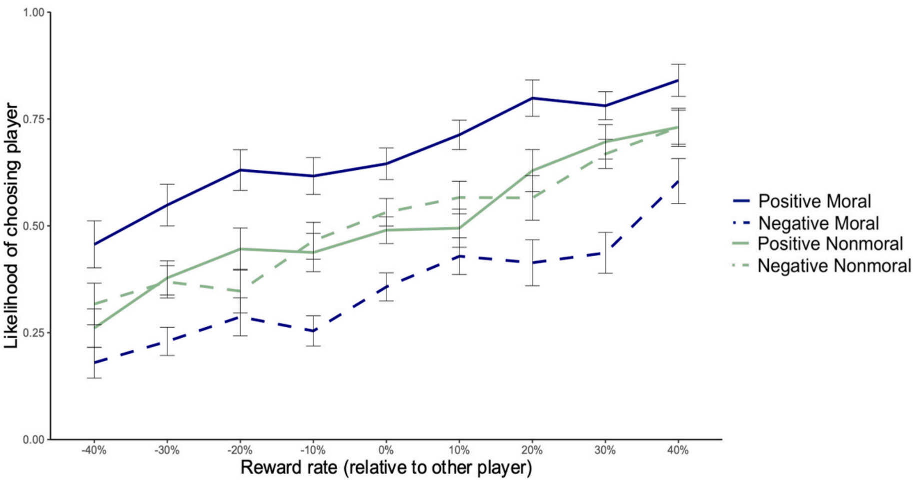

Fig. 3. Study 1 choice behavior during test phase.
Note. Choice behavior in the test phase as a function of training-phase reward rate relative to the alternative player, stereotype valence, and stereotype morality. Error bars depict standard error.

$$Q_{i,group}^{t+1} = Q_{i,group}^{t} + a_{group}\left(R^{t} - Q_{i,group}^{t}\right)$$

Where $Q_{i,group}$ is the action value of selecting player i from a specific group in trial t, R is the reward received in trial t, and a_{group} denotes the learning rate parameter, which determines the extent to which the prediction error is incorporated into an updated reward value, and which differs by the chosen player's group membership.

As in Schultner et al. (2024), (also Traast et al., 2024, 2025), 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 (see SI for an overview).

Model Fit. A model fit analysis provided the main test of our mechanistic hypothesis. To test model fit, the model is used to simulate trial-by-trial data for each participant based on that participant's specific sequence of task stimuli and behavioral responses in the training and test phases. Model fit is determined by assessing the fit of each participants' actual data to this model-simulated data, using a maximum likelihood approach, and as a function of model complexity, using the Akaike Information Criterion (AIC). Lower AIC scores indicate better fit combined with model simplicity (and thus reduced risk of overfitting).

Supporting our hypothesis, model fit comparisons indicated that the stereotype-learning model provided the best fit to the data (see full model fits in Table S2). Furthermore, as hypothesized, this model fit better to data in the moral stereotype condition ($\text{Mdn}_{\text{AIC}} = 80.99$) relative to the nonmoral stereotype condition ($\text{Mdn}_{\text{AIC}} = 92.09$; see Table S3 in SI).

Model validation. To validate our interpretation of the model fit results, we examined parameter estimates derived from the best-fitting stereotype-learning model as well as simulated test phase data to determine whether these correspond to observed patterns of behavior.

First, a Wilcoxon signed rank test of parameter estimates revealed a stronger parameter value for the prior in the moral stereotype condition ($Mdn = 0.26$, $SD = 39.65$) compared with the nonmoral stereotype condition ($Mdn = -0.01$, $SD = 30.25$), $W = 415$, $p < .045$. This pattern corresponded to the stronger effect of moral stereotypes on initial behavioral choices relative to the nonmoral condition observed in behavior.

Next, we examined learning rate parameter estimates. Although our theorizing would suggest a lower learning rate over time in the moral condition, it is notable that the behavioral data showed a steeper initial change in choice preference (from priors toward 50 %) in the moral condition due to its more extreme initial expectancies, relative to the nonmoral condition. This initial change was a reflected in parameter estimates, such that the learning rate was marginally higher in the moral ($\text{Mdn} = 0.317$, $\text{SD} = 0.35$) than the nonmoral condition ($\text{Mdn} = 0.210$, $\text{SD} = 0.32$), $t(134) = 1.98$, $p = .050$ (see Table S6 in SI for full results). As such, this analysis does not inform our main question regarding the persistence of moral stereotypes; rather, our hypothesis regarding updating is directly supported by the test phase choice behavior reported above. Nevertheless, this pattern of learning rate parameter estimates serves to validate model fit by approximating behavior during the initial stage of learning.

Finally, we examined the simulated test phase choice behavior produced by the stereotype-learning model. These simulated data closely replicated the observed data, illustrating a stronger group effect on choice preference in the moral stereotype condition, relative to the nonmoral condition, $B = 0.29$, $SE = 0.11$, $t = 2.56$, $p = .013$, in addition to a main effect of player reward rate, $B = 0.005$, $SE = 0.0006$, $t = 8.99$, $p < .001$ (Figure S3, panels a and b). These results further demonstrate that the best-fitting model provided a valid account of behavior.

4.4. Discussion

Study 1 supported the hypothesis that moral stereotypes have stronger and more persistent effects on group member impressions than nonmoral stereotypes. First, we found that while stereotypes created divergent expectancies for the behavior of positively- and negatively-stereotyped group members, these expectances were more extreme for moral stereotypes, as evidenced by participants' initial choice preferences during the learning phase of the interaction task. Second, we found that moral stereotypes were more persistent. That is, although moral stereotypes were updated to a degree following initial expectancies, they were never fully updated to match individual players' actual behavior, as shown in test phase choice behavior. By contrast, competence stereotypes were fully updated, such that by the test phase, preferences reflected only individuals group members' behavior and were no longer influenced by the stereotype.

These results revealed an effect of moral stereotypes on group member impressions, extending prior work on individual-level moral content effects to the domain of groups and the effects of stereotypes.

Furthermore, these results demonstrate the effect of moral stereotypes on impressions formed through direct interaction, involving the exchange of choice and feedback, which complements prior work on impression formation based on verbal description or behavioral observation.

Computational modeling provided further insight into the mechanism underlying the enhanced impact of moral stereotypes on interaction-based impression formation. First, model comparison analysis replicated prior findings that stereotypes affect impression formation by creating group-based expectancies, modeled as opposing group priors, and by updating the reward value of individuals according to their respective group-level value representations, modeled as separate group-based learning rates. Second, we found that this model provided a better fit to choice preference formation in the moral stereotype condition than in the competence stereotype condition—a pattern further supported by parameter estimates and model-simulated test phase data. Complementing the behavioral results, these findings suggest that the moral stereotype effect on impression formation and updating reflects the creation of more extreme group expectancies and greater reliance on group-level representations for updating, relative to nonmoral stereotypes.

It is notable that the effect of moral stereotypes on impressions were observed while controlling for any effects in valence associated with moral and nonmoral stereotype descriptions. Valence was controlled through both task design, such that stereotype descriptions were written to be matched on valence, and statistically, whereby valence ratings from an independent pilot study were covaried in our main analyses. This approach allows us to conclude that the observed effects of moral stereotypes were due to the moral content of the stereotype and not merely to differences in valence.

A potential limitation of our approach is that the task context—in which players share points—may be more relevant to moral traits (e.g., generosity) than to competence traits. This form of reward feedback was used because it corresponds most closely to the hypothesized instrumental learning mechanism that underlies interaction-based preference formation (Hackel et al., 2015). Nevertheless, it is possible that moral stereotypes had a stronger effect on reward-based learning because reward feedback may be seen as a moral response. One way in which we addressed this potential confound is to present the task feedback as reflecting the player's prior responses in a past experiment. In this way, the player's responses do not reflect the direct sharing of point with the participant; rather, the participant is predicting which player will share and then receiving feedback on whether that prediction was correct.

To further evaluate whether task relevance presents an alternative explanation for our results, we considered whether our results are consistent with this alternative. A task-relevance account could indeed explain our finding of more extreme initial expectancies in the moral than the nonmoral condition. However, it does not explain our main finding that moral stereotypes are more resistant to updating compared with competence stereotypes. That is, if reward feedback is more relevant to impressions of members from the moral-stereotype group, relative to the competence-stereotype group, then there should be greater updating in the moral condition. Contrary to this alternative, and consistent with our hypothesis, we found that updating was impaired in the moral condition. Furthermore, an explicit motivation in both conditions was to earn money by making choices based on players' actual feedback. This motivation applied equally to both conditions. If reward-based learning were more task relevant in the moral stereotype condition, then this motivation would be more strongly expressed in that condition; however, again, we see the opposite, such that participants in the moral condition were relatively worse at choosing players based on their actual reward feedback.

Finally, it is possible that participants in the moral stereotype condition might have attended less to individual player feedback if they believed that task behavior was more strongly determined by the group's moral stereotype, relative to a competence stereotype in the nonmoral condition. This might explain the lack of updating in the moral condition. However, our data show that participants discerned the relative reward rates of individual players in both the moral and competence conditions; that is, the main effect of player reward rate was not moderated by stereotype morality, valence, or their interaction, and the slopes representing individual player reward learning were similar across conditions (see Fig. 3). This pattern suggests that participants attended to player feedback to a similar extent across conditions, contradicting this alternative explanation.

The results of Study 1 provide initial evidence that moral content in group stereotypes has a stronger effect on impressions of individual group members formed through direct interaction. Furthermore, the influence of moral stereotypes on impressions was more resistant to updating in response to stereotype-disconfirming behavioral feedback from group members that is inconsistent with the stereotype. Our behavioral and computational results, in combination, suggest this moral stereotyping effect is due to its creation of more extreme expectations combined with the tendency to update group member preferences according to a group-level representations as opposed to individual-level representations.

5. Study 2

The goal of Study 2 was to replicate the findings of Study 1 while further examining the implications of these effects for subsequent social judgments decisions and their generalization to novel group members.

5.1. Method

5.1.1. 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 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 for the mixed-factors Stereotype Valence x Stereotype Morality interaction indicated that for $N = 118$, the minimum detectable effect size was $d = 0.08$ ($\alpha = 0.05$, power $= 0.80$).

5.2. Design and procedure

The design and procedure of the social decision task were the same as in Study 1. Following the decision task, participants in this study additionally completed post-task ratings.

Post-task ratings of previously-experienced players. To test whether preferences formed of group members during the decision task generalized to non-economic social decisions, participants rated each of the players encountered during the task regarding their attitudes toward the player and willingness to interact with the player in a noneconomic context. These ratings included likeability (i.e., "*How much do you like [this player]?*"), willingness to work together (i.e., "*How much would you want to work together with [this player]?*"), hiring likelihood (i.e., "*How likely would you be to hire [this player] for a job?")*, and helping likelihood (i.e., "*How likely would you be to help [this player]?*"). Participants made these ratings for each player of both groups. When making each rating, the player's avatar was displayed above the rating item.

Ratings were made on 7-point Likert-type scales ranging from 1 (extremely unlikely/not at all) to 7 (extremely likely/very much). Ratings across the four social decision measures were highly correlated and showed high internal consistency (αs ≥ 0.81). Rather than collapse these ratings into a single average score for each player, separate ratings were included as repeated measures in the multilevel regression model, which had the advantage of controlling for variation between rating type as a factor in the model (this was not preregistered; full model results are reported in SI).

Post-task ratings of novel group members. After rating the group members encountered in the task, participants made the same set of ratings for two novel individuals—one member of each group presented in their condition. Novel group members were not represented by an avatar, but were instead referred to as "another member of Group A or Group B" in the item text (e.g., "we now ask you about whether you would consider interacting with a new member of each group (A and B) you haven't played with in the game"). As with the ratings of familiar group members, the four ratings for each novel group member were highly intercorrelated (αs ≥ 0.83), and thus were included as separate scores within a "rating type" factor in the multilevel regression model for analysis.

5.3. Results

5.3.1. 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 regression model as in Study 1, focusing on the first 30 trials of training phase responses (Fig. 4; see unsmoothed choice behavior in SI).[4] To avoid overfitting, random slopes were excluded for reward rate in this model.

This analysis revealed main effects of relative reward rate, $B = 0.19$, $SE = 0.05$, $z = 4.00$, $p < .001$, $d = 0.79$, stereotype valence, $B = 1.54$, SE $= 0.23$, $z = 6.70$, $p < .001$, $d = 1.27$, and stereotype morality, $B = 0.56$, $SE = 0.19$, $z = 3.04$, $p = .002$, $d = 0.58$, on choice behavior. Importantly given our theoretical question, the Group Valence x Stereotype Morality interaction was significant, $B = -1.06$, $SE = 0.34$, $z = -3.12$, $p = .002$, $d = -0.60$, indicating the valence effect of stereotypes was stronger for moral stereotypes, $B = 1.49$, $SE = 0.23$, $z = 6.44$, $p < .001$, $d = 1.72$, than for nonmoral stereotypes, $B = 0.47$, $SE = 0.24$, $z = 2.00$, $p = .045$, $d = 0.52$. This interaction effect remained significant when stereotyping valence ratings were covaried, $B = -0.53$, $SE = 0.16$, $z = -3.29$, $p < .001$, $d = -0.62$. These results replicated Study 1 findings and again suggest that moral stereotypes had a stronger influence on initial group preferences than nonmoral stereotypes.

5.3.2. 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.[5] This analysis produced main effects of relative reward rate, $B = 0.86$, $SE = 0.09$, $z = 9.79$, $p < .001$, $d = 1.95$, stereotype morality, $B = 1.01$, $SE = 0.29$, $z = 3.47$, $p < .001$, $d = 0.63$, and stereotype valence, $B = 1.33$, $SE = 0.39$, $z = 3.43$, $p < .001$, $d = 0.65$, on choice behavior. Critically, the Stereotype Valence x Stereotype Morality interaction was significant, $B = -1.62$, $SE = 0.57$, $z = -2.84$, $p = .004$, $d = -0.48$ (Fig. 5). Simple effects analyses indicated a valence-based effect of the stereotype only when the stereotype had moral content, $B = 1.29$, $SE = 0.37$, $z = 3.48$, $p < .001$, $d = 0.89$ and not when stereotypes did not have moral content, $B = -0.27$, $SE = 0.44$, $z = -0.61$, $p = .545$, $d = -0.07$. This critical Stereotype Valence x Stereotype Morality interaction remained significant when stereotype valence ratings were covaried, B $= -0.73$, SE $= 0.27$, z $= -2.7$, $p = .007$, $d = -0.6$). Thus, replicating Study 1, these results show that nonmoral stereotypes were updated in response to player feedback whereas moral stereotypes persisted.

Additionally, as in Study 1, the Stereotype Morality x Reward Rate interaction was not significant, $B = -0.21$, $SE = 0.19$, $z = -1.10$, $p = .272$. This result indicated that participants' learning of players' relative reward feedback did not differ between conditions, contradicting the possibility that moral stereotypes reduced participants' attention to individual player feedback.

5.3.3. Computational modeling

As in Study 1, we used computational modeling to examine the cognitive mechanisms involved in the observed stereotype effects.

Model fit. Model fit comparisons again determined that behavioral choice data were fit best by the stereotype-learning model, which includes a group-based biased prior and separate learning rates for each group, relative to all alternative models. Replicating Study 1, we again observed better fit to this model in the moral stereotype condition ($\text{Mdn}_{\text{AIC}} = 80.99$) relative to the nonmoral stereotype condition ($\text{Mdn}_{\text{AIC}} = 91.25$; see Table S4 in SI).

Model validation. As in Study 1, we validated our interpretation of the model by examining its parameter estimates and simulated test-phase data. Replicating Study 1, parameter estimates for group-based priors were larger for moral stereotypes ($Mdn = 0.13$, $SD = 17.20$) than for nonmoral stereotypes ($Mdn = -0.01$, $SD = 14.40$), $W = 1227$, $p = .006$, matching observed behavior. In this study, learning rate parameters did not differ significantly as a function of stereotype morality ($Mdn_{\text{moral}} = 0.317$, $SD = 0.35$, $Mdn_{\text{competence}} = 0.21$, $SD = 0.32$), $t(228) = 0.027$, $p = .636$ (see SI). However, learning rates were numerically higher in the moral condition, directionally consistent with Study 1 and again matching the pattern of a larger change in choice preference following initial expectancies in the moral condition relative to the nonmoral condition.

Finally, as in Study 1, simulated data replicated the observed data, showing a stronger group effect on choice preference in the moral stereotype condition compared with the nonmoral condition, $B = 0.198$, $SE = 0.08$, $t = 2.47$, $p = .015$, and a main effect of player reward rate, $B = 0.006$, $SE = 0.0005$, $t = 11.70$, $p < .001$ (Figure S3, panels c and d). Together, these results support the model's validity and interpretation, replicating Study 1.

5.3.4. Post-task ratings

Next, we asked whether the preferences formed during the social decision task generalized to non-economic preference judgments and social decisions concerning both players and novel group members.

Ratings of Previously-Experienced Players. To test whether stereotype-based group preferences generalized to non-economic judgments and decisions concerning players of the task, we fit a mixed effects regression predicting participants' ratings. Predictors included (a) player reward rate (standardized and centered), (b) stereotype valence, (c) stereotype morality, (d) rating type, and (e) 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 random slopes for type of rating led to singular fit).

This analysis produced a significant main effect of reward rate on ratings, $B = 0.38$, $SE = 0.04$, $t = 8.55$, $p < .001$, $d = 1.58$, such that participants reported more positive preferences for players who were more rewarding in the main task, controlling for the type of rating.

There was also a main effect of stereotype valence, $B = 0.35$, $SE = 0.14$, $t = 2.59$, $p = .011$, $d = 0.47$, whereby players from positively-stereotyped groups were rated more positively than players from

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

[5] 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 not found and was then dropped from the preregistered replication of the main task (presented here as Study 1); the issue is discussed in the General Discussion. (Note that Studies 1 and 2 were re-ordered because Study 2 included additional measures of generalization; aside from these measures, the studies are identical.)

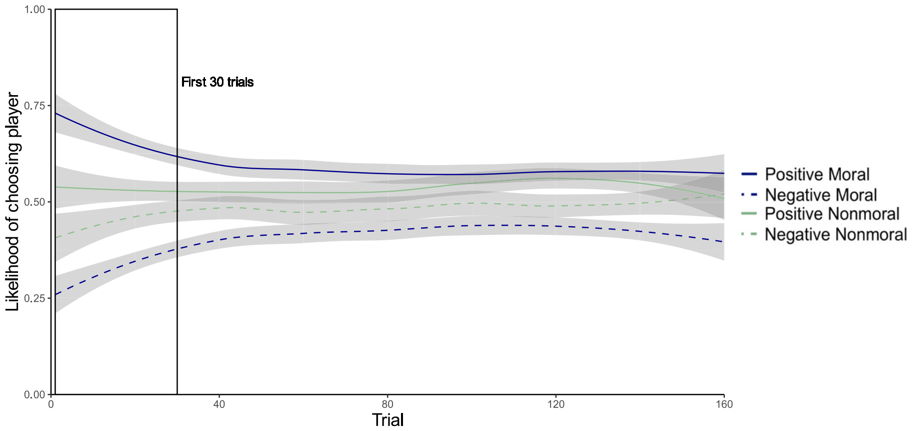

Fig. 4. Study 2 choice behavior during training phase over time.

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 box indicates the first 30 trials of the training phase and grey shading shows confidence intervals.

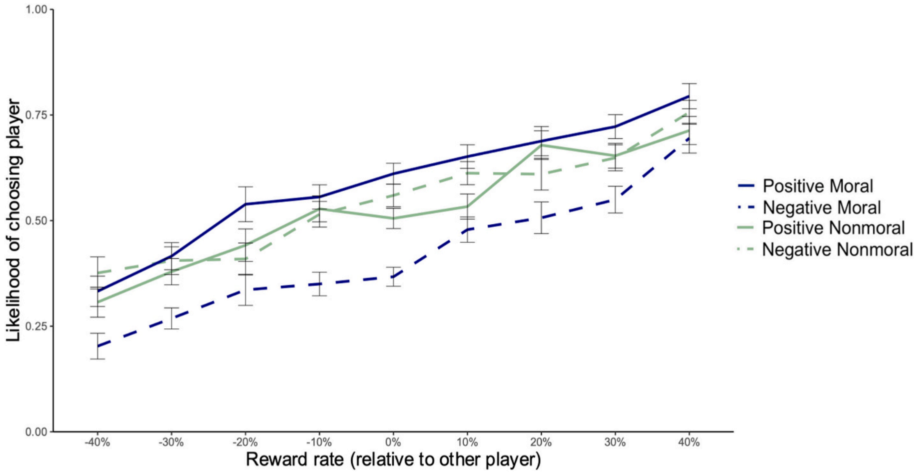

Fig. 5. Study 2 choice behavior during test phase.

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

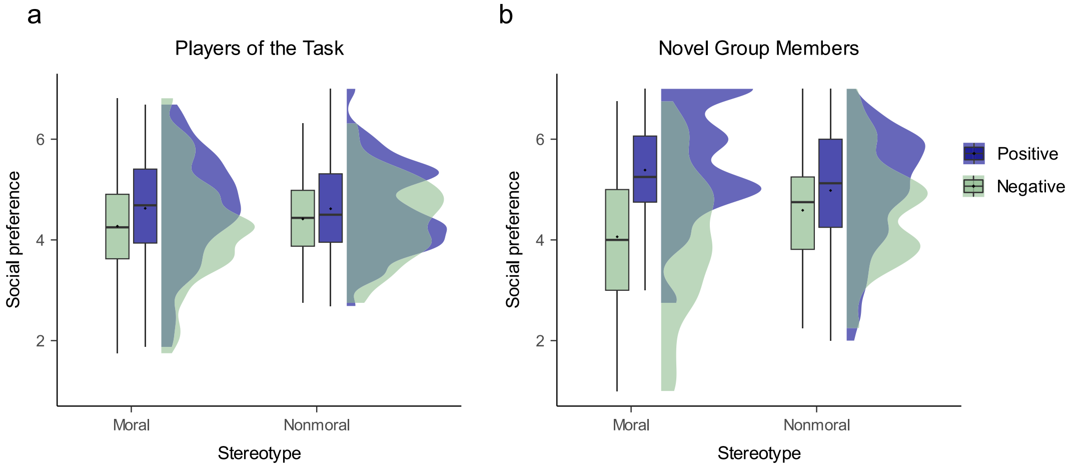

Fig. 6. Post-task ratings of players and novel group members.

Note. Social preferences toward (a) familiar group members from the task and (b) novel group members, as a function of the Stereotype Morality x Stereotype Valence interaction, showing boxplots, mean scores, and half-density distributions.

negatively-stereotyped groups (Fig. 6). However, the interaction between stereotype valence and stereotype morality was not significant, $B = -0.14$, $SE = 0.20$, $t = -0.72$, $p = .474$, $d = -0.13$, indicating that moral content did not have a greater influence on self-reported preferences for future interaction, relative to competence stereotypes, in contrast to the pattern observed for choice behavior.

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 rate, since these novel targets did not have a reward history). This analysis produced main effects of stereotype valence, $B = 1.32$, $SE = 0.21$, $t = 6.34$, $p < .001$, $d = 1.17$, and stereotype morality, $B = 0.53$, $SE = 0.24$, $t = 2.21$, $p = .029$, $d = 0.41$. Importantly, these main effects were qualified by a Stereotype Valence x Stereotype Morality interaction, $B = -0.93$, $SE = 0.31$, $t = -3.01$, $p = .003$, $d = -0.55$ (Fig. 6). This interaction suggested that the impact of stereotypes on preferences toward novel group members was stronger in the moral condition than in the nonmoral condition: simple effects indicated a significant effect of stereotype in the moral condition, $B = 1.32$, $SE = 0.22$, $t = 6.10$, $p < .001$, $d = 1.53$, but not in the competence condition, $B = 0.39$, $SE = 0.22$, $t = 1.82$, $p = .074$, $d = 0.50$.

Interestingly, unlike behavioral choice preferences observed in the interaction task, these self-reported preferences appeared to show an asymmetry effect—a larger impact of negative than positive moral stereotypes: preferences were significantly worse toward novel members of groups with negative moral stereotypes ($M = 4.06$, $SD = 1.60$) than those with negative competence stereotypes ($M = 4.59$, $SD = 1.37$), $B = 0.525$, $SE = 0.24$, $t = 2.21$, $p = .029$, but preferences toward novel members of positively stereotyped groups differed less as a function of moral content, ($M_{\text{moral}} = 5.39$, $SD_{\text{moral}} = 1.22$, $M_{\text{competence}} = 4.98$, $SD_{\text{competence}} = 1.35$), $B = -0.405$, $SE = 0.20$, $t = -2.00$, $p = .047$. We discuss this asymmetry further in the General Discussion.

Finally, we tested whether choice preferences formed toward group members in learning task directly predicted preferences toward novel group members. A regression analysis indicated that test phase group preference, across morality and valence condition (as these effects are already represented in group preference scores), did indeed predict future interaction preferences toward novel group members, $B = 1.38$, $SE = 0.48$, $t = 2.90$, $p = .004$, $d = 0.54$. This result demonstrates that social-interactive instrumental learning about existing group members, which is enhanced by moral stereotypes, generalizes to reported preferences for novel group members.

5.4. Discussion

Study 2 had two aims: to replicate the effects of Study 1 and to examine their implications for subsequent social judgments. First, the results of Study 2 closely matched those of Study 1. Behavioral data showed that moral stereotypes had a stronger influence on initial expectancies for group members' behavior and induced impressions that were more resistant to updating, relative to nonmoral stereotypes. Also, as in Study 1, computational model fits indicated that the effect of stereotypes on choice preferences was best explained by a model that included divergent group expectancies combined with updating according to separate group representations. Moreover, consistent with our main hypothesis, this model characterized choice behavior observed in the moral stereotype condition more closely than in the nonmoral stereotype condition. Together, these findings replicated Study 1 and further demonstrated the enhanced effect of moral stereotypes on impression formation and updating.

Extending the aims of Study 1, Study 2 tested whether moral stereotype effects on instrumental learning generalized to judgments of both familiar and novel group members in new contexts. When making social judgments about existing group members regarding future interaction, with whom participants had extensive experience, these post-task judgments reflected players' actual reward feedback, as in prior research (Hackel et al., 2015, 2020, 2022), as well as the positivity or negativity of the stereotype associated with their group, a novel finding. It is possible that, when making more deliberative self-report judgments about these players, participants could more effectively apply relevant individual-specific information to their judgments, in comparison to the binary approach/avoid-type decisions required for their choice behaviors in the learning task.

By contrast, judgments of novel group members were influenced by the moral content of the stereotype, in addition to stereotype valence. This pattern may reflect that fact that participants could not draw from prior experiences with novel members and thus relied more heavily on the group stereotype when making judgments. This interpretation is consistent with our finding that judgments of novel group members were associated with the reward associations participants had formed toward existing group members, but that this influence was relatively smaller than the effect of moral stereotypes. An intriguing implication of this finding is that while participants could revise their self-reported impressions of group members following extensive experiences with them, their self-reported judgments of novel members continued to reflect the stereotype.

6. General discussion

Social stereotypes are often moral in tone. We asked whether moral content is what leads many 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 divergent group-based expectancies and then updating impressions according to separate group-level representations—but that 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 demonstrate that it can produce the extreme and persistent effects often associated with racial and ethnic stereotypes.

In addition to these main findings, this research contributes broader advances to the study of intergroup attitudes. First, it presents a novel integration of theory and questions from the stereotyping, impression formation, and moral psychology literatures to address the societal phenomenon of moralized group portrayals. Second, it examined stereotype effects on impression formation in the context of direct socio-interactive learning, in which participants formed impressions of group members through repeated rounds of choice and feedback—an active form of impression formation that complements previously-studied passive (e.g., instructional or observational) forms of impression formation (Amodio, 2019). Third, this research employed a combination of behavioral experiments, involving repeated exposure to target behaviors and the repeated measure of updating, with computational models that formalized a specific theory of stereotype effects on learning. By adapting this approach to the context of stereotyping and moral impression formation, it offers greater precision in the measurement of impression updating and in the test of theoretical mechanism. We elaborate on these contributions in what follows.

6.1. 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 proposed that moral stereotypes may be considered more diagnostic of a group and thus more essential to the group's identity (Brambilla et al., 2019; Goodwin, 2015). Furthermore, like moral traits, moral stereotypes may be viewed as "other-profitable," in the sense that they have greater implications for others who interact with a target, relative to competence-related traits that may be primarily relevant to targets themselves (Wojciszke, 2005). 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 resistant 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. Although moral components of stereotype content have long been recognized (Phalet & Poppe, 1997), it is their influence on impression formation that may help to explain their impact on prejudice and discrimination.

6.2. *Valence effects on moral impression formation*

Because moral traits are especially relevant to perceivers, due to their diagnosticity about a target's character as well as the other-profitability of the target's behavior, moral traits are often perceived as more extreme in valence than nonmoral traits (Brambilla et al., 2019; Cone & Ferguson, 2015; Peeters & Czapinski, 1990; Reeder & Coovert, 1986; Skowronski & Carlston, 1987; Wojciszke et al., 1998, Wojciszke, 2005); thus it may be difficult to distinguish effects of moral content from valence on impressions, raising the question of whether our findings—that moral stereotypes induce stronger expectancies and are more persistent than nonmoral stereotypes—is simply due to moral stereotypes being more extreme in valence. The present work addressed this issue in three ways. First, the stereotype messages were designed to be matched in valence, and independent ratings of these stereotype descriptions' content confirmed that negative moral and nonmoral stereotypes did not differ in extremity, although positive moral and nonmoral stereotypes differed slightly. Second, to control for any difference in valence between moral and competence stereotypes, we statistically adjusted for these valence ratings and found that our results remained robust (see SI), replicating prior research in which moral attitude effects remained after adjusting for attitude strength (e.g., Luttrell et al., 2022; Skitka et al., 2005). 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. These additional findings strengthen our conclusion that moral content enhances the impact and durability of stereotypes beyond any effect of valence. Importantly, however, while this approach isolates the effect of moral content, it does not rule out a role for valence as part of the process through which moral content influences impression formation and updating.

An aspect of moral impression formation that was not emphasized in this research is the asymmetrical impact of negative information, relative to positive information, on impressions (Cone & Ferguson, 2015; Skowronski & Carlston, 1987)—an effect that, in some research, is enhanced for moral content (Mende-Seidlecki et al., 2013). We speculate that we did not observe this valence asymmetry because our primary measure of preference was a dichotomous behavioral choice between members of a positively- or negatively-stereotyped group. In such choice decisions, a preference toward one group is a nonpreference for the other. By contrast, prior observations of the asymmetry effect come from independent assessments of positive and negative impressions, such as those based on self-reports or implicit tasks. This difference in approach is not a methodological limitation, but rather reflects the nature of different expressions of a preference—that is, in preference judgments as opposed to choice behaviors. However, consistent with this explanation, it is notable that a valence asymmetry was observed in participants' self-reported preferences toward novel group members in Study 2, such that negative moral stereotypes had a stronger influence on preferences than positive moral content, relative to the nonmoral condition. Thus, our findings are not inconsistent with prior evidence for valence asymmetry in moral impressions but rather extend them, suggesting that such asymmetries are less likely to be expressed in behavioral decisions that involve a dichotomous choice.

6.3. *Constraints on generalizability*

Although our experiments were designed to permit tightly-controlled tests of our hypotheses regarding the effects of moral stereotypes on impression formation, it is important to consider the limitations of this approach for generalizability. One limitation concerns the generalization of the stereotypes used in our studies. These specific descriptions were designed to manipulate the moral content of stereotypes, and therefore they might not reflect the complexity or specific characteristics associated with stereotypes for real groups in different cultures and societies.

A second potential limitation concerns the ecological validity of the task procedure. The procedure prioritized construct validity and internal validity, such that is was designed to be engaging and believable while precisely manipulating and controlling theoretical variables of interest. Our prioritization of construct and internal validity limits the task's ecological validity, in that the nature of interactions with players was minimal (although no participant reported suspicion about these manipulations). While this procedure served as an experimentally-controlled proxy for real-world social interactions, everyday interactions involve more nuance and complexity, and impressions formed through interaction incorporate other kinds of information beyond stereotypes and feedback such as existing knowledge about a person and one's emotional responses (Amodio, 2025). Further research is needed to understand how instrumental learning interacts with these other sources of information to shape interaction-based impressions.

7. Conclusion

Social stereotypes characterized by moralized content are often the most insidious. Our findings provide direct evidence for this observation and describe its underlying mechanism: we found that moral stereotypes do indeed have stronger and more persistent effects on impression formation of group members than stereotypes without a moral component, and that this pattern is due to a heightened effect of moral stereotypes on initial expectations group members' behavior and how one interprets and learns from group members' actual responses. These findings identify a potent effect of moral stereotypes on impression formation and updating and suggest that efforts to reduce their effects may require targeting this moral content.

Open science 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/. 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.

CRediT authorship contribution statement

Inga K. Rösler: Writing – review & editing, Writing – original draft, Visualization, Resources, Project administration, Methodology,

Investigation, Formal analysis, Data curation, Conceptualization. **Isabel Kerber:** Investigation, Data curation. **David M. Amodio:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Funding

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

Declaration of competing interest

Authors declare no conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2025.104750.

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. (2025). Learning and memory mechanisms underlying impression formation and updating. *Nature Reviews Psychology*.

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/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*.

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

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

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), Article 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), Article 103948. https://doi.org/10.1016/j.jesp.2019.103948

Hamilton, D. L., Sherman, S. J., & Ruvolo, C. M. (1990). Stereotype-based expectancies: Effects on information processing and social behavior. *Journal of Social Issues, 46*, 35–60. https://doi.org/10.1111/j.1540-4560.1990.tb01922.x

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*, 237–271. 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. (2017b). An R2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics, 44*, 1086–1105. https://doi.org/10.1080/02664763.2016.1193725

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017a). An *R*2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics, 44*(6), 1086–1105. https://doi.org/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

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology, 93*(2), 234–249. https://doi.org/10.1037/0022-3514.93.2.234

Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes, 126*, 88–106.

Luttrell, A., Petty, R. E., Briñol, P., & Wagner, B. C. (2016). Making it moral: Merely labeling an attitude as moral increases its strength. *Journal of Experimental Social Psychology, 65*, 82–93. https://doi.org/10.1016/j.jesp.2016.04.003

Luttrell, A., Sacchi, S., & Brambilla, M. (2022). Changing impressions in competence-oriented domains: The primacy of morality endures. *Journal of Experimental Social Psychology, 98*, Article 104246. https://doi.org/10.1016/j.jesp.2021.104246

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social cognitive and affective neuroscience, 8*, 623–631.

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

Peeters, G., & Czapinski, J. (1990). Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology, 1*, 33–60.

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., … 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

R Core Team. (2024). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/.

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* (pp. 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

Traast, I. J., Schultner, D. T., Doosje, B., & Amodio, D. M. (2024). *Race effects on impression formation in social interaction: An instrumental learning account*. https://doi.org/10.31234/osf.io/3j2rm

Traast, I. J., Doosje, B., & Amodio, D. M. (2025). Impression formation through social interaction: The effect of ethnicity in the Dutch context. *Group Processes & Intergroup Relations*. https://doi.org/10.1177/13684302241305054. in press.

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), Article e48693. https://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. (2005). Morality and competence in person- and self-perception. *European Review of Social Psychology, 16*(1), 155–188. https://doi.org/10.1080/10463280500229619

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