Running Head: Stereotypes and Learning

The power of the unexpected:

Prediction errors enhance stereotype-based learning

Johanna K. Falbén,[1,2] Marius Golubickis,[1] Dimitra Tsamadi,[1]
Linn M. Persson,[1] C. Neil Macrae[1]

[1]School of Psychology, University of Aberdeen, Aberdeen, Scotland, UK
[2]Department of Psychology, University of Warwick, Coventry, England, UK

Address Correspondence To:

Johanna Falbén
Department of Psychology
University of Warwick
Coventry CV4 7AL
England, UK

Email: johanna.falben@warwick.ac.uk

Abstract

Stereotyping is a ubiquitous feature of social cognition, yet surprisingly little is known about how group-related beliefs influence the acquisition of person knowledge. Accordingly, in combination with computational modeling (i.e., Reinforcement Learning Drift Diffusion Model analysis), here we used a probabilistic selection task to explore the extent to which gender stereotypes impact instrumental learning. Several theoretically interesting effects were observed. First, reflecting the impact of cultural socialization on person construal, an expectancy-based preference for stereotype-consistent (vs. stereotype-inconsistent) responses was observed. Second, underscoring the potency of unexpected information, learning rates were faster for counter-stereotypic compared to stereotypic individuals, both for negative and positive prediction errors. Collectively, these findings are consistent with predictive accounts of social perception and have implications for the conditions under which stereotyping can potentially be reduced.

Keywords: stereotyping, person perception, reinforcement learning, prediction errors, drift diffusion model.

Declarations

Funding: None.

Conflicts of interest/Competing interests: None.

Data availability: Available on OSF https://osf.io/9ajcz/

Code availability: Available on OSF https://osf.io/9ajcz/

The power of the unexpected:

Prediction errors enhance stereotype-based learning

1. Introduction

Stereotypes exert a pervasive influence on both thinking and doing. Although several reasons have been advanced for why this may be the case, one account has dominated contemporary theorizing on the topic, stereotyping spares people the trouble of thinking deeply about others (Allport, 1954; Brewer, 1988; Correll et al., 2017; Fiske & Neuberg, 1990; Freeman & Ambady, 2011; Hilton & von Hippel, 1996; Macrae & Bodenhausen, 2000). Corroborated by decades of research — in challenging task settings and absent the operation of offsetting motives or goals — stereotyping streamlines core aspects of social-cognitive functioning, including resource allocation, memorial processing, and impression formation (e.g., Bodenhausen & Lichtenstein, 1987; Correll et al., 2015; Eberhardt et al., 2004; Falbén et al., 2019; Macrae et al., 1994; Stern et al., 1984). Specifically, once activated, category-related expectancies facilitate the detection, encoding, and accessibility of confirmatory (i.e., stereotype-consistent) material, thereby driving the stereotype-based outcomes that punctuate interpersonal and intergroup exchanges.

Beyond the economizing effects that stereotypes exert on information processing and decision-making, a complementary line of inquiry has focused on expectancy-violating individuals (i.e., counter-stereotypes) and the pivotal role they play in the reduction of group-based inequalities (Diekman & Eagly, 2000; Fitzgerald et al., 2019; Olsson & Martiny, 2018; Wood & Eagly, 2012). Advocating that members of minority/disadvantaged groups should occupy positions of prominence and authority in all corners of society, it has been suggested that counter-stereotypes are an essential tool in the drive to eliminate discriminatory practices and create equal opportunities for all (Dennehy & Dasgupta, 2017; Eagly & Steffen, 1984; Morgenroth et al., 2015). To give but a single pertinent example, by promoting the acceptability of non-traditional life choices, gender-incongruent individuals (e.g., women in positions of power, men in nurturing professions) enable children and

adolescents to overcome the psychological barriers that otherwise limit their educational and occupational ambitions (Olsson & Martiny, 2018; Wood & Eagly, 2012). Put simply, role models generate role aspirants.

Lending empirical support to this viewpoint, laboratory and field research have revealed that encountering or merely imagining incongruent individuals weakens stereotype-based responding across a range of tasks and measures (e.g., Beaman et al., 2009; Dasgupta & Asgari, 2004; Hastie et al., 1990; Kunda et al., 1990; Prati et al., 2015; Rudman & Phelan, 2010). For example, exposure to women in counter-stereotypic leadership roles reduces the expression of implicit gender stereotypes (Dasgupta & Asgari, 2004). The power of counter-stereotypes derives from their challenge to prevailing cultural beliefs, thus informational value. Exposed to an endless stream of stereotype-consistent individuals (e.g., women performing domestic functions, men studying the sciences), people's extant knowledge and understanding of the world remains intact and unopposed. Throw some counter-stereotypes into the mix, however, and the situation suddenly changes, such that prior category-related convictions no longer adequately capture external reality, thereby triggering the modification of stereotype-based beliefs (Hinton, 2017; Otten et al., 2017; Wood & Eagly, 2012).

The utility of unexpected information — as exemplified by counter-stereotypes — has been widely acknowledged across the psychological and neurosciences (Johnston & Hawley, 1994; McClelland et al., 1995). One of the brain's foremost capacities is the ability to learn from previous experiences to generate predictions about future states of the world (Bar, 2007; Clark, 2013; O'Callaghan et al., 2017; Otten et al., 2017). Computationally, the precision of these forecasts is enhanced when mismatches are detected between expected and actual outcomes — that is, when prediction errors arise (i.e., the surprising omission [negative prediction error] or surprising occurrence [positive prediction error] of an expected outcome). When experienced repeatedly, these prediction-related discrepancies are used to update the brain's beliefs about the world to produce better forecasts and minimize surprise (Bar, 2007; Clark, 2013). Critically, functioning in this way, prediction errors comprise a cornerstone of reinforcement learning (RL) whereby, underpinned by

phasic activity of midbrain dopaminergic neurons (Garrison et al., 2013), stimulus-outcome associations are acquired and revised through cumulative experiences (Gershman, 2015; Pearce & Hall, 1980; Schultz & Dickinson, 2000).

Remarkably, despite the widespread applicability of this computational account of RL (Gershman & Daw, 2017; O'Doherty et al., 2017), whether similar principles apply during person construal remains unknown. This is surprising given the likely products of error-based learning and their societal significance. First, given the potency of unexpected outcomes (Clark, 2013; Johnston & Hawley, 1994; Otten et al., 2017), stereotype-inconsistent (vs. stereotype-consistent) individuals (e.g., female engineers, male nursery teachers) should be advantaged during RL. That is, learning should be accelerated by prediction errors. If observed, such an effect has potential implications for interventions designed to reduce discriminatory practices outside the laboratory, as acquiring knowledge pertaining to individuals in counter-stereotypic roles and learning that stereotypic assumptions are inaccurate are primary pathways through which stereotyping is believed to be diminished (Diekman & Eagly, 2000; Hinton, 2017; Olsson & Martiny, 2018; Wood & Eagly, 2012). Given its operational characteristics, error-based learning may therefore serve as a key process through which this information is procured. Second, consideration of how stereotype-related beliefs influence RL would further underscore the value of computational approaches in elucidating the dynamics of core facets of social cognition (Allidina & Cunningham, 2021; Amodio, 2019; Golubickis & Macrae, 2022; Hackel & Amodio, 2018; Hackel et al., 2015, 2020; Lindström et al., 2015; Lockwood & Klein-Flügge, 2020).

1.1. The Current Research

To explore how prior stereotype-based beliefs influence RL, here we employed a probabilistic selection task (PST; Frank et al., 2004, 2007) in conjunction with computational modeling. In a modified PST, three different stimulus pairs comprising a female and male face (AB, CD, EF) were presented, and over the course of multiple trials participants were required to learn which was the

correct face in each pairing. Crucially, prior to commencing the task, participants were informed that one of the individuals in each pairing was more likely to enjoy ballet (or boxing) as their favorite pastime. In other words, the correct target comprised either a stereotype or a counter-stereotype (Falbén et al., 2019; Quadflieg et al., 2011; Wood & Eagly, 2012). Feedback was given after each trial to indicate whether the selected face was correct or incorrect, but this information was probabilistic (Frank et al., 2004, 2007). In AB trials, choosing A led to correct (i.e., positive) feedback in 80% of the trials, whereas B was accompanied by incorrect (i.e., negative) feedback in these trials (feedback was reversed for the remaining 20% of AB trials). Learning was more difficult for the other stimulus pairs, such that C was the correct response in 70% of CD trials, and E was the correct response in only 60% of EF trials. In this PST, learning could be accomplished via either positive (e.g., A is correct), negative (e.g., B is incorrect), or both types of feedback.

When, for example, participants were tasked with establishing which person in each pairing was more likely to enjoy ballet as their favorite pastime, they could select either a female or male face. As such, learning could arise in one of two ways. If, based on prior stereotype-related beliefs, the female face was selected (i.e., expect a stereotype, likely to be correct) but this turned out to be incorrect, this would constitute a negative prediction error (i.e., the outcome was worse than expected). In contrast, if participants went against established stereotype-related knowledge and chose the male face (i.e., expect a counterstereotype, unlikely to be correct), and this was in fact the correct response, this would represent a positive prediction error (i.e., the outcome was better than expected). Thus, in the current PST, learning could be enhanced when stereotypic selections were incorrect, counter-stereotypic choices were correct, or both outcomes were experienced.

To identify the mechanisms underpinning learning, computational modeling was undertaken on the data. Specifically, based on recent developments, a Reinforcement Learning Drift Diffusion Model (RL-DDM) analysis was adopted (Fontanesi et al., 2019; Pedersen & Frank, 2020; Pedersen et al., 2017). Integrating sequential sampling and RL models (Miletić et al., 2020; Pedersen & Frank, 2020), the RL-DDM pinpoints the latent psychological operations that underpin decision-making (i.e.,

choice selection) and how these are adjusted as learning progresses. This is achieved through the simultaneous hierarchical Bayesian modeling of response time and choice data. A scaling parameter (i.e., drift rate, v) measures sensitivity to feedback by taking both the expected outcome and speed of evidence accumulation into account, such that higher values indicate confident learning, whereas lower values imply uncertainty regarding the expected outcome or low motivation to learn. As such, differences in drift rate scaling reflect variability in the integration of the expected outcomes that contribute to the speed of evidence accumulation toward the chosen option. Additionally, drift rate scaling is equivalent to the inverse temperature parameter in classic instrumental learning models (Pedersen et al., 2017). A learning rate parameter (η) — ranging from zero to one — quantifies how quickly individuals learn, with larger values indicating the utilization of current feedback (i.e., fast learning), and smaller values reflecting reduced updating from recently experienced outcomes (i.e., slow learning). In the current version of the PST, the learning rate captures how efficiently participants learn which face in each of the pairings is most likely to be correct. In this regard, either a single learning rate (η) that captures all learning, or separate learning rates for negative and positive prediction errors (η^- & η^+ respectively) can be estimated (Miletić et al., 2020; Pedersen & Frank, 2020; Pedersen et al., 2017). The model also establishes how much evidence is needed to make a decision (i.e., threshold separation, a), with larger (vs. smaller) values indicating greater response caution. Finally, the non-decision time (t_0; i.e., components that are not part of the decision-making process, such as stimulus encoding and response execution) is also estimated in the RL-DDM.

Based on prior work, several effects were expected to emerge. First, reflecting the influence of pre-existing stereotype-related beliefs (i.e., women are more likely to enjoy ballet, men are more likely to enjoy boxing), we anticipated that participants would display an expectancy-based bias in favor of stereotype-consistent (vs. stereotype-inconsistent) responses (Diekman & Eagly, 2000; Olsson & Martiny, 2018; Wood & Eagly, 2012). During decision-making, evidence is gradually accumulated for the response options based on the subjective value (i.e., expected reward) associated with each outcome (i.e., Q-values). Once a critical evidential threshold has been reached, a response is

selected (Miletić et al., 2020; Pedersen & Frank, 2020). As stereotypes comprise beliefs about the likely traits, characteristics, and behavioral proclivities of group members (Macrae & Bodenhausen, 2000), stereotype-consistent responses should initially be associated with greater subjective reward (i.e., larger *Q*-values) than stereotype-inconsistent outcomes. Second, given the potency of unexpected outcomes (i.e., expectancy violation – stereotypic choice selections are incorrect/counter-stereotypic choice selections are correct), we expected learning rates to be faster for counter-stereotypes compared to stereotypes, both for positive (η^+) and negative (η^-) prediction errors (Bar, 2007; Clark, 2013). Third, in terms of the operations that underpin decisional processing, counter-stereotypic (vs. stereotypic) choice selections were anticipated to be accompanied by greater response caution (i.e., larger threshold separation, *a*). That is, the stability of pre-existing stereotype-based beliefs should increase the evidential requirements of stereotype-inconsistent (vs. stereotype-consistent) choice selections (Freeman & Ambady, 2011; Hilton & von Hippel, 1996; Macrae & Bodenhausen, 2000). Finally, in line with previous research, it was expected that non-decision times would be faster for stereotype-consistent (vs. stereotype-inconsistent) responses (Frenken et al., 2022; Persson et al., 2022).

2. Method

2.1. Participants and Design

Sixty participants (47 females, 12 males, 1 other; M_{age} = 24.30, *SD* = 3.51), with normal or corrected-to-normal visual acuity, took part in the research. Six participants (5 females) failed to learn the probabilities associated with each face during the learning task, thus were excluded from the analyses. Data collection was conducted online using Prolific Academic (www.prolific.co), with each participant receiving compensation at the rate of £7.50 (~$10) per hour. Informed consent was obtained from participants prior to the commencement of the experiment and the protocol was reviewed and approved by the Ethics Committee at the School of Psychology, University of Aberdeen. The experiment had a 2 (Pastime: ballet or boxing) X 2 (Correct Target: stereotype or

counter-stereotype) mixed design with repeated measures on the second factor. To detect a significant two-way interaction, a sample of sixty participants afforded 90% power for a medium to large effect size (i.e., d = .65; PANGEA, v .0.2).

2.2. Stimulus Materials and Procedure

Participants performed two blocks of a PST (Frank et al., 2004, 2007), with each comprising a learning phase in which three pairs of faces (denoted as AB, CD, and EF, see Figure 1) were presented. Each pairing consisted of a female and male face, and prior to the commencement of the task additional stereotype-related information was provided (Falbén et al., 2019; Quadflieg et al., 2011; Wood & Eagly, 2012). Specifically, participants were informed that one of the individuals in each paring was more likely to enjoy ballet (or boxing) as their favorite pastime (i.e., feminine [or masculine] stereotype). Participants were instructed that they were required to learn which face in each pair was most likely to be correct (i.e., reinforced) based on feedback (i.e., correct vs. incorrect) provided after each selection. Critically, the sex of the faces associated with positive reinforcement was manipulated. Whereas in one of the blocks, female targets were more likely to be correct, for the other block, male targets were more likely to comprise the correct response. Thus, across the various permutations of the PST, stereotypic or counter-stereotypic individuals were more likely to be correct. Participants were randomly assigned to either the ballet or boxing condition in which they did either a stereotype-confirming (i.e., female [male] faces more likely to be correct in the ballet [boxing] condition) or stereotype-disconfirming (i.e., male [female] faces more likely to be correct in the ballet [boxing]) block of trials. The order of the stereotype-confirming and stereotype-disconfirming blocks was counterbalanced across the sample.

The probabilities indicating which face was more likely to be correct followed the standard version of the PST (Frank et al., 2004, 2007). Specifically, for the AB pair, A was 80% likely to be correct (20% for B), for the CD pair, C was 70% likely to be correct (30% for D), and finally, for the EF pair, E was 60% likely to be correct (40% for F, see Figure 1). Over numerous choice selections,

participants learned which item in each pairing was more likely to be correct (i.e., A, C, E rather than B, D, F) based on the feedback provided. The task was completed when participants reached sufficient levels of accuracy for each pairing (i.e., AB, 60% or above; CD, 55% or above; EF, 50% or above; Frank et al., 2004, 2007).

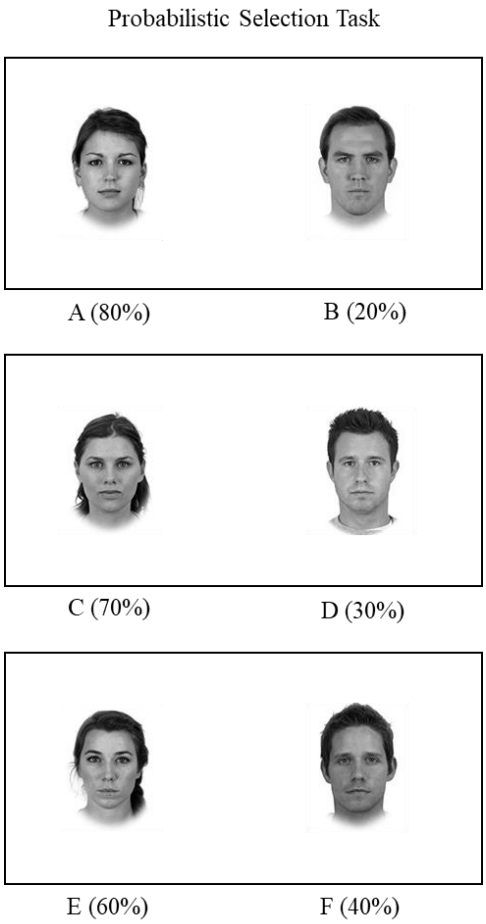

Figure 1. Example of the stimulus pairs and the probabilities of correct responses during the probabilistic selection task.

Each trial began with the presentation of a face pair that remained on the screen until the participant made a response. After the participant selected one of the faces, both textual (i.e., the word

'Correct' in green or 'Incorrect' in red) and auditory (i.e., a high-pitched beep for a correct response or a low-pitched beep for an incorrect response) feedback were presented for 1000 ms, followed by a blank screen for 500 ms, after which the next trial commenced. Participants had to select a face by pressing the appropriate button on the keyboard (i.e., 'A' for the face on the left side of the screen, 'L' for the face on the right side of the screen). The faces in each pair were equally likely to be presented on the left or right side of the screen. The faces (6 female & 6 male) were taken from the Chicago Face Database (Ma et al., 2015), were 140 x 176 in size, grayscale, and depicted young adults aged 20-30 years. These faces were divided into two sets which were counterbalanced, such that they were presented in either the stereotype-confirming or stereotype-disconfirming block. The experiment was conducted using Inquisit Web. Participants completed blocks of 60 trials in which each of the three face pairs appeared randomly, equally often, until accuracy reached a satisfactory level. The maximum number of learning blocks was set to six (i.e., 360 trials in total) if the participant did not reach satisfactory levels of accuracy earlier in the task (Frank et al., 2007). If the participant's accuracy was not sufficient after six blocks of trials, they were excluded from the analyses. On completion of the experiment, participants were debriefed and thanked.

3. Results

3.1. Behavioral Data

Decision Time. Responses faster than 200 ms and slower than 4 seconds were excluded from the analysis (Frank et al., 2007), eliminating approximately 4% of the trials. To explore how stereotype-based beliefs influenced decision times and learning performance, a 2 (Pastime: ballet vs. boxing) X 2 (Correct Target: stereotype vs. counter-stereotype) X 3 (Stimulus Pair: AB vs. CD vs. EF) multilevel model was used (see Table 1 for the treatment means). Analyses were conducted using the R package 'lme4' (Pinheiro et al., 2015). Pastime, Correct Target, and Stimulus Pair were treated as categorical fixed effects, with participants as a crossed random effect (Judd et al., 2012).

Analysis of the decision times yielded a main effect of Correct Target (b = -.042, SE = .005, t = -8.14, p < .001), such that responses were faster to stereotypes (M = 961 ms, SD = 332 ms) than counter-stereotypes (M = 1,033 ms, SD = 367 ms). Additionally, a significant main effect of Stimulus Pair (b = -.019, SE = .005, t = -3.30, p = .001) indicated that responses were faster to faces with higher probabilities of being correct or incorrect (M_{AB} = 973 ms, SD_{AB} = 337 ms, M_{CD} = 1,002 ms, SD_{CD} = 345 ms, M_{EF} = 1,017 ms, SD_{EF} = 365 ms). Finally, a significant Pastime X Correct Target interaction was also observed (b = -.017, SE =.005, t = 3.35, p = .001). Further inspection of the interaction revealed that, in the ballet condition, responses were faster to stereotypes than counter-stereotypes (b = -.025, SE = .007, t = -3.35, p = .001), an effect that was also observed in the boxing condition (b = -.059, SE = .007, t = -8.22, p < .001).

Learning Performance. The multilevel analysis revealed a significant main effect of Stimulus Pair (b = .161, SE = .024, z = 6.70, p < .001), such that learning was more accurate for faces with higher probabilities of being correct or incorrect (M_{AB} = 77%, SD_{AB} = 13%, M_{CD} = 72%, SD_{CD} = 13%, M_{EF} = 68%, SD_{EF} = 13%). A significant Pastime X Correct Target interaction was also observed (b = .044, SE = .021, z = 2.05, p = .040). Further inspection of the interaction yielded no significant differences between stereotypes and counter-stereotypes in either of the pastimes.[1]

[1] Although no differences emerged in learning performance across the experimental conditions, the fact that decision times varied significantly indicates that learning occurred during the PST. It is likely that, over time, learning was accompanied by reduced response caution, resulting in comparable levels of accuracy across the task (see Miletic et al., 2020). Crucially, as the RL-DDM considers both decision time and accuracy when estimating parameters, this highlights the benefits of this analytical approach.

Table 1. Decision time (ms) and learning performance (%) as a function of Pastime, Correct Target, and Stimulus Pair.

Pastime	Correct Target	Decision Time	Learning Performance
AB pair			
ballet	stereotype	975 (272)	77 (13)
	counter-stereotype	1,054 (407)	78 (12)
boxing	stereotype	892 (351)	75 (14)
	counter-stereotype	969 (319)	77 (13)
CD pair			
ballet	stereotype	1,010 (295)	73 (13)
	counter-stereotype	1,051 (368)	72 (13)
boxing	stereotype	928 (365)	69 (14)
	counter-stereotype	1,021 (352)	74 (13)
EF pair			
ballet	stereotype	1,012 (308)	72 (12)
	counter-stereotype	1,070 (403)	67 (13)
boxing	stereotype	949 (398)	67 (15)
	counter-stereotype	1,036 (352)	67 (12)

Note. Standard deviations (*SD*) appear within parentheses.

3.2. Reinforcement Learning Drift Diffusion Model Analysis

To identify the processes underpinning learning, data were submitted to a RL-DDM analysis (Pedersen & Frank, 2020; Pedersen et al., 2017). This analysis combines the strengths of RL and sequential-sampling models (SSMs) to elucidate the operations that support task performance. Specifically, although RL models account for changes in the relative proportion of choice probabilities over the course of learning, they do not speak to concurrent differences in response latencies, a fundamental and important dimension of the available data (e.g., as learning takes place, decision times decrease). In this respect, SSMs (e.g., drift diffusion model; Ratcliff & Smith, 2004;

Ratcliff et al., 2016) are useful as they provide a mechanistic account of binary decision-making by explaining how choice accuracy and response latencies collectively arise from a common set of latent cognitive processes (e.g., rate of evidence accumulation, response caution). In essence, these models assert that evidence is gathered for each choice option (e.g., face A vs. face B) until a critical evidential threshold is reached, at which point a response is made. Thus, crucially, the RL-DDM extends standard RL models by explicating the processes through which learning unfolds over time (Fontanesi et al., 2019; Miletić et al., 2020; Pedersen & Frank, 2020; Pedersen et al., 2017).

To estimate model parameters, an extension of the Bayesian hierarchical drift diffusion toolbox was adopted (Pedersen & Frank, 2020; Wiecki et al., 2013). Models were response-coded, such that the upper threshold corresponded to responses to stimuli that were positively reinforced (i.e., faces corresponding to the letters A, C, & E) and the lower threshold to stimuli that were negatively reinforced (i.e., faces corresponding to the letters B, D, & F; Pedersen & Frank, 2020). Specifically, in the stereotype-confirming conditions, the upper response boundary represented the letters corresponding to stereotype-consistent individuals (i.e., females in the ballet condition, males in the boxing condition) and the lower boundary to stereotype-inconsistent persons (i.e., females in the boxing condition, males in the ballet condition). Conversely, in the stereotype-disconfirming conditions, the upper response boundary corresponded to counter-stereotypic individuals and the lower boundary to stereotypic persons. Bayesian posterior distributions were modeled using a Markov chain Monte Carlo (MCMC) with 10,000 samples (including 5,000 burn). Model comparison was performed using the Deviance Information Criterion (DIC) as this approach is routinely adopted when comparing hierarchical Bayesian models (Spiegelhalter et al., 2002). Lower DIC values favor models with the highest likelihood and least number of parameters.

3.2.1. RL-DDM Model Comparison

To identify the processes underpinning task performance, six RL-DDM models were selected for comparison. To establish whether participants held pre-existing beliefs about which gender was

more likely to enjoy ballet (or boxing) as a favorite pastime (Diekman & Eagly, 2000; Olsson & Martiny, 2018; Wood & Eagly, 2012), the initial values of expected outcomes (Q) of the delta learning rule in RL (i.e., $Q_{chosen\text{-}option\ (t)} = Q_{\text{chosen-}option\ (t\text{-}1)} + \eta(\text{feedback}_{(t\text{-}1)} - Q_{chosen\text{-}option\ (t\text{-}1)})$, where η = the speed of learning) were varied across the experimental conditions. Of interest was whether participants displayed: (i) an expectancy-based bias to assume that stereotype-consistent responses would be correct (i.e., the initial Q-value shifted toward the stereotype-consistent response threshold); (ii) no bias toward either stereotype-consistent or stereotype-inconsistent responses (i.e., initial Q-value centered between the response thresholds), or (iii) an expectancy-based bias to assume that stereotype-inconsistent responses would be correct (i.e., initial Q-value shifted toward the stereotype-inconsistent response threshold).

As previous modeling research has demonstrated that expectancy-based biases during stereotype-related decision-making are underpinned by differences in the relative starting point of decisional processing (Falbén et al., 2019; Persson et al., 2021, 2022; Tsamadi et al., 2020), the stereotype-bias model (Model 1) was defined as a shift of the initial expected reward (Q) toward the stereotype-consistent response boundary. This resulted in setting the initial expected outcome value to 0.6 for the stereotypic conditions (as the upper boundary corresponded to stereotype-consistent responses) and 0.4 for the counter-stereotypic conditions (as the lower boundary corresponded to counter-stereotypic responses). In contrast, in the counter-stereotypic model (i.e., Model 3), these values were reversed, such that the initial expected reward value (Q) was 0.4 in the stereotypic conditions and 0.6 in the counter-stereotypic conditions. Finally, in the no bias model (Model 2), the initial expected reward value (Q) was fixed to 0.5 in both the stereotypic and counter-stereotypic conditions.

To establish whether learning was driven by a single or dual learning rate, for each combination of the initial expected reward (i.e., Q-values, single [η] and dual-learning rate [η^- & η^+]), models were compared in which the learning rates varied across Pastime (i.e., ballet vs. boxing) and Correct Target (i.e., stereotype vs. counter-stereotype). In these models, drift rate scaling (v) varied by

Pastime and Correct Target, while threshold separation (a) and non-decision time (t_0) varied only by Correct Target. This parametrization was selected as previous research has shown that stereotypic (vs. counter-stereotypic) decisions require less evidence (e.g., Falbén et al., 2019; Persson et al., 2021, 2022; Tsamadi et al., 2020) and are associated with faster non-decision times (e.g., Frenken et al., 2022; Persson et al., 2022). As can be seen from Table 2, the dual learning rate version of Model 1 provided the best fit (i.e., the lowest DIC value; Spiegelhalter et al., 2002). This model also converged well across three chains of 10,000 samples and 5,000 burn ($\hat{R}$ = 1.005; Gelman & Rubin 1992; Pedersen et al., 2017).

To further evaluate the best fitting model, a Posterior Predictive Check (PPC) was performed (Pedersen & Frank, 2020; Wiecki et al., 2013). From the best fitting model, posterior distributions of the estimated parameters were used to simulate data. The quality of model fit was then assessed by plotting the observed data against the simulated data for the choice proportions and decision times for each stimulus pair (i.e., AB, CD, & EF; Pedersen & Frank, 2020; Pedersen et al., 2017). Visual inspection of the PPC indicated a good model fit (see Supplementary Material for plots). In addition, the RL-DDM parameter recovery analysis was also undertaken. Specifically, the estimated parameters from the best-fitting model were used to simulate data which were then refitted by the RL-DDM. This analysis indicated successful model parameter recovery (see Supplementary Material for further details).

Table 2. Model comparison of the initial expected outcome values (*Q*) and learning rates.

Model	Stereotype (Ballet & Boxing)	Counter-Stereotype (Ballet & Boxing)	Single Learning Rate Model DIC	Dual Learning Rate Model DIC
1.	0.6	0.4	32785	32765
2.	0.5	0.5	32788	32771
3.	0.4	0.6	32792	32777

Note. Model 1 = stereotype bias; Model 2 = no bias; Model 3 = counter-stereotype bias. DIC = Deviance Information Criterion.

3.2.2. RL-DDM Parameter Posterior Distributions

Examination of the posterior distributions revealed differences in learning rates for negative and positive prediction errors (η^- & η^+), threshold separation (a), drift rate scaling (v), and non-decisional processes (t_0). See Figures 2 and 3 and Supplementary Material for the parameter estimates.

***Negative Prediction Errors* (η^-).** In the ballet condition, comparison of the posterior distributions yielded evidence that learning was faster for counter-stereotypes compared to stereotypes (p_{Bayes}[counter-stereotypes > stereotypes] < .001, BF_{10} > 1000).[2] An identical effect was observed in the boxing condition (p_{Bayes}[counter-stereotypes > stereotypes] = .009, BF_{10} = 113). Thus, across both pastimes, the learning rate was faster for counter-stereotypes than stereotypes (p_{Bayes}[counter-stereotypes > stereotypes] < .001, BF_{10} > 1000).

***Positive Prediction Errors* (η^+).** In the ballet condition, evidence was observed for faster learning for counter-stereotypes compared to stereotypes (p_{Bayes}[counter-stereotypes > stereotypes] = .078, BF_{10} = 12), an effect that also emerged in the boxing condition (p_{Bayes}[counter-stereotypes >

[2] Bayesian *p* values quantify the degree to which the difference in the posterior distribution is consistent with the hypothesis. For example, a Bayesian *p* of .05 indicates that 95% of the posterior distribution supports the hypothesis that the parameter posteriors differ across the conditions (Marsman & Wagenmakers, 2017).

stereotypes] = .035, BF_{10} = 28). For both pastimes, therefore, the learning rate was faster for counter-stereotypes than stereotypes (p_{Bayes}[counter-stereotypes > stereotypes] =.016, BF_{10} = 62).

***Threshold Separation* (*a*).** Comparison of the posterior distributions indicated that threshold separation was wider for counter-stereotypes compared to stereotypes (p_{Bayes}[stereotypes < counter-stereotypes] = .013, BF_{10} = 76), indicating that response caution was greater for the former decisions.

***Non-decision Time* (t_0).** Comparison of the posterior distributions indicated that the non-decision time was faster for stereotypes than counter-stereotypes (p_{Bayes}[stereotypes < counter-stereotypes] < .001, BF_{10} > 1000).

***Drift Rate Scaling* (*v*).** Comparison of the posterior distributions indicated that drift rate scaling was larger for stereotypes in the ballet condition (p_{Bayes}[stereotypes > counter-stereotypes] < .001, BF_{10} > 1000), but for counter-stereotypes in the boxing condition (p_{Bayes}[counter-stereotypes > stereotypes] = .200, BF_{10} = 4).

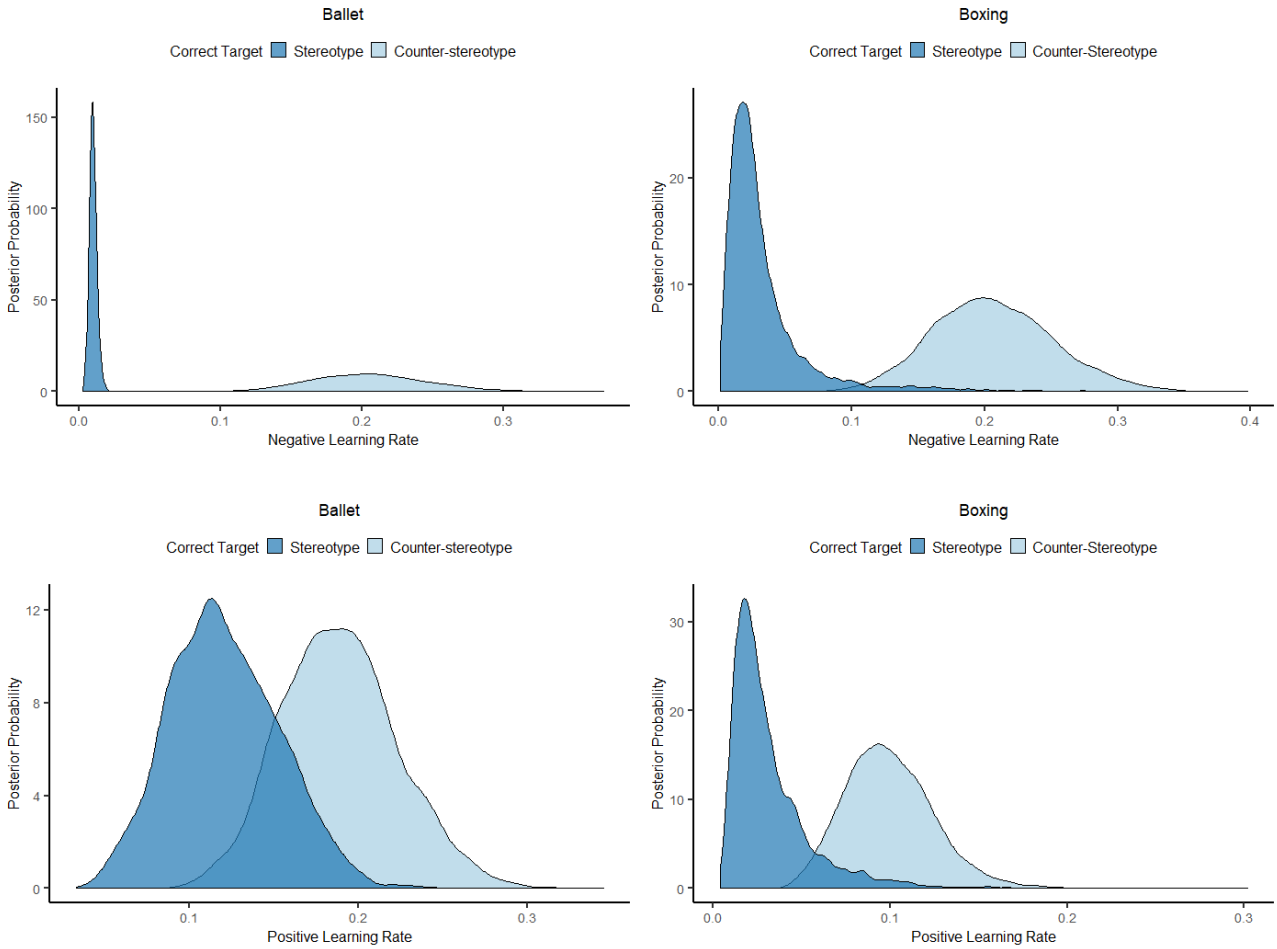

Figure 2. Posterior probabilities of the Reinforcement Learning Drift Diffusion Model Parameters (Learning Rates).

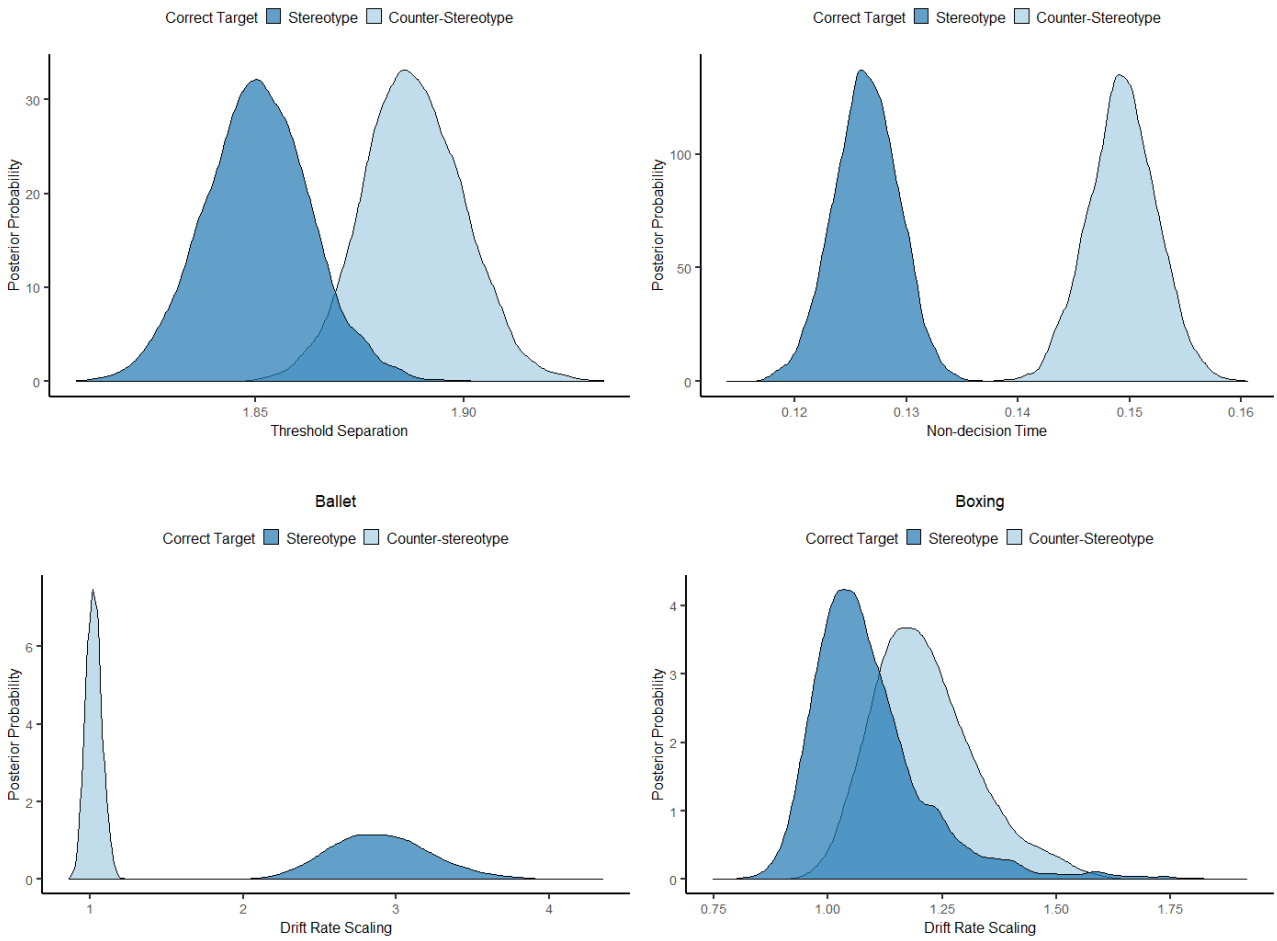

Figure 3. Posterior probabilities of the Reinforcement Learning Drift Diffusion Model Parameters (Threshold Separation, Non-decision Time and Drift Rate Scaling).

4. Discussion

Using a PST (Frank et al., 2004, 2007), the current inquiry explored how stereotype-based beliefs influence the acquisition of person-related knowledge. Several noteworthy effects emerged. First, reflecting the impact that cultural socialization exerts on person construal, participants were initially biased to assume that stereotype-consistent (vs. stereotype-inconsistent) responses would be positively reinforced. Specifically, pre-existing expectations impacted the starting values of the expected rewards (Diekman & Eagly, 2000; Olsson & Martiny, 2018; Wood & Eagly, 2012). Similarly, threshold setting was less cautious for stereotype-consistent compared to stereotype-

inconsistent responses. As such, the current findings corroborate previous research which has indicated that stereotypes decrease the evidential requirements of response selection by biasing participants toward stereotype-consistent outcomes (e.g., Falbén et al., 2019; Persson et al., 2021, 2022; Tsamadi et al., 2020). Second, choice selections during the PST were faster for stereotypes than counter-stereotypes. Third, notwithstanding the economizing effects that stereotypes exerted on decisional processing (Fiske & Neuberg, 1990; Freeman & Ambady, 2011; Macrae & Bodenhausen, 2000), as revealed by the RL-DDM analysis, counter-stereotypes were learned more rapidly than stereotypes, both for negative and positive prediction errors. Thus, extending extant work, unexpected person-related outcomes facilitated learning (Gershman & Daw, 2017; O'Doherty et al., 2017). Finally, replicating previous research, non-decisional processes were faster for stereotype-consistent compared to stereotype-inconsistent responses (Frenken et al., 2022; Persson et al., 2022).

Underpinned by the benefits of a mind that is both stable and flexible, the current findings can be understood in terms of predictive accounts of social perception (Bach & Schenke, 2017; Clark, 2013; Hinton, 2017; O'Callaghan et al., 2017; Otten et al., 2017). Whereas expectancy-consistent outcomes reveal nothing about the world that was not already presumed, expectancy-inconsistent outcomes, in contrast, challenge conventional wisdom, thus attract additional scrutiny to resolve the apparent prediction error (Johnston & Hawley, 1994; Otten et al., 2017; Sherman et al., 1998, 2000). This enhanced processing entails an assessment of whether existing forecasts are imprecise and updating (i.e., learning) is required. When encountered repeatedly, the brain eventually rejects the possibility that errors were caused by random noise and predictions are adjusted accordingly (Clark, 2013; Hinton, 2017; O'Callaghan et al., 2017; Otten et al., 2017). The results reported here resonate with this viewpoint. For both negative and positive prediction errors, counter-stereotypes were learned more rapidly than stereotypes. Specifically, learning was enhanced when stereotypic choice selections were unexpectedly incorrect (i.e., negative prediction errors) and counter-stereotypic choice selections were surprisingly correct (i.e., positive prediction errors), although stronger evidence was observed for the former effect (i.e., learning was speeded when responses based on prior beliefs were

disconfirmed). This pattern of results is interesting as it highlights that learning outcomes were more potent not when counter-stereotypic responses were correct (e.g., a woman enjoys boxing), but rather when stereotypic responses were incorrect (e.g., a man does not enjoy boxing). As such, stereotype negation may serve as productive tactic for modifying group-related beliefs (Kawakami et al., 2000).

That unexpected (vs. expected) information speeded learning corroborates and extends prior work on person construal. For example, at least when tasked with forming impressions of others, a memorial bias for unexpected information is commonly observed (e.g., Hastie & Kumar, 1979; Macrae et al., 1993; Stangor & Duan, 1991). Underpinning the emergence of this recollective preference is an effortful (i.e., resource consuming) cognitive process termed inconsistency resolution (Srull & Wyer, 1989). Specifically, when unexpected material is encountered, elaborative processing is initiated in an attempt to reconcile the discrepant information with pre-existing group-related beliefs (Crocker et al., 1983; Macrae et al., 1999). As such, surprising material is advantaged in memory. In a similar way, here we demonstrated that, when prediction errors challenged stereotype-related beliefs during a PST, learning was enhanced. In particular, when the knowledge yield was greatest, learning was accelerated. Underscoring the potency of unexpected stimulus inputs during person perception (Sherman et al., 1998, 2000), this suggests that learning is facilitated when there is the most to be learned.

Favoring the acquisition of expectancy-discrepant knowledge, RL has potentially important implications for the reduction of stereotype-based responding and creation of equitable societal opportunities for women and men. Through exposure to gender-congruent individuals in media portrayals and daily life (e.g., women in unpaid roles, men in salaried positions), children rapidly become cognizant of the conduct expected of them, expectations that guide their behavior in a restrictive stereotype-confirmatory manner. As such, to broaden their horizons and prospects, numerous initiatives and interventions have focused on encouraging individuals to contemplate non-stereotyped educational/occupational choices (Fitzgerald et al., 2019; Olsson & Martiny, 2018). Crucially, the negation of stereotypic beliefs and the observation (and learning) of counter-stereotypes

is critical in this regard (Pettigrew & Tropp, 2006; Wood & Eagly, 2012). Social Role Theory (SRT; Wood & Eagly, 2012), for example, asserts that following the perception of non-traditional divisions of labor, women and men are associated with counter-stereotypic characteristics and aspirations. As a result, stereotype-incongruent individuals have the capacity to modify both group-related beliefs and future behavioral choices (e.g., academic major, preferred hobby) in desirable ways. If, as revealed by the current findings, instrumental learning operates in such a way as to enhance the acquisition of counter-stereotypic compared to stereotypic knowledge, this provides an important pathway through which prior gender beliefs can be updated following errant person-related predictions.

Of course, updating stereotypes in everyday life does not share the characteristics of probabilistic selection tasks in which people must actively choose which of two competing alternatives is most likely to be correct (i.e., rewarding) over multiple trials. Rather, outside the laboratory, targets are simply encountered who either confirm or disconfirm prevailing stereotype-related beliefs. The potential importance of the current findings, however, lies in the demonstration that counter-stereotypes were learned more rapidly than stereotypes, a finding that speaks to the strategies and tactics that could be used in attempts to attenuate stereotyping through training and education. For example, a commonplace intervention has been to expose children to counter-stereotypic role models (either through media portrayals or live interaction) in the hope this will challenge their stereotype-related beliefs (see Olson & Martiny, 2018). In this regard, an intriguing extension of the current work would be to devise games or puzzles in which children must judge, for example, which of two targets is most likely to work in a particular occupation, possess a specific personality characteristic, or enjoy a certain pastime. In other words, translate the features of the PST (i.e., error-based learning) into an engaging classroom activity. The benefits of such an approach would be considerable, but most notably it would make salient (in a cost-effective manner) the existence of prediction errors across multiple stereotype-related dimensions, a prerequisite of meaningful stereotype change (Fitzgerald et al., 2019).

Notwithstanding accelerated learning rates for counter-stereotypes compared to stereotypes, it should be noted that group-based beliefs remain stubbornly resistant to modification in the face of disconfirmation (Maurer et al., 1995; Richards & Hewstone, 2001; Weber & Crocker, 1983). The limiting factor is a process termed subtyping. Subtyping occurs when atypical exemplars (i.e., stereotype-disconfirming group members) are clustered together in memory to form a distinct subgroup (Hewstone & Hamberger, 2000; Kunda & Oleson, 1995). For example, one may generate subtypes that represent female plumbers or male secretaries. By considering such individuals as exceptions to the rule, they are segregated from the group as a whole with the result that pre-existing stereotype-based beliefs can be preserved. Of course, subtyping becomes progressively difficult if stereotype-discrepant persons become increasingly numerous and/or stereotypic assumptions about group members turn out to be unfounded. Under such conditions, error-based learning may play a contributory role in the reduction of stereotypical thinking.

The updating of stereotype-based knowledge via prediction errors raises several interesting issues. Most notably, as belief modification is sensitive to the strength of prediction errors, paradoxically, the effect of unexpected outcomes should be greatest for potent stereotypes that are held with the utmost conviction. That is, prediction errors should exert most influence when — and for whom — they are least expected. Interestingly, presaging current predictive processing frameworks (Bach & Schenke, 2017; O'Callaghan et al., 2017; Otten et al., 2017), the Encoding Flexibility Model (EFM; Sherman et al., 1998) of stereotyping advanced a similar observation. According to this account, because expectancy-consistent material confirms prior beliefs, it has little informational value, thus attracts less attention than expectancy-inconsistent information during stimulus encoding, an effect that is amplified in demanding task contexts and when expectations are strongly (vs. weakly) endorsed. In essence, to enhance efficiency and maximize cognitive flexibility, processing favors stimuli that generate the largest knowledge gain (Hinton, 2017; Johnston & Hawley, 1994; McClelland et al., 1995).

Collectively, these theoretical viewpoints suggest that prediction errors are most informative when they are triggered by established compared to emerging stereotypes and among individuals who hold strong rather than weak stereotype-related beliefs. Thus, counter-intuitively, entrenched beliefs may be the easiest to modify, unless of course additional motivational factors (e.g., system justification, identity-maintenance) dilute the significance, hence impact, of prediction errors (Jost & Hunyady, 2002; Tajfel, 1982). For example, as noted previously, people may subtype atypical (i.e., stereotype-inconsistent) group members to preserve the superordinate stereotype (Richards & Hewstone, 2011). Future research should explore this important theoretical and practical matter for the acquisition of person-related knowledge across a diverse range of stereotypes (e.g., race, age, occupational, class) and individuals (e.g., persons high or low in sexism). One intriguing possibility is that stereotypes may exert distinct effects on components of decisional processing. Here, for example, beliefs about gender-related pastimes influenced the drift rate scaling parameter in opposing ways. Whereas, for ballet, learning was more confident for stereotypes compared to counter-stereotypes, this effect was reversed for boxing (i.e., counter-stereotypes > stereotypes). This difference may reflect variability in the strength with which women and men are associated with these activities, hence the potency of prediction errors (Wood & Eagly, 2012). In addition, extending the current investigation, computational approaches should be adopted that explore stereotype-based learning in more complex task settings. A basic limitation of the RL-DDM model is that it can only address binary decision-making, thereby potentially underestimating the nuanced ways in which stereotypes bias learning. Overcoming this restriction, application of complementary analytical methods — for example the Reinforcement Learning Advantage Racing Diffusion (RL-ARD) Model — would elucidate the dynamics of stereotype-based learning in task contexts in which multiple decisional outcomes are possible (Miletić et al., 2021).

In combination with different analytical techniques, modification of the current PST may yield valuable insights into the vagaries of stereotype-based learning. For example, given the pivotal role that stereotypes play in facilitating processing efficiency (Macrae & Bodenhausen, 2000; Sherman et

al., 2000), what would happen if to-be-learned material was encountered in a demanding task setting (e.g., cognitive load, time constraints; see Rae et al., 2014; Sewell & Stallman, 2000)? In particular, would the knowledge yield be greater following negative or positive prediction errors? Given the demonstration that expectancy-inconsistent (vs. expectancy-consistent) material is prioritized when attentional resources are scarce (Sherman et al., 1998), it is conceivable that positive (i.e., counter-stereotypic choice selections) rather than negative (stereotypic choice selections) prediction errors may exert greater influence on learning under these conditions. Additionally, as it was potentially possible for learning to transfer across the stimulus pairs in the current PST (i.e., positively [negatively] reinforced outcomes comprised same-sex faces), it would be interesting to explore RL in a task context in which only same-sex faces were utilized and stereotype fit (i.e., stereotype vs. counter-stereotype) was manipulated in a different way.

Consideration should also be given to the neural activity that underpins stereotype-driven learning. A rapidly emerging literature is successfully delineating how computational modeling can be used to identify the specific processes that underpin core aspects of social-cognitive functioning (e.g., impression formation, mentalizing, self-referential mentation) and how these are related to neural activity and behavior (Hackel & Amodio, 2018; Lockwood & Klein-Flügge, 2020). Moving beyond abstract stimuli and inconsequential judgments, work of this kind probes the possibility that distinct regions of the brain may be recruited when decision-making is inherently social (vs. non-social) — the so-called social brain hypothesis (Dunbar, 2009). In the context of stereotype-guided instrumental learning, for example, in addition to activity in the ventral striatum (VS) correlating with prediction errors (Garrison et al., 2013), so too other cortical regions (e.g., temporal parietal junction [TPJ], medial prefrontal cortex [mPFC], anterior cingulate cortex [ACC]) may be sensitive to specific aspects of the to-be-learned material (e.g., affective tone, value, personal relevance) and requirements of the prevailing task setting (Apps et al., 2016; Zaki et al., 2016). Research of this kind is theoretically important as it will establish whether the areas of the brain involved in stereotype-based RL are uniquely social or instead reflect the operation of domain general processes.

5. Conclusion

The way in which we think about, and interact with, other people is profoundly influenced by stereotypic knowledge acquired through many years of cultural socialization (Wood & Eagly, 2012). Crucially, however, these group-related preconceptions are not immutable, but rather evolve continuously based on new learning experiences. Using computational modeling techniques, here we showed that RL is enhanced for the very targets that ultimately weaken stereotype-based beliefs — counter-stereotypes. In this way, error-based learning may serve as a potential pathway through which prediction-mismatching experiences modify people's stereotype-related presumptions, thus the character of their social exchanges and life choices.

References

Allidina, S., & Cunningham, W. A. (2021). Avoidance begets avoidance: A computational account of negative stereotype persistence. *Journal of Experimental Psychology: General, 150*, 2078-2099.

Allport, G.W. (1954). *The nature of prejudice*. Reading, MA: Addison- Wesley.

Amodio, D. M. (2019). Social cognition 2.0: An interactive memory systems account. *Trends in Cognitive Sciences, 23*, 21-33.

Apps, M. A. J., Rushworth, M. F. S., & Chang, S. W. C. (2016). The anterior cingulate gyrus and social cognition: Tracking the motivation of others. *Neuron, 90,* 692-707.

Bach, P., & Schenke, K. C. (2017). Predictive social perception: Towards a unifying framework from action observation to person knowledge. *Social and Personality Psychology Compass, 11*, 1-17.

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences, 11*, 280-289.

Beaman, L., Chattopadhyay, R., Duflo, E., Pande, R., & Topalova, P. (2009). Powerful women: Does exposure reduce bias? *The Quarterly Journal of Economics, 124,* 1497-1540.

Bodenhausen, G. V., & Lichtenstein, M. (1987). Social stereotypes and information-processing strategies: The impact of task complexity. *Journal of Personality and Social Psychology, 52*, 871-880.

Brewer, M. B. (1988). A dual process model of impression formation. *Advances in Social Cognition, 1*, 1-36.

Clark. A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36,* 181-253.

Correll, J., Hudson, S. M., Guillermo, S., & Earls, H. A. (2017). Of kith and kin: Perceptual enrichment, expectancy, and reciprocity in face perception. *Personality and Social Psychology Review*, *21*, 336-360.

Correll, J., Wittenbrink, B., Crawford, M. T., & Sadler, M. S. (2015). Stereotypic vision: How stereotypes disambiguate visual stimuli. *Journal of Personality and Social Psychology, 108*, 219-233.

Crocker, J., Hannah, D. B., & Weber, R. (1983). Person memory and causal attribution. *Journal of Personality and Social Psychology*, *44*, 55-66.

Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology, 40,* 642-658.

Dennehy, T. C., & Dasgupta, N. (2017). Female peer mentors early in college increase women's positive academic experiences and retention in engineering. *Proceedings of the National Academy of Sciences, 114,* 5964-5969.

Diekman, A. B., & Eagly, A. H. (2000). Stereotypes and dynamic constructs: Women and men of the past, present, and future. *Personality and Social Psychology Bulletin, 26,* 1171-1188.

Dunbar, R. I. M. (2009). The social brain hypothesis and its implications for social evolution. *Annals of Human Biology, 36,* 562-572.

Eagly, A. H., & Steffen, V. J. (1984). Gender stereotypes stem from the distribution of women and men into social roles. *Journal of Personality and Social Psychology, 46,* 735–754.

Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing black: Race, crime, and visual processing. *Journal of Personality and Social Psychology, 87*, 876-893.

Falbén, J. K., Tsamadi, D., Golubickis, M., Olivier, J. L., Persson, L. M., Cunningham, W. A., & Macrae, C. N. (2019). Predictably confirmatory: The influence of stereotypes during decisional processing. *Quarterly Journal of Experimental Psychology*, 72, 2437-2451.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology, 23*, 1-74.

Fitzgerald, C., Martin, A., Berner, D., & Hurst, S. (2019). Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: A systematic review. *BMC Psychology, 7,* 1-12.

Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin and Review, 26,* 1099-1121.

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchinson, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings in the National Academy of Sciences, 104*, 16311-16316.

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science, 306,* 1940-1943.

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review, 118,* 247-279.

Frenken, M., Hemmerich, W., Izydorczyk, D., Scharf, S., & Imhoff, R. (2022). Cognitive processes behind the shooter bias: Dissecting response bias, motor preparation, and information accumulation. *Journal of Experimental Social Psychology, 98,* 104230.

Garrison, J., Erdeniz, B., & Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews, 37,* 1297-1310.

Gelman, A., & Rubin. D (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457-511.

Gershman, S. J., (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin and Review*, *22*, 1320-1327.

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology, 68,* 101-128.

Golubickis, M., & Macrae, C. N. (2022). Sticky me: Self-relevance slows reinforcement learning. *Cognition*, *227*, 105207.

Hackel, L. M., & Amodio, D. M. (2018). Computational neuroscience approaches to social cognition. *Current Opinion in Psychology*, *24*, 92-97.

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*, 1233-1235.

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,* 103948.

Hastie, R., & Kumar, P. (1979). Person memory: Personality traits as organizing principles in memory for behaviors. *Journal of Personality and Social Psychology*, *37*, 25-38.

Hastie, R., Schroeder, C., & Weber, R. (1990). Creating complex social conjunction categories from simple categories. *Bulletin of the Psychonomic Society*, *28*, 242-247.

Hewstone, M., & Hamberger, J. (2000). Perceived variability and stereotype change. *Journal of Experimental Social Psychology*, *36*, 103-124.

Hilton, J. L., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology, 47,* 237-271.

Hinton, P. (2017). Implicit stereotypes and the predictive brain: Cognition and culture in "biased" person perception. *Palgrave Communications, 3,* 17086.

Johnston, W. A., & Hawley, K. J. (1994). Perceptual inhibition of expected inputs: The key that opens closed minds. *Psychonomic Bulletin & Review, 1*, 56-72.

Jost, J. T., & Hunyady, O. (2002). The psychology of system justification and the palliative function *European Review of Social Psychology, 13*, 111–153.

Judd, C.M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*, 54-69.

Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): Effects of training on the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology, 78*, 871-888.

Kunda, Z., Miller, D. T., & Claire, T. (1990). Combining social concepts: The role of causal reasoning. *Cognitive Science, 14,* 551-577.

Kunda, Z., & Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: Constructing grounds for subtyping deviants. *Journal of Personality and Social Psychology, 68*, 565-579.

Lindström, B., Golkar, A., & Olsson, A. (2015). A clash of values: Fear-relevant stimuli can enhance or corrupt adaptive behavior through competition between Pavlovian and instrumental valuation systems. *Emotion, 15*, 668-676.

Lockwood, P. L., & Klein-Flugge, M. C. (2020). Computational modelling of social cognition and behaviour: A reinforcement learning primer. *Social Cognitive and Affective Neuroscience, 16,* 761-771.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*, 1122-1135.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology, 51*, 93-120.

Macrae, C. N., Bodenhausen, G. V., Schloerscheidt, A. M., & Milne, A. B. (1999). Tales of the unexpected: Executive function and person perception. *Journal of Personality and Social Psychology, 76,* 200-213.

Macrae, C. N., Hewstone, M., & Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology, 23*, 76-87.

Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology*, *66,* 37-47.

Marsman, M., & Wagenmakers, E.-J. (2017). Three insights from a Bayesian interpretation of the one-sided *p* value. *Educational and Psychological Measurement, 77,* 529-539.

Mauer, K. L., Park, B., & Rothbart, M. (1995). Subtyping versus subgrouping processes in stereotype representation. *Journal of Personality and Social Psychology, 69*, 812-824.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419-457.

Miletić, S., Boag, R. J., & Forstmann, B. U. (2020). Mutual benefits: Combining reinforcement learning with sequential sampling models. *Neuropsychologia, 136,* 1-11.

Miletić, S., Boag, R. J., Trutti, A. C., Stevenson, N., Forstmann, B. U., & Heathcote, A. (2021). A new model of decisional processing in instrumental learning tasks. *eLife,* 10:e63055.

Morgenroth, T., Ryan, M. K., & Peters, K. (2015). The motivational theory of role modelling: How role models influence role aspirants' goals. *Review of General Psychology, 19,* 465-483.

O'Callaghan, C., Kverga, K., Shine, J. M., Adams, R. B., & Bar, M. (2017). Predictions penetrate perception: Converging insights from brain, behaviour, and disorder. *Consciousness and Cognition, 47,* 63-74.

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, reward and decision-making. *Annual Review of Psychology, 68,* 73-100.

Olsson, M., & Martiny, S. E. (2018). Does exposure to counterstereotypical role models influence girls' and women's gender stereotypes and career choices? A review of social psychological research. *Frontiers in Psychology, 9,* 1-12.

Otten, M., Seth, A. K., & Pinto, Y. (2017). A social Bayesian brain: How social knowledge can shape visual perception. *Brain and Cognition, 112,* 69-77.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87,* 532–552.

Pedersen, M. L., & Frank, M. J. (2020). Simultaneous hierarchical Bayesian parameter estimation for reinforcement learning and drift diffusion models: A tutorial and links to neural data. *Computational Brain & Behavior, 3,* 458–471.

Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin & Review, 24,* 1234-1251.

Persson, L. M., Golubickis, M., Dublas, D., Mastnak, N., Falbén, J. K., Tsamadi, D., Caughey, S., Svensson, S., & Macrae, C. N. (2021). Comparing person and people perception: Multiple group members do not increase stereotype-based priming. *Quarterly Journal of Experimental Psychology*, *78*, 1418-1431.

Persson, L. M., Falbén, J. K., Tsamadi, D., & Macrae, C. N. (in press). People perception and stereotype-based responding: Task context matters. *Psychological Research*.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology, 90*, 751–783.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Development Core Team. (2015). *nlme: Linear and nonlinear mixed effects models*. The Comprehensive R Archive Network (CRAN), Vienna, Austria.

Prati, F., Vasiljevic, M., Crisp, R. J., & Rubini, M. (2015). Some extended psychological benefits of challenging social stereotypes: Decreased dehumanization and a reduced reliance on heuristic thinking. *Group Processes and Intergroup Relations, 18*, 801-816. ~Social

Rudman, L. A., & Phelan, J. E. (2010). The effect of priming gender roles on women's implicit gender beliefs and career aspirations. *Social Psychology*, *41*, 192-202.

Quadflieg, S., Flannigan, N., Waiter, G. D., Rossion, B., Wig, G. S., Turk, D. J., & Macrae, C. N. (2011). Stereotype-based modulation of person perception. *NeuroImage, 57*, 549-557.

Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 1226-1243.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*, 333-367.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences, 20,* 260-281.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9,* 438-481.

Richards, Z., & Hewstone, M. (2001). Subtyping and subgrouping: Processes for the prevention and promotion of stereotypes. *Personality and Social Psychology Review, 5,* 52-73.

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience, 23,* 473-500.

Sewell, D. K., & Stallman, A. (2020). Modeling the effect of speed emphasis in probabioisitc category learning. *Computational Brain & Behavior, 3,* 129-152.

Sherman, J. W., Lee, A. Y., Bessenoff, G. R., & Frost, L. A. (1998). Stereotype efficiency reconsidered: Encoding flexibility under cognitive load. *Journal of Personality and Social Psychology, 75,* 589-606.

Sherman, J. W., Macrae, C. N., & Bodenhausen, G. V. (2000). Attention and stereotyping: Cognitive constraints on the construction of meaningful social impressions. *European Review of Social Psychology, 11,* 145-175.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B, 64,* 583-639.

Srull, T. K., & Wyer, R. S., Jr. (1989). Person memory and judgment. *Psychological Review, 96,* 58-83.

Stangor, C., & Duan, C. (1991). Effects of multiple task demands upon memory for information about social groups. *Journal of Experimental Social Psychology, 27,* 357-378.

Stern, L. S., Marrs, S., Millar, M. F., & Cole, E. (1984). Processing time and the recall of inconsistent and consistent behaviors of individuals and groups. *Journal of Personality and Social Psychology*, *47*, 253-262.

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual Review of Psychology, 33*, 1-39.

Tsamadi, D., Falbén, J. K., Persson, L. M., Golubickis, M., Caughey, S., Sahin., B., & Macrae, C. N. (2020). Stereotype-based priming without stereotype activation: A tale of two priming tasks. *Quarterly Journal of Experimental Psychology*, 11, 1939-1948.

Weber, R., & Crocker, J. (1983). Cognitive processes in the revision of stereotypic beliefs. *Journal of Personality and Social Psychology*, 45, 961-977.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics, 7*, 1-10.

Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. *Advances in Experimental Social Psychology*, *46*, 55-123.

Zaki, J., Kallman, S., Wimmer, G. E., Ochsner, K., & Shohamy, D. (2016). Social cognition as reinforcement learning: Feedback modulates emotion inference. *Journal of Cognitive Neuroscience, 28,* 1270-1282.