Effects of Moral Stereotypes on the Formation and Persistence of Group Preferences: **Supplementary Information**

Supplemental Methods

Stereotype descriptions

Moral stereotypes. 'Individuals from Group [*A/B*] grew up in a society with a government [*low/high*] in corruption and [*low/high*] in transparency. Members of this society are known for being [*moral/immoral, trustworthy/untrustworthy, honest/dishonest, and fair/unfair*].'

Nonmoral stereotypes. 'Individuals from Group [*A/B*] grew up in a society with an education system with [*low/high*] inefficiency and [*low/high*] in performance. Members of this society are known for being [*competent/incompetent, successful/unsuccessful, intelligent/unintelligent, and ambitious/unambitious*].'

Supplemental Results

Stereotype Valence Validation Study Method

To assess the possibility of a valence extremity difference between the moral and nonmoral descriptions, a separate sample (N = 100, M_{age} = 36.83, SD_{age} = 11.75; 47 female, 50 male, 2 non-binary/third gender, 1 prefer not to say; 78% White/Caucasian, 8% Black or African, 7% Asian/ Pacific Islander, 3% Hispanic, 2% Multiple ethnicities/ Other, 2% prefer not to say) was recruited to rate each of the four stereotype descriptions (moral-positive, moral-negative, nonmoral-positive, nonmoral-negative). In accordance with the experimental design, stereotype valence was manipulated within-subjects and stereotype morality between-subjects and order was counterbalanced.

Stereotype Valence Covariate Analyses

To address the possibility that the effect of moral vs. nonmoral stereotypes was due to more extreme valence in moral stereotypes, we reran analyses in the main text while adjusting for stereotype valence ratings obtained in the independent validation study. This covariate analysis was performed in each study for the Stereotype Morality x Stereotype Valence interactions observed in (a) the first 30 trials of learning and (b) test phase choices.

To perform this analysis in the glmm framework, we restructured the model reported in the main text to regress stereotype valence-consistent choice (i.e., choosing positively stereotyped groups over negatively stereotyped groups, or vice versa) onto additive terms including morality and pretest stereotype valence. Here, the prediction was for a morality main effect, which in this restored analysis was analogous to the Stereotype Morality x Stereotype Valence interaction reported in the main test. This restructuring yielded a degree of freedom required for the inclusion of the stereotype valence covariate. In each case, as reported in the main text,

the interaction effect remained significant, suggesting that the effect of morality could not be explained by differences in valence.

Computational model fits and parameter values

Overview of models. In Tables S1 and S2 below, our hypothesized stereotype learning model—which includes a group-based symmetrical prior and separate group learning rates—is listed as Model 4.

Models 1-3 refer to models described in the main text. Model 1 is a standard Q-learning model. Model 2 includes a group-based prior but a single learning rate. Model 3 includes separate learning rates for each group but not prior.

Models 5-10 include variations of these models that distinguish between learning rates for gains and loses; none of these fit better than the main stereotype-learning model.

Table S1

Overview of RL Models and Parameters Studies 1 & 2

Model Number	Numbers of Parameter	Parameters
1	2	α,β
2	3	α,β,P
3	3	$\alpha_{Pos},\alpha_{Neg},\beta$
4	4	$\alpha_{Pos},\alpha_{Neg},P,\beta$
5	3	$\alpha^{+},\alpha^{-},\beta$
6	4	$\alpha^{+},\alpha^{-},P,\beta$
7	5	$\alpha_{Pos}^{+},\alpha_{Pos}^{-}$, α_{Neg}^{+}, α_{Neg}^{-},β
8	6	$\alpha_{Pos}^{+},\alpha_{Pos}^{-}$, $\alpha_{Neg}^{+},\alpha_{Neg}^{-},P$, β
9	3	$\alpha_{PE}^{+},\alpha_{PE}^{-},\beta$
10	6	$\alpha_{Pos}^{+},\alpha_{Pos}^{-}$, α_{Neg}^{+}, α_{Neg}^{-}, C, β

Note. Each model includes the basic parameters α (learning rate) and β (inverse temperature). Models 2-10 add P (prior), $\alpha_{Pos/Neg}$ (distinct learning rate positively/negatively stereotyped), $\alpha^{+/-}$ (distinct learning rate rewards/losses), $\alpha_{PE+/-}$ (distinct learning rate prediction error rewards/losses), combinations between learning rates (e.g. α_{Pos}^{+}, distinct learning rates for rewards from positively stereotyped group members), and C (confirmation bias parameter, C, $1 \leq C \leq 10$), which amplifies gains and reduces losses for interactions with the positively-stereotyped group)

Table S2

Mean and Median Model Fit Indices by Model and Study

	Study 1				Study 2			
	AIC		BIC		AIC		BIC	
Model Number	Mean	Median	Mean	Median	Mean	Median	Mean	Median
1	115.05	120.99	120.11	126.10	115.90	120.96	120.95	126.05
2	83.38	90.25	90.98	96.59	93.63	103.07	101.21	110.56
3	102.56	108.90	110.17	116.60	108.14	116.35	115.72	124.04
4	**79.36**	**88.00**	**89.50**	**98.13**	**89.33**	**98.37**	**99.44**	**108.58**
5	90.87	99.50	98.48	107.19	99.09	106.05	106.67	113.71
6	82.97	88.90	93.11	99.16	92.29	100.24	102.40	110.48
7	84.01	89.47	96.69	102.08	91.85	100.16	104.48	112.96
8	81.72	85.35	96.93	100.74	91.43	99.99	106.59	115.35
9	83.68	90.44	91.29	98.14	92.69	102.57	100.27	109.92
10	101.53	117.13	114.21	129.96	108.23	121.15	120.86	133.94
11	102.20	111.84	117.41	127.10	110.07	119.00	125.23	134.17

Note. Each model was fit 50 times using random starting points. Best fitting model in bold. For model 11 in Study 2 two participants were excluded because the estimated C parameters did not fall into the range of our parameter boundaries.

Table S3

Mean and Median Model Fit Indices by Model and Stereotype Morality Study 1

	AIC				BIC			
	Moral		Non-moral		Moral		Non-moral	
Model Numb	Mean	Median	Mean	Median	Mean	Median	Mean	Median
1	115.28	118.35	114.88	121.53	120.39	123.48	119.91	126.61
2	78.46	87.82	86.94	94.08	86.13	95.39	94.5	101.78
3	100.16	106.01	104.3	116.86	107.83	113.71	111.86	124.52
4	**73.99**	**80.99**	**83.24**	**92.09**	**84.22**	**91.25**	**93.32**	**101.44**
5	83.94	94.92	95.90	105.90	91.61	102.61	103.46	113.58
6	77.20	87.28	87.16	98.86	87.43	97.36	97.24	109.03
7	77.99	83.92	88.38	96.09	90.77	96.74	100.97	108.74
8	76.01	79.4	85.86	94.15	91.34	94.60	100.97	108.18
9	78.67	88.09	87.35	95.3	86.34	95.76	94.9	102.95
10	90.87	107.52	109.27	121.72	103.65	120.19	121.87	134.33
11	93.26	108.59	108.68	123.8	109.69	125.08	116.15	131.51

Note. Each model was fit 50 times using random starting points. Best fitting model in bold.

Table S4

Mean and Median Model Fit Indices by Model and Stereotype Morality Study 2

	AIC				BIC			
	Moral		Non-moral		Moral		Non-moral	
Model Numb	Mean	Median	Mean	Median	Mean	Median	Mean	Median
1	115.07	120.66	116.88	122.07	121.95	125.75	120.1	127.16
2	91.87	99.24	95.69	104.88	99.42	106.75	103.3	112.14
3	109.09	115.25	107.02	117.29	116.64	122.93	114.63	124.81
4	**87.52**	**95.02**	**91.47**	**103.35**	**97.59**	**105.21**	**101.62**	**113.3**
5	96.11	105.06	102.62	106.86	103.65	112.31	110.23	114.54
6	90.25	95.42	94.71	104.27	100.32	105.62	104.86	114.22
7	89.85	97.16	94.21	105.04	102.43	109.90	106.9	117.47
8	89.26	90.64	94.00	103.72	104.36	105.74	109.23	119.07
9	90.76	94.25	95.06	106.79	98.31	101.91	102.67	114.26
10	106.04	119.12	110.83	124.37	118.62	131.94	123.51	136.8
11	108.8	114.12	111.54	125.6	123.89	129.41	126.76	140.52

Note. Each model was fit 50 times using random starting points. Best fitting model in bold. For model 11 two participants were excluded whose C parameters did not fall into the range of our parameter boundaries.

Table S6

Median parameter values for best fitting RL models per study

	Study 1		Study 2	
Parameters	Moralized	Non-moralized	Moralized	Non-moralized
Learning rate positively stereotyped	0.13	0.02	0.07	0.01
Learning rate negatively stereotyped	0.16	0.02	0.07	0.05
Prior	0.26	-0.01	0.13	-0.01
Beta	0.19	0.05	0.19	0.14

Table S7

Overview Results Posttask Measures Combined Measures for Players of the Task Study 2

Predictors	Estimates	CI	p
(Intercept)	4.63	4.38 – 4.88	**<0.001**
Reward rate	0.38	0.29 – 0.47	**<0.001**
Dimension	0.14	-0.21 – 0.50	0.424
Group Valence	0.35	0.09 – 0.61	**0.010**
Scale Hire	-0.43	-0.53 – -0.33	**<0.001**
Scale Like	-0.52	-0.62 – -0.42	**<0.001**
Scale Work	-0.49	-0.59 – -0.39	**<0.001**
Dimension * Group Valence	-0.14	-0.53 – 0.24	0.472
Random Effects			
σ^2	1.24		
$\tau_{00\ id}$	0.89		
$\tau_{11\ id.sharing\_s}$	0.20		
$\tau_{11\ id.group\_facePositive}$	1.04		
ρ_{01}	-0.06		
	-0.52		
ICC	0.47		
N_{id}	118		
Observations	3776		
Marginal R^2 / Conditional R^2	0.083 / 0.515		

Table S8

Overview Results Posttask Measures Combined Measures for Novel Group Members Study 2

Predictors	Estimates	CI	p
(Intercept)	4.47	4.14 – 4.80	**<0.001**
Reward rate	0.53	0.06 – 0.99	**0.027**
Group Valence	1.32	0.91 – 1.73	**<0.001**
Scale Hire	-0.61	-0.75 – -0.47	**<0.001**
Scale Like	-0.58	-0.72 – -0.44	**<0.001**
Scale Work	-0.44	-0.59 – -0.30	**<0.001**
Dimension * Group Valence	-0.93	-1.54 – -0.32	**0.003**
Random Effects			
σ^2	0.63		
$\tau_{00\ id}$	1.49		
$\tau_{11\ id.group\_valencepositive}$	2.48		
$\rho_{01\ id}$	-0.76		
ICC	0.67		
N_{id}	118		
Observations	944		
Marginal R^2 / Conditional R^2	0.143 / 0.715		

Table S9

Significance of Interaction Effect Stereotype Valence x Stereotype Dimension in Training Phase Data

	Study 1					Study 2				
	10 trials	20 trials	30 trials	50 trials	All trials	10 trials	20 trials	30 trials	50 trials	All trials
Stereotype Valence x Stereotype Dimension	0.001**	0.024 *	0.027 *	0.008 **	0.002 **	< 0.001 ***	< 0.001 ***	0.002 **	< 0.001 ***	0.028 *

Note. The models of training phase data for 10 and 20 trails of Study 1 and for 10 trails of Study 2 had singular fit, excluding random effects did not resolve this issue.

Figure S1

Unsmoothed Choice Behavior of the Training Phase Over Time Study 1

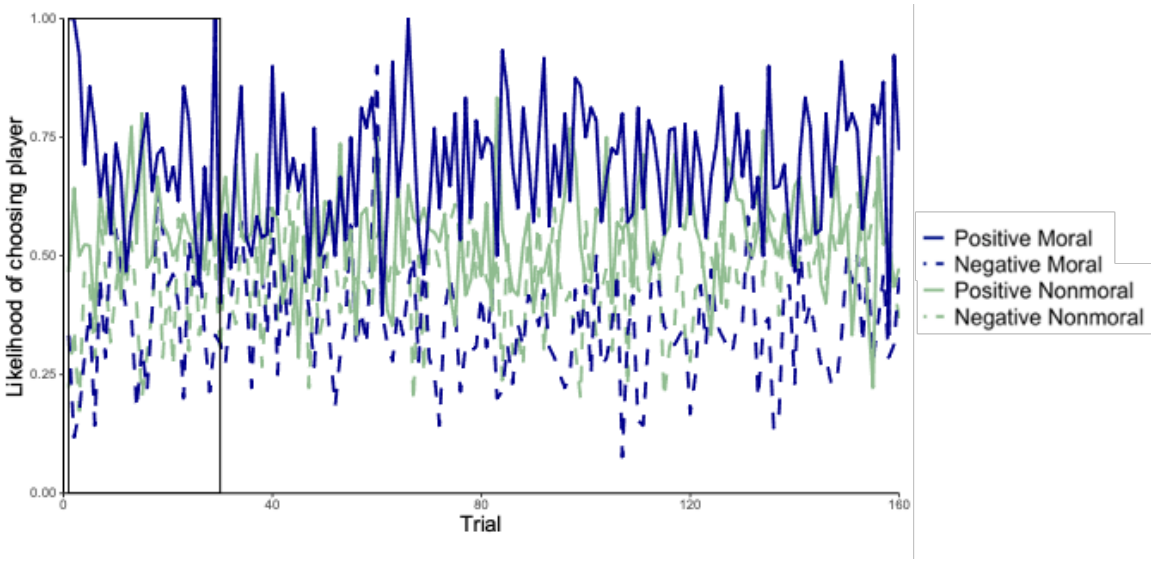


Figure S2

Unsmoothed Choice Behavior of the Training Phase Over Time Study 2

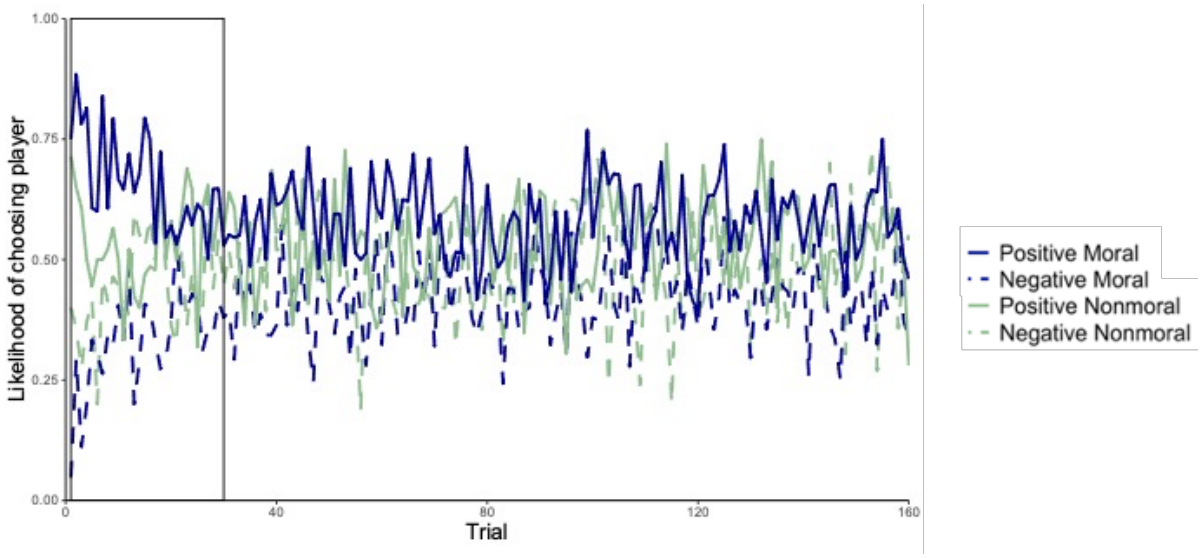
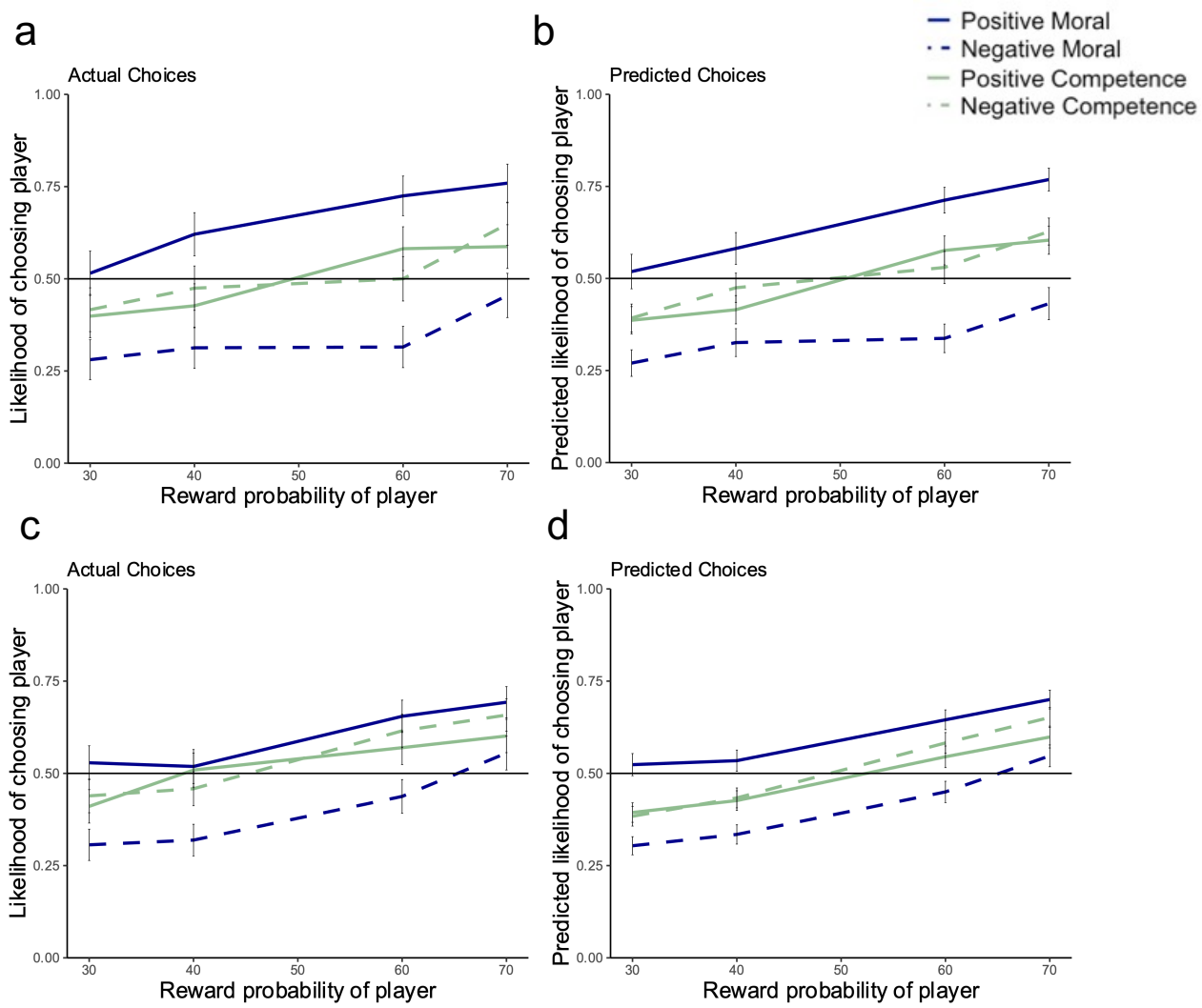


Figure S3

Predictions Test Phase Choices From the Best Fitting RL Model (Studies 1 & 2)



Note. Actual (a) and predicted test phase choices (b) of Study 1 and actual (c) and predicted test phase choices (d) of Study 2 by the reward probability of a player. The best-fitting model combined biased group-based preferences (i.e., priors) and learning (i.e., learning rates).