HHS Public Access

Author manuscript

J Exp Soc Psychol. Author manuscript; available in PMC 2022 September 01.

Published in final edited form as:

J Exp Soc Psychol. 2022 September ; 102: . doi:10.1016/j.jesp.2022.104380.

Untested assumptions perpetuate stereotyping: Learning in the absence of evidence

William T.L. Cox[a,b,*], **Xizhou Xie**[a], **Patricia G. Devine**[a]

[a]Department of Psychology, University of Wisconsin – Madison, United States of America

[b]Inequity Agents of Change Nonprofit, www.biashabit.com, United States of America

Abstract

In the present work, we set out to assess whether and how much people learn in response to their stereotypic assumptions being confirmed, being disconfirmed, or remaining untested. In Study 1, participants made a series of judgments that could be influenced by stereotypes and received feedback that confirmed stereotypes the majority of the time, feedback that disconfirmed stereotypes the majority of the time, or no feedback on their judgments. Replicating past work on confirmation bias, patterns in the conditions with feedback indicated that pieces of stereotype-confirming evidence exerted more influence than stereotype-disconfirming evidence. Participants in the Stereotype-Confirming condition stereotyped more over time, but rates of stereotyping for participants in the Stereotype-Disconfirming condition remained unchanged. Participants who received no feedback, and thus no evidence, stereotyped more over time, indicating that, matching our core hypothesis, they learned from their own untested assumptions. Study 2 provided a direct replication of Study 1. In Study 3, we extended our assessment to memory. Participants made judgments and received a mix of confirmatory, disconfirmatory, and no feedback and were subsequently asked to remember the feedback they received on each trial, if any. Memory tests for the no feedback trials revealed that participants often misremembered that their untested assumptions were confirmed. Supplementing null hypothesis significance testing, Bayes Factor analyses indicated the data in Studies 1, 2, and 3 provided moderate-to-extreme evidence in favor of our hypotheses. Discussion focuses on the challenges these learning patterns create for efforts to reduce stereotyping.

Keywords

Stereotypes; Stereotyping; Learning; Bias; Hebbian learning

Stereotyping restricts opportunities for members of stereotyped groups, and stereotypes are strongly linked to biases, prejudice, discrimination, and oppression (Allport, Clark,

*Corresponding author at: Department of Psychology, University of Wisconsin, Madison, United States of America. william.cox@biashabit.com (W.T.L. Cox).

Appendix A. Supplementary data

Details on stimulus development, trial order/randomization, and additional analyses. Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2022.104380.

& Pettigrew, 1954; Cox, Abramson, Devine, & Hollon, 2012; Cox, Devine, Bischmann, & Hyde, 2016; Devine, 1989; Devine & Sherman, 1992; Fiske, 1998; Pratto, Sidanius, Stallworth, and Malle, 1994; Sue et al., 2007). For these reasons, stereotyping is discouraged by both social norms and many people's personal nonprejudiced values and standards, and people often struggle to reduce the influence of stereotypes on their thoughts, feelings, and behavior (Blanchard, Lilly, & Vaughn, 1991; Cox et al., 2012; Devine, 1989; Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002; Monteith, Deneen, & Tooman, 1996; Plant & Devine, 1998). Nevertheless, stereotypes persist, in culture and within individuals' minds. Many people report having stereotypes "pop to mind", in opposition to their consciously espoused personal convictions that oppose stereotypes and biases (Cox et al., 2012; Devine, 1989). Indeed, for many whose values oppose bias, automatically activated stereotypes constitute a serious personal dilemma, resulting in sustained efforts to reduce stereotypes (Cox et al., 2012; Cox & Devine, 2019; Devine, 1989; Devine, Forscher, Austin, & Cox, 2012). These efforts, however, often fail. Stereotypes persist and affect people's judgments and treatment of others, in spite of personal, institutional, or cultural forces that oppose stereotypes.

Because stereotypes are the cognitive component of intergroup biases, many advances in the stereotyping literature draw on established concepts and work in the cognitive and cognitive neuroscience literature (Cox & Devine, 2015; Hamilton, Sherman, Wyer, & Srull, 1994; Hilton & Von Hippel, 1996). Continuing that trend in the present work, we draw on advances in cognitive psychology and cognitive neuroscience to formulate and test a theoretical case for how basic neural learning processes contribute to stereotype persistence. Specifically, we identify how known Hebbian and hedonic (reward/aversion) learning processes operate when people make stereotypic inferences, then test key predictions from that model in three experiments.

1. Confirmed, disconfirmed, or untested stereotypic assumptions

The most commonly discussed way that learning contributes to stereotype persistence is *confirmation bias* — perceivers accord evidence that confirms stereotypic expectations greater weight than evidence that disconfirms stereotypic expectations (Nickerson, 1998; Pohl, Pohl, & editors., 2004). Confirmation biases thereby perpetuate stereotypes even when people are presented with equal amounts of confirmatory and disconfirmatory evidence (e.g., Darley & Gross, 2004). In many cases, however, a perceiver makes an inference but receives no evidence, never learning whether the inference was correct or incorrect. These circumstances are ripe for stereotyping because one of the major functions of stereotypes is to "fill in gaps" when the evidence available to perceivers is incomplete. These stereotypic inferences are *untested assumptions* — predictions or hypotheses that are neither confirmed nor disconfirmed.

Consider how untested assumptions may guide behavior in some fairly typical situations: A pedestrian may rely on the *Black→Criminal* stereotype to predict that a Black man is dangerous and cross the street to avoid him, stereotypes about gender and leadership may lead an employer to think that a woman cannot be an effective leader and decline to hire her, or a straight woman may assume that a fashionable man is gay and fail to ask him on a

date (Cox & Devine, 2015; Devine, 1989; Dunning & Sherman, 1997; Oliver, 2003; Steele, 2011). Not only did stereotype-supplied assumptions serve a basis for behavior in these examples, but the behaviors actually precluded the stereotypic assumption from being tested. In other cases, people may make an assumption but never have an opportunity to learn whether or not their assumption was correct. We argue that there are many circumstances in which perceivers never learn whether their stereotypic assumptions were correct or incorrect.

What effects do these untested assumptions have on the future likelihood of making similar stereotypic inferences? From a purely rational learning perspective, these untested assumptions should have no impact on the likelihood of future stereotypic inferences. Because the perceivers have no objective evidence about their assumptions, no rational learning can occur. Human learning processes, however, are not always rational (e.g., Gibson, Rogers, & Zhu, 2013; Kalish, Rogers, Lang, & Zhu, 2011; McClelland, 2006; McClelland et al., 2010). Drawing on established theory and evidence in the cognition and cognitive neuroscience literature, reviewed below, we propose that merely making a stereotypic assumption strengthens the stereotypic associations, which increases the likelihood of the same assumption being made in the future. In other words, people learn from their own untested assumptions, in the absence of external evidence.

Our theorizing in the present work is rooted in biologically-based cognitive theory (e.g., connectionism, parallel distributed processing models, emergentist approaches, see McClelland et al., 2010). A core tenet of this discipline is that, because cognitive processes emerge from brain activity, consideration of neural mechanisms can and should inform our theorizing about how cognitive processes unfold (Cox & Devine, 2015; McClelland et al., 2010). Mental concepts, for example, are encoded across distributed clusters of neurons in the brain, thus our understanding of how neural clusters interact in the brain can be useful to derive theoretical predictions for how mental concepts interact. In the following sections, we briefly review basic details about two key neural learning mechanisms, which then serve as a basis to make predictions about stereotyping processes. Specifically, Hebbian (i.e., activation-based) learning should strengthen stereotypic associations when someone makes a stereotypic assumption. If the perceiver subsequently receives confirmation or disconfirmation of that stereotypic assumption, hedonic (i.e., reward/aversion) learning will then subsequently moderate that initial Hebbian learning. Consideration of the independent and joint contributions of these two learning processes provides a foundation from which to derive predictions in the present work, which we then test in three experiments.

2. Hebbian learning: learning via activation

Our predictions about *learning from untested assumptions* are tightly linked to Hebbian mechanics (Hebb, 1949), which form the neural basis for association learning (Cox & Devine, 2015; Markram, Lübke, Frotscher, & Sakmann, 1997; McClelland, 2006; McClelland et al., 2010). When neuron or neural cluster *A* is activated immediately before neuron/neural cluster *B*, spike-timing-dependent plasticity strengthens the synaptic connections from *A* to *B* (Markram et al., 1997). This Hebbian synaptic strengthening is most commonly discussed in terms of pre-synaptic and post-synaptic activations both being

initiated by external stimuli (e.g., stimulus A activates *A,* then stimulus B activates *B*, thus $A \rightarrow B$ is strengthened). See Fig. 1.

Importantly for the present work, Hebbian learning also occurs when the post-synaptic activation is generated internally (e.g., stimulus A activates *A*, which activates *B,* thus $A \rightarrow B$ is strengthened, without stimulus B being externally present) (Antony, Ferreira, Norman, & Wimber, 2017; Markram et al., 1997; McClelland, 2006). See Fig. 2. This *inferential Hebbian learning*, therefore, involves internally-generated activations strengthening preexisting notions in the absence of external evidence. Prior work has demonstrated the effects of this mechanistic learning process within the machine learning literature, showing its effects on category assumptions in computational models (cf. "semi-supervised learning; Gibson et al., 2013; Zhu & Goldberg, 2009). Others have translated this computational work to human learning processes, showing that inferential Hebbian learning also occurs in human participants (Gibson et al., 2013; Kalish et al., 2011; Palmeri & Flanery, 1999, 2002; Zaki & Nosofsky, 2007; Zhu, Rogers, Qian, & Kalish, 2007). This past work with human participants uses categorization tasks involving novel objects or fictitious social groups. To our knowledge, no work has yet tested these processes with perceptions of real social groups and pre-existing stereotypes.

In the present article, therefore, we extend this prior work, arguing that these mechanics should operate during stereotyping (e.g., seeing a Black person activates *Black*, which activates *Criminal*, thus $Black \rightarrow Criminal$ is strengthened). When someone makes a stereotypic inference, inferential Hebbian learning will strengthen neural associations that make up the stereotype even if no evidence is presented, making the same stereotypic inference more likely in the future. In this way, we propose that people learn from their untested stereotypic assumptions.

3. Hedonic learning: learning via feedback

In cases where a perceiver *does* receive feedback about their assumptions, additional learning mechanisms come into play. Receiving feedback about a probabilistic inference involves hedonic (reward/aversion) learning processes: confirmatory feedback yields a positive prediction error, which is rewarding, and disconfirmatory feedback yields a negative prediction error, which is aversive (Frank, Seeberger, & O'reilly, 2004; McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002; McClelland, 2006; Reggev, Chowdhary, & Mitchell, 2021; Tricomi, Delgado, McCandliss, McClelland, & Fiez, 2006; Tricomi & Fiez, 2008). Once feedback is received, this hedonic learning initiates dopaminergic processes that modulate the Hebbian synaptic strengthening (McClelland, 2006; Montague, Dayan, & Sejnowski, 1996; Schultz, 2016).

When learners make an inference and subsequently receive feedback, these hedonic learning processes would operate *in addition to* the inferential Hebbian learning discussed above. The joint contributions of inferential Hebbian learning and hedonic learning could explain the differential weighting of confirmatory and disconfirmatory evidence observed in confirmation bias patterns. As noted above, when someone makes an inference, inferential Hebbian learning is hypothesized to strengthen the association (+1). If the learner

subsequently receives confirmatory feedback, the hedonic learning reward signal further strengthens the association (+2), yielding a higher combined net learning weight (+3). And if the learner receives disconfirmatory feedback, the inferential Hebbian learning (+1) and the hedonic learning aversion signal (−2) work against one another, such that the net learning weight (−1) is weaker for disconfirmatory than confirmatory evidence (Gilbert, 1991; McClelland, 2006).[1] See Fig. 3. In this way, the initial inferential Hebbian learning essentially interferes with the disconfirmation, yielding a confirmation bias pattern.

4. The present work

Studies 1 and 2 were designed to explore how learning processes perpetuate stereotyping in the presence and absence of evidence. We hypothesized that, when people make a stereotypic assumption but receive no evidence about it, merely making the stereotypic assumption will strengthen the stereotypic association, leading to higher rates of stereotyping over time. We also hypothesized that, consistent with past work on confirmation bias, as well as our model elaborated above, when people receive evidence about their stereotypic assumptions, confirmatory evidence has a greater influence than disconfirmatory evidence. To test these hypotheses, participants made a series of sequential judgments that could be influenced by stereotypes. Participants were randomly assigned either to receive no feedback regarding the accuracy of their judgments, to receive a pattern of feedback that mostly confirmed stereotypes, or to receive a pattern of feedback that mostly disconfirmed stereotypes. Learning was assessed by examining whether and how participants' rate of stereotyping changed over time as a function of this feedback manipulation. Study 3 extended Studies 1 and 2 to test participants' memory involving untested assumptions.

5. Study 1

5.1. Method

5.1.1. Participants and design—Undergraduate participants (N = 469, 69% Female, 89% White, 98% Straight) consented to participate and completed the experiment in person, in exchange for extra credit in their introductory psychology course. All participants received the same amount of extra credit compensation; there were no additional incentives based on how they performed in the task. A priori power analyses indicate that to achieve power of 1-β = 0.95 to detect the predicted effect (estimated at f = 0.15, based on rounding down a comparable effect size using a similar task in Study 5 of Cox et al., 2016), each condition should contain 141 participants, which we rounded to 150 participants per condition (power calculated with G*Power; Faul, Erdfelder, Buchner, & Lang, 2009). We stopped data collection at the end of the week after we hit our target percondition sample size.

[1]These numerical values are placeholders to allow relative comparison across the different possible outcomes of a stereotypic assumption (i.e., remaining untested, being disconfirmed, or being confirmed). They are not intended to be indicators that can be applied to make precise numerical predictions, like one might explore in computational learning models.

Participants made a series of sequential judgments that could be affected by stereotypes, and we assessed learning by comparing rates of stereotyping on trials in the Training block versus the Test block. The key hypothesis test centered on comparisons between three between-subjects feedback conditions, to which participants were randomly assigned. In the *No Feedback* condition (n = 154), participants received no feedback on their judgments. In the *Stereotype-Confirming* (n = 156) and *Stereotype-Disconfirming* (n = 159) conditions, participants received feedback after each trial in the Training block, with most of the trials confirming stereotypes or disconfirming stereotypes, respectively (explained in further detail below). The present experiment therefore had a 3 (Feedback Condition: Stereotype-Confirming vs. Stereotype-Disconfirming vs. No Feedback) × 2 (Block: Training vs. Test) mixed factorial design, with Feedback Condition between-subjects and Block within-subjects.

5.1.2. Stereotyping categorization task—Testing the present hypotheses required an experimental task that had the following features. First, the task must involve participants making a series of sequential judgments, so that we could evaluate learning over time. Second, participants' judgments needed to be informed by prior stereotypic expectations. Third, participants needed to find the task plausible (i.e., they needed to believe that they could make the inferences based on information supplied). Fourth, the task needed to involve judgments that would not strongly raise social desirability concerns. Fifth and finally, the presumed accuracy of judgments required some uncertainty so that feedback about the accuracy of the inferences could be manipulated without the task losing believability. The Stereotyping Categorization Task (SCT; Cox et al., 2016), meets these criteria and was adapted for the present study.

The SCT is framed as a task evaluating the extent to which people can accurately categorize men as gay or straight based on their (ostensibly real) social media profiles. People tend to rely on gay-stereotypic interests (e.g., fashion, shopping) as stereotypic cues to judge men as gay, and straight-stereotypic interests (e.g., sports, athletics) as stereotypic cues to judge men as straight (Cox et al., 2016; Cox & Devine, 2015). Importantly, participants often do not think of this categorization process as "stereotyping," — a cultural phenomenon known as the gaydar myth relabels this kind of stereotyping as "gaydar", a purported "sixth sense" that enables people to judge who is gay or straight (Cox et al., 2016). Widespread cultural prevalence of the gaydar myth therefore reduces concerns about the social undesirability of stereotyping in this task (Cox et al., 2016; Cox & Devine, 2014, 2015). The SCT's social media cover story was also well-suited to our purposes, because social media profiles are specifically designed to simply and directly convey personal information in a standardized format. We therefore could easily and believably manipulate the profiles to display characteristics of our targets that serve as the basis for social judgments. The social media cover story places the task in a context that is highly familiar to most participants, who use social media to make judgments in their daily lives and are often interested in testing their ability to accurately intuit information based on social media.

In the present version of the SCT, participants made judgments based on screenshots of fabricated social media profiles. Each profile screenshot displayed a profile picture of a White male, information saying he attended the university at which the study was conducted,

and five interests (i.e., “likes”) of the profile owner. Each profile had two “filler” interests that we pretested to be neutral relative to gay and straight male stereotypes (details of stimulus pretesting is reported in the Supporting Information). The remaining three interests differentiated the profiles into one of three categories: Shopping profiles, Sports profiles, or Neutral profiles. The Neutral profiles contained only stereotype-irrelevant interests and served as filler trials; they are discussed no further. The Shopping profiles displayed three interests related to shopping/fashion, because, as noted above, past work shows that people tend to stereotype men with shopping-related interests as gay (Cox et al., 2016; Cox & Devine, 2014, 2015). The Sports profiles displayed three sports-related interests, because past work shows that people tend to stereotype men with sports-related interests as straight (Cox et al., 2016; Cox & Devine, 2015). All participants saw a mix of these Neutral, Shopping, and Sports profiles.

Participants were told that “about half” of the men were gay, to counteract any prior base rate expectations they may have. In the SCT, participants pressed “G” or “S” on the keyboard to categorize men as gay or straight. Reminders of the response keys remained on-screen for the duration of the experiment. Participants made judgments about a total of 30 profiles, broken into a Training block (20 trials) and a Test block (10 trials). No profile appeared more than once. Importantly, the stimulus profiles seen by participants did not differ by feedback condition. The only differences by condition were whether participants received feedback and the overall pattern of feedback participants received.

The feedback manipulation operated on the Training block’s trials. Participants in the No Feedback condition received no feedback on their judgments. In the Stereotype-Confirming condition, participants received feedback after each trial in the Training block, with 66–75% of the trials confirming stereotypes and 25–33% of trials disconfirming stereotypes (percentages exclude the Neutral filler trials). The Stereotype-Disconfirming condition reversed this pattern, with 66–75% disconfirmatory and 25–33% confirmatory trials in the Training block. Further details about the trial type distributions and randomization schemes are reported in the Supporting Information.

In the Stereotype-Confirming and Stereotype-Disconfirming conditions, feedback was given immediately after each judgment, with the stereotypic expectation for that trial either being confirmed (i.e., a man with sports interests was revealed to be straight/a man with shopping interests was revealed to be gay) or disconfirmed (i.e., a man with sports interests was revealed to be gay/a man with shopping interests was revealed to be straight). The profile remained on the screen, and feedback appeared below the profile and remained on the screen for 5 s, allowing the participant time to encode the feedback and look over the profile again. If the participant’s judgment was “correct” relative to the orientation assigned to a given profile, the feedback read, in green font, “Correct - That profile is a (Straight/Gay) man.” If the participant’s judgment was “incorrect,” the feedback read, in red font, “Error - That profile is a (Straight/Gay) man.” In the No Feedback condition, after participants made a categorization, the task moved on to the next trial.

After the Training block, there was a brief pause to reiterate the task instructions, then participants completed the Test block, which contained 5 Shopping profiles and 5 Sports

profiles, on which no one received feedback. Rates of stereotyping were computed as the percentage of trials on which participants made stereotype-congruent judgments (i.e., judging sports profiles as straight and shopping profiles as gay) in each block.

5.2. Results

Across conditions, participants stereotyped at a rate of 64.1% in the Training block, which is higher than 50% chance, $t(468) = 20.280$, $p < 0.001$, indicating that participants' categorizations were indeed influenced by the stereotypes, as in past versions of the SCT (Cox et al., 2016).

The key hypothesis test was a 3 (Condition: Stereotype-Confirming vs. Stereotype-Disconfirming vs. No Feedback) × 2 (Block: Training vs. Test) mixed ANOVA on the stereotyping percentage rates, with Feedback Condition between-subjects and Block within-subjects. A sensitivity power analysis for this test ($\alpha = 0.05$; $1\text{-}\beta = 0.80$) with the observed repeated-measures correlation ($r_{within} = 0.211$) revealed that we have sufficient power to detect effects as small as $\eta_p^2 = 0.0081$. See descriptive statistics in Table 1 and ANOVA statistics in Table 2.

The main effects of Feedback Condition and Block were qualified by the predicted Condition × Block interaction. Rates did not differ by condition in the Training block, $F(2,466) = 0.43$, $p = 0.651$, $\eta_p^2 = 0.002$; the interaction was driven by participants' rates of stereotyping differing by condition in the Test block, $F(2,466) = 9.09$, $p < 0.001$, $\eta_p^2 = 0.038$. Stereotyping rates in the Test block of the No Feedback condition did not statistically differ from those of the Stereotype-Confirming condition, $F(1, 308) = 2.36$, $p = 0.126$, $\eta_p^2 = 0.008$, but the Stereotype-Disconfirming condition's Test block rates were significantly lower than both the No Feedback condition, $F(1,311) = 6.53$, $p = 0.011$, $\eta_p^2 = 0.021$, and the Stereotype-Confirming condition, $F(1,313) = 19.471$, $p < 0.001$, $\eta_p^2 = 0.059$. The Condition x Block interaction indicated that learning patterns differed based on feedback condition. In the Stereotype-Confirming ($d = 0.53$) and No Feedback ($d = 0.26$) conditions, participants stereotyped more over time, but there was no statistically significant change in the Stereotype-Disconfirming condition ($d = 0.04$). See Fig. 4.

5.2.1. Bayesian analyses—Our a priori hypotheses were formulated using the mixed between- and within- subjects design, and mixed ANOVA analysis, as reported above. Reviewers helpfully suggested that an even stronger test of our hypothesis would be to conduct Bayes factor analyses within each condition. One of the main benefits of using Bayes Factor for hypothesis testing is that it provides an indicator of the relative likelihood that the alternate vs null hypothesis is true, rather than simply rejecting the null hypothesis under the *p*-value approach (Wagenmakers et al., 2018). We used JASP (2020) to conduct a one-sided Bayesian Paired Samples *t*-test to test the alternate hypothesis that the No Feedback participants' stereotyping rates in the Test block were higher than their stereotyping rates in the Training Block. Because we do not have a basis for predetermining a Bayesian prior, we left the Cauchy prior scaled at the recommended default value, 0.707. The estimated Bayes Factor, $BF_{-0} = 5.293$, indicated the data were 5.293 times more

likely to occur under a model where, matching our hypothesis, the rates of stereotyping in test block are higher than those in training block, rather than the null hypothesis model where the rates of stereotyping in test block equal those in training block. Following Lee & Wagenmakers, 2013 rule of thumb, this BF for the No Feedback condition indicates that our data provided *moderate* evidence in favor of our hypothesis. We conducted the same analysis for the Stereotype-Confirming condition, and its $BF_{-0} > 999$ provided *extreme* evidence that rates increased in that condition.

The lack of statistically significant change in the Stereotype-Disconfirming condition cannot be used as evidence of a lack of change in stereotyping rates. Another benefit of a Bayesian analysis, however, is that it can provide evidence of no change (i.e., evidence in favor of the null hypothesis). To that end, we conducted a two-sided paired t-test in the Stereotype-Disconfirming condition, yielding $BF_{01} = 10.521$, which indicates our data provided *moderate* evidence in favor of the null hypothesis that the train and test rates are not different. This pattern supports our interpretation that rates in the Stereotype-Disconfirming condition did not change.

5.3. Discussion

Study 1 supported the hypotheses derived from our model. In the No Feedback condition, rates of stereotyping increased, despite participants receiving no evidence related to the accuracy of the stereotypes or their judgments. Supporting our central hypothesis, participants learned from their own assumptions, in the absence of objective evidence related to the accuracy of the stereotypes. Participants in the No Feedback condition learned from their untested assumptions, stereotyping more over time, with no rational or external evidentiary basis for this increase.

In the Stereotype-Confirming condition, participants stereotyped more over time, but in the Stereotype-Disconfirming condition, participants' rates of stereotyping did not change. Comparison of these two conditions demonstrates that stereotype-confirming evidence carried more weight than stereotype-disconfirming evidence, consistent with our model and past work on confirmation bias. The lack of change observed in the Stereotype-Disconfirming condition is striking, given that this condition contained two-to-three times as many disconfirmatory trials as confirmatory trials. The pattern observed in the Stereotype-Disconfirming condition suggests that each piece of confirmatory evidence carried up to three times the weight of each piece of disconfirmatory evidence. This imbalance conceptually replicates past work on confirmation bias (Darley & Gross, 2004; Lord, Ross, & Lepper, 1979).

In the Stereotype-Confirming and Stereotype-Disconfirming conditions, after participants made a categorization, the stimuli remained on the screen for 5 s, to allow the participants to receive and encode the feedback. After participants in the No Feedback condition made a categorization, however, the task immediately moved on to the next trial. The task procedure, therefore, included a potential confound, because the No Feedback condition participants were exposed to the stimuli for a shorter time period than participants in the other two conditions. This potential confound does not call into question the core finding that people learned from their untested assumptions. It could, however, limit our ability

to interpret the learning patterns when compared across the three conditions, given that participants in the No Feedback condition saw the stimuli for less time. We corrected this issue in Study 2, replicating Study 1's test of our hypotheses without this potential confound.

6. Study 2

6.1. Method

We conducted Study 2 as a higher-powered replication of Study 1. Undergraduate participants (N = 682; 62% Female, 78% White, 95% Straight) were randomly assigned to one of three between-subjects conditions (Stereotype-Confirming, Stereotype-Disconfirming, or No Feedback). This sample size granted power of 1-β =0.999 to detect the key effect from Study 1.

Study 2 was nearly identical to Study 1, with one change in the SCT. In the No Feedback condition, the stimuli remained on-screen for 5 s after each categorization, so that No Feedback participants saw the stimuli for the same amount of time as participants in the other two conditions. After participants in the No Feedback condition made their judgment on each trial, the program acknowledged the participants' categorization, displaying "You selected ('Straight'/'Gay') for this profile" in black font while the stimulus remained on the screen. Otherwise, procedures were identical to Study 1.

6.2. Results and discussion

As in Study 1, we conducted a 3 (Condition: Stereotype-Confirming vs. Stereotype-Disconfirming vs. No Feedback) x 2 (Block: Training vs. Test) mixed ANOVA on the stereotyping percentage rates, with Condition between-subjects and Block within-subjects. Sensitivity power analysis for this test (α = 0.05; 1-β = 0.80) with the observed repeated-measures correlation (r_{within} = 0.251) revealed that we have sufficient power to detect effects as small as η_p^2 = 0.0053. See descriptive statistics in Table 3 and ANOVA statistics in Table 4. Replicating Study 1, the main effects of Feedback Condition and Block were qualified by the predicted Condition x Block interaction, with participants' rates of stereotyping increasing in the No Feedback (d = 0.18) and the Stereotype-Confirming (d = 0.52) conditions, but there was no evidence of change in the Stereotype-Disconfirming (d = − 0.11) condition. Test rates in the Stereotype-Confirming condition were significantly higher than test rates of the No Feedback condition, F(1,448) = 8.53, p = 0.004, η_p^2 = 0.019, and the Stereotype-Disconfirming condition F(1,452) = 68.81, p < 0.001,η_p^2 = 0.132. The test rates of the No Feedback condition were significantly higher than those of the Stereotype-Disconfirming condition, F(1,458) = 32.15, p < 0.001,η_p^2 = 0.066. See Fig. 5.

Because the Study 2 procedure directly replicates Study 1, we used the Replication Bayes Factor to quantify the replication attempt given the prior data (Ly, Etz, & Marsman, 2019). The replication Bayes Factor BF(d_{rep} | d_{orig}) is a measurement of replication success computed by setting the prior distribution for the replication study as the calculated posterior distribution from the original study (Ly et al., 2019). It is calculated by dividing the complete Bayes factor from Studies 1 and 2, BF_{+0} (d_{orig}, d_{rep}), by the Bayes factor from

Study 1, BF_{+0} (d_{orig}) (Ly et al., 2019). Again using JASP, the same one-sided Bayesian Paired Samples *t*-test was performed on the complete data from the No Feedback condition of Studies 1 and 2, with default Cauchy prior scaled at 0.707. The complete Bayes Factor, BF_{+0} (d_{orig}, d_{rep}) = 36.61, indicates that together, Study 1 and 2's data provide *very strong* evidence in favor of the alternative hypothesis. See Fig. 6. Dividing the complete BF by Study 1's original BF_{+0} (d_{orig}) = 5.293 yields the replication BF_{+0}(d_{rep} | d_{orig}) = 6.92. This replication BF indicates that, with Study 1's data as a prior, Study 2's data were predicted by the alternative hypothesis 6.92 times better than by the null hypothesis. These patterns indicate that Study 2 replicates Study 1 *moderately* well in the No Feedback condition.

In the Stereotype-Confirming condition, the complete BF_{-0} (d_{orig}, d_{rep}) > 999 provided *extreme* evidence that rates increased, and yielded a replication BF_{+0}(d_{rep} | d_{orig}) > 999. For the Stereotype-Disconfirming condition, the two-sided paired t-test complete BF_{01} (d_{rep}, d_{orig}) = 13.628 provided *strong* evidence in favor of the null and yielded a replication BF_{01} (d_{rep} | d_{orig}) = 1.295.

7. Study 3

In Study 3, we extended our exploration of untested assumptions to memory. It is unclear the extent to which people are aware of the influence of their untested assumptions, and whether they will recognize and remember that an assumption remained untested. To justify acting on their assumptions, perceivers may consciously or tacitly assume their untested judgments are correct (Bem, 1972; Kunda, 1990). Or, perceivers may consciously recognize that their assumption was untested and should not be used as evidence (even though the learning processes may operate outside awareness and lead them to stereotype more). We predict that people will disproportionately misremember their untested assumptions as having been confirmed. Study 3 was designed to test this hypothesis, assessing how often perceivers accurately remembered that their assumption was untested, compared to how often they misremembered their untested assumptions as having been verified.

In Study 3, all participants completed the SCT and received a mix of stereotype-confirming feedback, stereotype-disconfirming feedback, or no feedback. Then, they were shown each stimulus again, then asked to report whether they had received feedback and, if so, what that feedback was. Of interest was their incorrect memories of the No Feedback trials. Specifically, we examined whether they misremembered their untested assumptions as having been externally validated.

7.1. Method

We recruited 289 undergraduates (58% Female, 67% White, 94% Straight). The procedure began the same as in Studies 1 and 2, with participants making gay/straight categorizations. On each of the 20 trials in the Training block, participants randomly received either no feedback, stereotype-confirming feedback, or stereotype-disconfirming feedback. On the No Feedback trials, after participants made a categorization, the profile remained on-screen for 5 s, with the message "Response Recorded." Stereotype-confirming and stereotype-disconfirming feedback was delivered using the same procedure as in Studies 1 and 2. All participants received no feedback on the 10 trials in the Test block.

After completing the Training and Test blocks, participants completed a memory task, which serves as the primary outcome of interest. Participants were not told in advance that they would be asked to remember the feedback they received in the first part of the study. In the memory task, participants were presented with the same 30 stimuli a second time, in a new, random order. For each stimulus, they were asked to report what feedback they received. They were presented with four choices: “I was told this man is Gay”, “I was told this man is Straight”, “I did not receive feedback on this profile”, or “I do not remember whether I received feedback”. Participants did not receive feedback about their responses to the memory trials.

7.2. Results

On average, of the non-neutral memory trials, participants correctly remembered 12.2 (*sd* = 4.07) trials (50.48%, *sd* = 16.606), reported that they did not remember 4.02 (*sd* = 3.38) trials (16.64%, *sd* = 13.956), and incorrectly remembered 7.92 (*sd* = 4.04) trials (32.88%; *sd* = 16.646). Out of the incorrectly remembered trials, 4.48 (*sd* = 3.19) of these incorrect memory responses (53.97%, *sd* = 23.549) matched their initial response, and 3.44 (*sd* = 2.15) did not match their original response (46.03%, *sd* = 23.549). The accuracy of participants’ memory differed by trial type, $F(2, 576) = 12.154$, $p < 0.001$, $\eta_p^2 = 0.040$. Specifically, participants’ memory was more accurate for stereotype-confirming trials (56.7% correct) than for no feedback trials (49.4% correct) or stereotype-disconfirming trials (46.8% correct).

Of key interest in Study 3 was whether participants misremembered their untested assumptions as confirmed. The following analyses therefore focus solely on the memory tests for the no feedback trials. Overall, for the No Feedback trials, participants correctly remembered that they received no feedback 49.4% of the time, they reported that they could not remember 18.6% of the time, and 32.0% of the time, they misremembered, reporting incorrectly that they had received feedback. There were 17 participants who never misremembered receiving feedback on a no feedback trial. Because our hypotheses focus specifically on misremembered no feedback trials, we excluded these 17 participants from the subsequent analyses. The remaining 272 participants correctly reported that they received no feedback 47.8% of the time, reported that they could not remember 18.1% of the time, and they misremembered 34.0% of the time. With these 272 participants, a sensitivity power analysis test ($\alpha = 0.05$; $1\text{-}\beta = 0.80$) revealed that both our paired *t*-tests and our one-sample t-tests have sufficient power to detect effects as small as $d = 0.17$.

7.2.1. Misremembering as assumption-confirming vs assumption-disconfirming—

For our purposes, misremembered no feedback trials can come in two forms: The participants could misremember that they received feedback that their untested assumption was confirmed, or they could misremember that their untested assumption was disconfirmed. If these misremembered responses were merely a product of random chance error in memory, we would expect that about 50% of the misremembered trials would be misremembered as assumption-confirming and 50% of the misremembered trials would be misremembered as assumption-disconfirming. If, however, >50% of the misremembered trials are assumption-confirming, that would indicate that making an untested assumption

influences participants to later misremember their untested assumptions as confirmed. Matching this hypothesized pattern, participants' false memories were much more likely to be assumption-confirming (M = 69.30%, sd = 28.34) than assumption-disconfirming (M = 30.70%), d = 1.36, and this percentage of assumption-confirming false memories was higher than 50% chance, d = 0.68, t (271) = 11.229, p < 0.001.

7.2.2. Misremembering as stereotype-confirming vs stereotype-disconfirming—Because participants tend to stereotype at rates higher than chance on the SCT, it is possible that participants were merely stereotyping again in the memory task, rather than being influenced by their prior untested assumptions. In other words, the effects described above could be due to forces operating during the memory recall, rather than during encoding. To evaluate this alternate explanation, we conducted an additional analysis. Instead of categorizing the misremembered no feedback trials by whether participants misremembered their *assumptions* being confirmed/disconfirmed, we categorized them by whether participants misremembered that the targets confirmed/disconfirmed *stereotypes*. Comparing rates of misremembering confirmation/disconfirmation of participants' assumptions versus stereotypes will help disentangle whether it is indeed participants' prior assumptions that lead to the false memories. On trials in which participants' untested assumptions matched stereotypes, there is no way to distinguish whether participants misremembered those trials as assumption-confirming versus stereotype-confirming. However, because participants did not stereotype on every trial, there will be some trials on which misremembering assumption-confirming feedback will yield a different pattern than misremembering stereotype-confirming feedback. In this analysis, we therefore tested whether participants' rates of misremembering their assumptions as confirmed were higher than their rates of misremembering stereotypes as confirmed. This pattern would indicate that participants were indeed falsely remembering their assumptions as validated, rather than merely stereotyping during the memory block, without being influenced by their prior untested assumptions.

Of the misremembered No Feedback trials, participants falsely remembered that targets were stereotype-confirming 61.19% (sd = 30.28) of the time, and misremembered them as stereotype-disconfirming 38.81% of the time. Although they misremembered targets as stereotype-confirming at a rate higher than 50% chance, t (276) = 6.095, p < 0.001, d = 0.37, their rate of misremembering their *assumptions* as confirmed (69.30%) was still higher than their rate of misremembering the *stereotype* as being confirmed (61.19%), t (271) = 3.890, p < 0.001, d = 0.28. In other words, their own prior untested assumptions (whether matching stereotypes or not) influenced participants' false memories more than any new stereotype activation during the memory block. See Fig. 7.

7.2.3. Bayesian analyses—We again replicated the three planned analyses, reported above, using Bayes Factor analyses with the Cauchy prior at the default value of 0.707. A one-sided Bayesian One Sample t-test tested the hypothesis that the rates of misremembering untested assumptions as confirmed were larger than 50%. The estimated Bayes Factor (BF_{+0} > 999) indicated our data provided *extreme* evidence in favor of the alternative hypothesis. Testing whether rates of misremembering no feedback trials as

stereotype-confirming were larger than 50% resulted in an estimated Bayes Factor (BF_{+0} > 999) likewise indicating *extreme* evidence in favor of the alternative hypothesis. To evaluate whether participants were influenced by their own untested assumptions more than novel stereotype activation, a one-sided Bayesian Paired Samples t-test was conducted. This estimated Bayes Factor (BF_{+0} = 198.626), also indicated that there was *extreme* evidence in favor of the alternative hypothesis that, when participants misremembered, it was more likely driven by their prior assumptions than by novel stereotype activation.

7.2.4. Count analyses—Our reviewers helpfully suggested that examining the Study 3 data as counts, rather than percentages, might strengthen our arguments. Analyses treating these data as counts, as shown in Table 5, support the same conclusions as the percentages, for both null hypothesis significance testing and Bayesian analytic approaches. Participants' misremembered No Feedback trials (M = 4.75) were much more likely to be assumption-confirming (M = 3.27, sd = 2.50) compared to assumption-disconfirming (M = 1.49, sd = 1.60), more likely to be stereotype-confirming (M = 2.83, sd = 2.20) compared to stereotype-disconfirming (M = 1.93, sd = 1.83), and more likely to be assumption-confirming than stereotype-confirming.

7.3. Discussion

Participants accurately remembered that their assumptions were untested only half the time. When they misremembered, participants were influenced by their untested assumptions, most often misremembering that their own untested assumptions had been externally confirmed to be true.

8. General Discussion

The present studies provide evidence that learning from one's assumptions is a strong contributor to the persistence of stereotypes. Replicating research in the cognitive literature (Gibson et al., 2013; Kalish et al., 2011), when perceivers lacked external evidence, they learned from their own internally-generated assumptions (Studies 1 & 2). Because this internally-generated learning must occur temporally before any feedback, it likely enhances the effect of confirmatory feedback but detracts from disconfirmatory feedback. Matching this prediction, when perceivers received external evidence about their assumptions, learning patterns indicated that confirmatory evidence had more influence than disconfirmatory evidence (Studies 1 & 2). This unequal weighting of confirmatory and disconfirmatory evidence conceptually replicates past work on confirmation bias (Nickerson, 1998; Pohl et al., 2004). Further, people incorrectly remembered their untested assumptions as having been confirmed (Study 3). Stereotypes may be so difficult to overcome because, in everyday life, they often lead to passing assumptions that remain untested — hypotheses apparently proven true solely by thinking them so (Uhlmann & Cohen, 2007).

8.1. Limitations and future directions

As with any experimental studies, the present work has potential limitations. Some limitations relate to external validity: Out in the real world, people are not often asked to explicitly categorize a series of people based on limited information, and social

environments are frequently more dynamic than the stimuli in our studies. Prior research has shown, however, that people do make tacit stereotypic inferences about others without being explicitly asked to do so (Dunning & Sherman, 1997), and in the case of stereotyping to infer orientation specifically, people use stereotypic cues to spontaneously make orientation judgments without any instruction to do so (Cox & Devine, 2014, 2015).

Another potential limitation is the low number of trials used (only 20 trials in the training block, 10 in the test block). With stereotyping experiments, the number of trials used often must be balanced against social desirability and demand characteristics concerns. A greater number of trials makes it more likely that participants will connect the task to the actual or imagined purpose of the study. Future work may seek to replicate the present effects in more externally valid tasks and find ways to include more trials without raising substantial demand characteristics.

Unexplored in the present work is any effect of individual differences on the processes at play. People differ in the extent to which they express or regulate stereotypic responses (e.g., Axt & Trawalter, 2017; Burns, Monteith, & Parker, 2017; Cox, 2015; Devine, 1989; Monteith, Ashburn-Nardo, Voils, & Czopp, 2002), and effortful stereotype regulation might interact meaningfully with the effects described here. People whose values oppose stereotyping and prejudice may put effort into working against any learning that might reinforce those biases (Axt & Trawalter, 2017). Further, *awareness* of potential biases has been identified as an important aspect of bias reduction efforts (e.g., see Carter, Onyeador, & Lewis Jr, 2020; Devine et al., 2012), and it is possible that teaching people about how untested assumptions perpetuate stereotypes may equip people to counteract these processes.

8.2. Possible alternate mechanisms

Our predictions in the present article were derived from the theoretical model outlined in the introduction, built on neural mechanisms established by work in the cognitive psychology, machine learning, and cognitive neuroscience literatures. Although highly consistent with predictions derived from that model, we acknowledge that our results show behavioral patterns and therefore do not provide direct tests of the neural mechanisms discussed. We, however, are compelled by the extensive research conducted by cognitive neuroscientists that provide more direct tests that establish the core principles and neural mechanisms applied in our model (e.g., Antony et al., 2017; Berns, McClure, Pagnoni, & Montague, 2001; Fiorillo, Tobler, & Schultz, 2003; Frank et al., 2004; Gibson et al., 2013; Kalish et al., 2011; Markram et al., 1997; McCandliss et al., 2002; McClelland, 2006; McClelland et al., 2010; Tricomi et al., 2006; Tricomi & Fiez, 2008).

As helpfully noted by a reviewer, although our model describes emergentist patterns that arise from “low-level” neural learning mechanisms, our results may also be consistent with an account involving more reflective, “high-level” learning. Specifically, it is possible that participants, upon completing the training block, reflect on their response strategy, and modify their performance in the test block accordingly. By this account, participants in the No Feedback condition learn in the training block that relying on stereotypes makes the task easier, and thus they choose to rely more on stereotypes in the test block. Similarly, in the Stereotype-Confirming condition, participants may learn explicitly that stereotyping

provides better results in the training block, thus they consciously choose to rely more on stereotyping in the test block. The results of Study 3 are inconsistent with this alternate explanation, however. If people were consciously choosing to rely more on stereotyping, we would expect that memory errors would be driven by novel stereotyping more than by participants' prior assumptions. We observed the opposite pattern in Study 3; participants were more likely to misremember their own assumptions as confirmed, rather than just relying more on stereotypes.

Although we do not claim that our data can directly eliminate this alternate explanation, it is unclear the extent to which it can similarly account for the effects in the Stereotype-Disconfirming condition. It seems to us that such a reflective learning process should lead participants in that condition to stereotype less. But, but as noted above, there is no change in that condition, and participants in the Stereotype-Disconfirming condition are still stereotyping at a rate significantly higher than 50% chance in the test block: Study 1, $t(158) = 9.535$, $p < 0.001$; Study 2, $t(231) = 9.578$, $p < 0.001$. Another reviewer suggested the pattern in the Stereotype-Disconfirming condition could involve participants consciously disbelieving and discounting the feedback in this condition. We acknowledge that these alternate, more reflective processes could occur in addition to the basic neural learning processes that served as the foundation for our predictions. What is lacking in these alternate interpretations, however, is any formulation that specifies why Hebbian or hedonic learning would *not* occur under these circumstances.

Although we still favor explaining the present effects in terms of low-level learning, these potential alternative explanations — that people are consciously changing their performance strategy or disregarding disconfirmatory feedback — do not contradict the core take-away conclusions of the present paper. These alternatives still involve processes that favor stereotype perpetuation; they merely differ in the suggested mechanisms. Whichever interpretation you prefer, the net take-away messages of the present work remain the same. When people make stereotypic assumptions but receive no feedback on the veracity of their assumptions, it makes them more likely to stereotype over time. When people do receive feedback, disconfirmatory feedback has less influence than confirmatory feedback.

8.3. Implications for interpretation of other measures and findings

The present findings also invite further consideration of learning that may happen during tasks commonly used to measure stereotyping or other biases, such as the implicit association test (IAT). As people make judgments or categorizations on individual trials in such cognitive tasks, they may learn from those judgments, such that completing tasks might itself change their later performance (Cochrane, Cox, & Green, 2022; Kattner, Cochrane, Cox, Gorman, & Green, 2017; Kattner, Cochrane, & Green, 2017). Greater consideration of learning that occurs *during* reaction time or categorization tasks might reveal new phenomena that can advance our understanding of both these tasks themselves and the real-world phenomena they are operationalized to represent.

At first glance, the relative impotence of disconfirmatory evidence demonstrated in Studies 1 and 2 may seem at odds with an abundance of evidence showing that exposure to counter-stereotypic exemplars reduces automatic biases (e.g., Cao & Banaji, 2016; Cone

& Calanchini, 2021; Dasgupta & Greenwald, 2001). Dasgupta and Greenwald (2001), for instance, exposed participants to information about admired Black and disliked White targets (i.e., stereotype-disconfirmatory evidence), then observed decreases in measured implicit bias. This apparent contradiction in findings, however, highlights an important aspect of our model. Exposure to counter-stereotypic exemplars involves the impact of purely external evidence (akin to the External Hebbian Learning shown in Fig. 1), which is a distinct circumstance from the focus of the present work. Crucially, our theorizing relates to disconfirmatory evidence *following* a stereotypic assumption. This process involves two steps: people first make an inference (which strengthens the stereotypic association via inferential Hebbian learning, as in Fig. 2), and it is then subsequently disconfirmed (which initiates hedonic, feedback-based learning that must work against the inferential Hebbian learning, as in Fig. 3). The present evidence, therefore, does not contradict past work on exposure to counter-stereotypic exemplars. Instead, our model specifies different circumstances under which people may receive disconfirmatory evidence (i.e., purely external, passive exposure to evidence vs. an active inference that is subsequently contradicted), each with distinct implications for learning and stereotype reduction or perpetuation. Key to the present work is understanding people as social perceivers who *actively* make stereotypic inferences and predictions (Clark, 2013) that may be subsequently confirmed, disconfirmed, or remain untested.

8.4. Implications for other groups and contexts

The present experiments used a task involving stereotyping to infer sexual orientation because people often do not think of this process as “stereotyping”, which reduces social desirability concerns (Cox et al., 2016; Cox & Devine, 2014, 2015). The processes demonstrated here, however, should play out in consequential ways for many different target groups and stereotypes. Consider what happens when stereotypes influence an employer to not hire or even interview an applicant (Bertrand & Mullainathan, 2004; Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012). If employers in scientific fields fail to consider hiring a woman because they consciously or tacitly assume her gender makes her unskilled at math/science, they never receive information about her actual scientific abilities (Moss-Racusin et al., 2012). If, as in the present studies, they learn from these untested assumptions, or falsely remember them as true, they may be even less likely to consider the next female applicant, perpetuating a recursive, pernicious pattern. Further, the imbalance in weighting of confirmatory and disconfirmatory evidence means that it will take several women who disconfirm stereotypes to counterbalance each woman who happens to match the stereotype. Assumptions that a Black person is dangerous, a Latin American person is an undocumented immigrant, or a Muslim person is a terrorist will very often, perhaps most often, remain untested, and thereby strengthen stereotypic associations in the absence of evidence. This learning process is likely a major contributor to the persistence and perpetuation of stereotypes and biases related to any target group.

One interesting possibility of learning in the absence of evidence is that it seems to imply that extinction will not occur. Although we would not go so far as to argue that extinction is impossible, when considering phenomena such as conspiracy theories, superstitions, the anti-vaccination movement, or other recent work on “fake news”, it seems possible, even

likely, that learning in the absence of evidence may indeed go unchecked. Once people have a preexisting belief or association that they use as a basis for inference, without another process operating to change it, the association may never extinguish — in the presence of evidence, confirmation biases will perpetuate it, and in the absence of evidence, it will self-perpetuate.

Indeed, it may be useful to apply insights from this cognitive process in other domains involving inferences that may remain untested, such as depression/anxiety (e.g., a person thinks, “no one likes me” and then fails to engage in social activities in which they could find out; Cox et al., 2012) and superstitions/conspiracy theories (e.g., “the more I think about this, the more it seems true!”). Learning from untested assumptions may also be useful to consider in work related to the recent proliferation of “fake news” (Gilbert, Tafarodi, & Malone, 1993; Schwarz, Newman, & Leach, 2016). Recently discussed issues related to scientific research practices and replicability may also arise in part due to these processes. Scientists may fail to properly weigh evidence that opposes their theories, and time spent reflecting on and developing theories may further entrench those theories in the scientists’ minds, without any new data (see also McClelland, 2006).

8.5. Additional consideration of hedonic learning mechanisms and stereotyping

Our model of the learning processes at play during stereotyping may lead to additional insights into stereotype persistence. One particular untapped area of inquiry is examining the role of the brain’s reward system in stereotyping. As noted earlier, learning from confirmatory and disconfirmatory feedback involves hedonic learning (Frank et al., 2004; McCandliss et al., 2002; McClelland, 2006; Reggev et al., 2021; Tricomi et al., 2006; Tricomi & Fiez, 2008). By definition, stereotypes operate with a wide degree of probabilistic uncertainty: Some people will confirm stereotypic expectations, and others will disconfirm them. It is in such cases of probabilistic uncertainty that reward-related striatal activity is strongest (Berns et al., 2001; Fiorillo et al., 2003). A stereotypic assumption is an inferential gamble about an individual, and it is rewarding when such gambles pay off, and aversive when they fail (Berns et al., 2001; Fiorillo et al., 2003; Frank et al., 2004; Garrison, Erdeniz, & Done, 2013).

Importantly for the present work, accuracy was not extrinsically incentivized — participants received the same external compensation for their participation no matter how they performed on the stereotyping categorization task. This study design choice was intentional. Extrinsic rewards (e.g., monetary compensation for accuracy) interact/interfere with the learning effects of intrinsic rewards processes. A core part of our model is built on the fact that stereotype confirmation is itself rewarding. This neural reward process intrinsically incentivizes stereotyping, whereas a performance incentive would extrinsically incentivize accuracy. Recent work by Reggev et al. (2021) suggests that stereotype confirmation is in fact *more* rewarding than monetary incentives. In their work, participants were given a choice between seeing a stimulus person who confirmed stereotypes or receiving more money. The overall pattern showed that participants preferred to see stereotype confirmation, even if it meant getting less money. In the present work, absent any external monetary incentive, the pattern in the stereotype–disconfirming condition of Studies 1 and 2 is highly

consistent with the interpretation that it is more satisfying to have a stereotypic assumption confirmed than to merely be correct. Participants' response patterns in that condition clearly favored stereotyping over accuracy.

In contrast to older theoretical models that conceptualize stereotyping as a largely non-affective, passive cognitive process, our model highlights stereotyping as an active predictive process, which inherently brings in affective components, in the form of these hedonic learning processes. Stereotypes facilitate perceptual fluency, which itself is affectively pleasing, bearing hedonic markers that elicit positive affect and aesthetic pleasure (Mendes, Blascovich, Hunter, Lickel, & Jost, 2007; Reber, Schwarz, & Winkielman, 2004). Providing more direct evidence that stereotype confirmation is intrinsically rewarding, the aforementioned work by Reggev et al. (2021) included neuroimaging studies showing increased activation in reward-related brain areas when participants viewed stimulus people who confirmed stereotypes. The counterpart to stereotype confirmation's reward value is that stereotype disconfirmation is aversive, which should lead people to avoid or resist opportunities for stereotype disconfirmation.

Reward/aversion learning are inherently linked to psychophysiological approach/avoidance processes. In fact, Mendes and her colleagues (Mendes et al., 2007) assessed psychophysiological indicators of approach/avoidance (i.e., ventricular contractility, cardiac output, and total peripheral resistance) in participants who interacted with a stereotype-confirming or stereotype-disconfirming person. Matching predictions from the present approach, participants interacting with a stereotype-confirming target showed increases on the physiological indicators of approach (and decreases on avoidance). Likewise, when participants interacted with a stereotype-disconfirming target, their pattern was reversed, showing increases on psychophysiological avoidance (and decreases in approach). The effects of these low-level hedonic processes, importantly, could be fully conscious or fully non-conscious. In either case, they would support the perpetuation of stereotypes in peoples' minds, judgments, and behavior. The involvement of these hedonic processes in stereotyping means that, at a fundamental level people should "like" to stereotype and even be compelled to stereotype, and also avoid or resist opportunities for stereotype disconfirmation.

There will always be some individuals who happen to match their group's stereotypes — women who prefer family over career, gay men who enjoy shopping, Black men who enjoy basketball. The existence of even a few such confirmatory cases means that stereotyping essentially operates with an intermittent reinforcement schedule, the most difficult learning pattern to extinguish (Reber et al., 2004). Like other bad habits (Devine, 1989), stereotyping is compelled by low-level hedonic processes, and greater attention to the role of these processes during stereotyping may be important to developing effective methods to combat stereotyping, prejudice, and discrimination.

9. Conclusion

The present work demonstrates that stereotyping, in essence, is self-perpetuating. These patterns further emphasize the necessary role of effortful, controlled processes in reducing stereotyping and biases (Cox et al., 2012; Cox & Devine, 2019; Devine, 1989; Devine et

al., 2012). Stereotypes, misinformation, and misconceptions spread in people's minds and in culture, leading many people to treat opinions, hunches, and tacit inferences as equivalent to, or even more important than, systematic data and scientific evidence. It is intuitive to think that the cure for misinformation is giving people more correct information, but, as the present work highlights, human learning processes are often "fact agnostic", favoring entrenchment and perpetuating preexisting associations that may be either correct or incorrect from an objective standpoint (Lord et al., 1979). Rather than adopting information deficit approaches that try to combat misinformation with additional facts (Cox, 2022; Suldovsky, 2017), perhaps what is needed is to raise awareness about people's potential for biased processing (Cox et al., 2012; Cox & Devine, 2019; Devine, 1989; Devine et al., 2012) and to change the way the public evaluates and consumes information. As scientists, we rely on the scientific method as a system of checks and balances to mitigate the influence of our subjective human biases and irrational learning processes on the acquisition of knowledge. If we can guide the public, as social perceivers and consumers of information, to hold their own minds to similar standards of evidence, perhaps we can enlist their conscious, controlled mental processes to mitigate the automatic learning processes that perpetuate stereotypes and stereotyping.

Open Practices

We reported all measures, manipulations, and exclusions. Procedures for all three studies were approved by the University of Wisconsin- Madison IRB. All experimental materials and data files will be available online at www.sciencecox.com. The hypotheses reported in-text were specified before any data collection began, as was the primary analytic approach (i.e., the 3 × 2 Mixed ANOVAs). After completion of Study 1, the hypotheses and design of Studies 2 and 3 were further documented in the first authors' grant submission. We have no "file drawer" studies testing these hypotheses that have been excluded from this paper.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The research in the present manuscript was supported by MIRA 1 R35 GM128888 from NIGMS at the NIH, awarded to W. T. L. Cox. Preparation of this article was additionally supported by a Wisconsin Alumni Research Foundation Professorship and MIRA 1 R35 GM127043-01 awarded to P. G. Devine. We would like to thank Eric Roman Beining, the graphic designer who crafted Figs. 1–5 and 7. We thank Adrienne Wood, Adam Safron, Chelsea Mitamura, Katharine Scott, and Kristina Kellett for comments this article. We also thank our colleagues Mark Seidenberg, Maryellen MacDonald, Jessica Montag, Chris Cox, Chuck Kalish, Lyn Abramson, Curtis Ryals, and Markus Brauer discussing this project with us. Special gratitude goes to Erica Nagy, Melanie Davies, Hannah Evans, and Lauryn Besasie for their tireless work developing the stimuli, as well as our amazing research assistants who conducted data collection and assisted in manuscript preparation, Nancy Lundin, Molly Maier, Amy Petermann, Shayla Krecklow, Alish Ridge-Montague, Kendra Lange, Mariah Apollon, Olivia Prager, Katie Platt, Mary Lu, Martha Morganstein, Ash Lyke, Saja Abu Hakmeh, Morgan Kubicek, Alex Mischler, Imani Wilson, Anna Kons, Chloe Wigul, Jack Berroug, Makena Meyers, Madison Thornton, Sophie Anderson, Noel Farmer, Gabby Janovsky, Parker Rosemeyer, Echo Fatsis, and Shannon Carnahan.

References

Allport GW, Clark K, & Pettigrew T (1954). The nature of prejudice. Reading MA: Addison- Wesley.

Antony JW, Ferreira CS, Norman KA, & Wimber M (2017). Retrieval as a fast route to memory consolidation. Trends in Cognitive Sciences, 21(8), 573–576. 10.1016/j.tics.2017.05.001 [PubMed: 28583416]

Axt J, & Trawalter S (2017). Whites demonstrate anti-black associations but do not reinforce them. Journal of Experimental Social Psychology, 70, 8–18.

Bem DJ (1972). Self-perception theory. Advances in Experimental Social Psychology, 6 (1), 1–62. 10.1016/S0065-2601(08)60024-6

Berns GS, McClure SM, Pagnoni G, & Montague PR (2001). Predictability modulates human brain response to reward. Journal of Neuroscience, 21(8), 2793–2798. 10.1523/JNEUROSCI.21-08-02793.2001 [PubMed: 11306631]

Bertrand M, & Mullainathan S (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. American Economic Review, 94(4), 991–1013. 10.1257/0002828042002561

Blanchard FA, Lilly T, & Vaughn LA (1991). Reducing the expression of racial prejudice. Psychological Science, 2(2), 101–105. 10.1111/j.1467-9280.1991.tb00108.x

Burns MD, Monteith MJ, & Parker LR (2017). Training away bias: The differential effects of counterstereotype training and self-regulation on stereotype activation and application. Journal of Experimental Social Psychology, 73, 97–110.

Cao J, & Banaji MR (2016). The base rate principle and the fairness principle in social judgment. Proceedings of the National Academy of Sciences, 113(27), 7475–7480.

Carter ER, Onyeador IN, & Lewis NA Jr. (2020). Developing & delivering effective anti-bias training: Challenges & recommendations. Behavioral Science & Policy, 6(1), 57–70.

Clark A (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36, 181–253. 10.1017/S0140525X12000477 [PubMed: 23663408]

Cochrane A, Cox WTL, & Green CS (2022). The Implicit Association Test is sensitive to learning: Exploring within-session modulations of IAT scores (Under review).

Cone J, & Calanchini J (2021). A process dissociation model of implicit rapid revision in response to diagnostic revelations. Personality and Social Psychology Bulletin, 47(2), 201–215. [PubMed: 32478605]

Cox WT (2015). Multiple determinants of prejudicial and nonprejudicial behavior. doctoral dissertation. The University of Wisconsin-Madison.

Cox WTL (2022). Developing scientifically validated bias and diversity trainings that work: Empowering agents of change to reduce bias, create inclusion, and promote equity. Management Decision. In Press.

Cox WTL, Abramson LY, Devine PG, & Hollon SD (2012). Stereotypes, prejudice, and depression: The integrated perspective. Perspectives on Psychological Science, 7(5), 427–449. 10.1177/1745691612455204 [PubMed: 26168502]

Cox WTL, & Devine PG (2014). Stereotyping to infer group membership creates plausible deniability for prejudice-based aggression. Psychological Science, 25(2), 340–348. 10.1177/0956797613501171 [PubMed: 24335602]

Cox WTL, & Devine PG (2015). Stereotypes possess heterogeneous directionality: A theoretical and empirical exploration of stereotype structure and content. PLoS One, 10(3). 10.1371/journal.pone.0122292

Cox WTL, & Devine PG (2019). The prejudice habit-breaking intervention: An empowerment-based confrontation approach. In Mallett RK, & Monteith MJ (Eds.), Confronting prejudice and discrimination: The science of changing minds and behaviors (pp. 249–274). London, UK: Academic Press.

Cox WTL, Devine PG, Bischmann AA, & Hyde JS (2016). Inferences about sexual orientation: The roles of stereotypes, faces, and the gaydar myth. The Journal of Sex Research, 53(2), 157–171. 10.1080/00224499.2015.1015714 [PubMed: 26219212]

Darley JM, & Gross PH (2004). A hypothesis-confirming bias in labeling effects. In Social Cognition (pp. 438–450). Psychology Press.

Dasgupta N, & Greenwald AG (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. Journal of Personality and Social Psychology, 81(5), 800. [PubMed: 11708558]

Devine PG (1989). Stereotypes and prejudice: Their automatic and controlled components. Journal of personality and social psychology, 56(1), 5–18.

Devine PG, Forscher PS, Austin AJ, & Cox WTL (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. Journal of Experimental Social Psychology, 48(6), 1267–1278. 10.1016/j.jesp.2012.06.003 [PubMed: 23524616]

Devine PG, Plant EA, Amodio DM, Harmon-Jones E, & Vance SL (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. Journal of Personality and Social Psychology, 82(5), 835. 10.1037/0022-3514.82.5.835 [PubMed: 12003481]

Devine PG, & Sherman SJ (1992). Intuitive versus rational judgment and the role of stereotyping in the human condition: Kirk or Spock? Psychological Inquiry, 3(2), 153–159. 10.1207/s15327965pli0302_13

Dunning D, & Sherman DA (1997). Stereotypes and tacit inference. Journal of Personality and Social Psychology, 73(3), 459. 10.1037/0022-3514.73.3.459 [PubMed: 9294897]

Faul F, Erdfelder E, Buchner A, & Lang AG (2009). Statistical power analyses using G* power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41(4), 1149–1160. 10.3758/BRM.41.4.1149 [PubMed: 19897823]

Fiorillo CD, Tobler PN, & Schultz W (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. Science, 299(5614), 1898–1902. 10.1126/science.1077349 [PubMed: 12649484]

Fiske ST (1998). Stereotyping, prejudice, and discrimination. In The handbook of social psychology (pp. 357–411). McGraw-Hill.

Frank MJ, Seeberger LC, & O'reilly RC (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. Science, 306(5703), 1940–1943. 10.1126/science.1102941 [PubMed: 15528409]

Garrison J, Erdeniz B, & Done J (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. Neuroscience & Biobehavioral Reviews, 37(7), 1297–1310. 10.1016/j.neubiorev.2013.03.023 [PubMed: 23567522]

Gibson BR, Rogers TT, & Zhu X (2013). Human semi-supervised learning. Topics in Cognitive Science, 5(1), 132–172. 10.1111/tops.12010 [PubMed: 23335577]

Gilbert DT (1991). How mental systems believe. American Psychologist, 46(2), 107. 10.1037/0003-066X.46.2.107

Gilbert DT, Tafarodi RW, & Malone PS (1993). You can't not believe everything you read. Journal of Personality and Social Psychology, 65(2), 221. 10.1037/0022-3514.65.2.221 [PubMed: 8366418]

Hamilton DL, Sherman JW, Wyer RS, & Srull TK (1994). Handbook of social cognition (pp. 2–68). Hillsdale: Lawrence Erlbaum Associates.

Hebb DO (1949). The organization of behavior: A neuropsychological approach. New York, NY: Wiley.

Hilton JL, & Von Hippel W (1996). Stereotypes. Annual Review of Psychology, 47(1), 237–271. 10.1146/annurev.psych.47.1.237

JASP. (2020). JASP Team.

Kalish CW, Rogers TT, Lang J, & Zhu X (2011). Can semi-supervised learning explain incorrect beliefs about categories? Cognition, 120(1), 106–118. [PubMed: 21474122]

Kattner F, Cochrane A, Cox CR, Gorman TE, & Green CS (2017). Perceptual learning generalization from sequential perceptual training as a change in learning rate. Current Biology, 27(6), 840–846. [PubMed: 28262488]

Kattner F, Cochrane A, & Green CS (2017). Trial-dependent psychometric functions accounting for perceptual learning in 2-AFC discrimination tasks. Journal of Vision, 17(11), 3.

Kunda Z (1990). The case for motivated reasoning. Psychological Bulletin, 108(3), 480. 10.1037/0033-2909.108.3.480 [PubMed: 2270237]

Lee MD, & Wagenmakers EJ (2013). Bayesian cognitive modeling: A practical course. Cambridge University Press.

Lord CG, Ross L, & Lepper MR (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. Journal of Personality and Social Psychology, 37(11), 2098. 10.1037/0022-3514.37.11.2098

Ly A, Etz A, Marsman M, & Wagenmakers EJ (2019). Replication Bayes factors from evidence updating. Behavior Research Methods, 51(6), 2498–2508. 10.3758/s13428-018-1092-x [PubMed: 30105445]

Markram H, Lübke J, Frotscher M, & Sakmann B (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science, 275(5297), 213–215. 10.1126/science.275.5297.213 [PubMed: 8985014]

McCandliss BD, Fiez JA, Protopapas A, Conway M, & McClelland JL (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. Cognitive, Affective, & Behavioral Neuroscience, 2(2), 89–108. 10.3758/cabn.2.2.89

McClelland J, Botvinick M, Noelle D, Plaut D, Rogers T, Seidenberg M, et al. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. Trends in Cognitive Sciences, 14(8), 348–356. [PubMed: 20598626]

McClelland JL (2006). How far can you go with Hebbian learning, and when does it lead you astray? Processes of a change in brain and cognition development: Attention and performance XXI, 21, 36–69.

Mendes W, Blascovich J, Hunter S, Lickel B, & Jost J (2007). Threatened by the unexpected: Physiological responses during social interactions with expectancy-violating partners. Journal of Personality and Social Psychology, 92(4), 698–716. 10.1037/0022-3514.92.4.698 [PubMed: 17469953]

Montague PR, Dayan P, & Sejnowski TJ (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. Journal of Neuroscience, 16 (5), 1936–1947. [PubMed: 8774460]

Monteith MJ, Ashburn-Nardo L, Voils CI, & Czopp AM (2002). Putting the brakes on prejudice: On the development and operation of cues for control. Journal of Personality and Social Psychology, 83, 1029–1050. [PubMed: 12416910]

Monteith MJ, Deneen NE, & Tooman GD (1996). The effect of social norm activation on the expression of opinions concerning gay men and blacks. Basic and Applied Social Psychology, 18(3), 267–288. 10.1207/s15324834basp1803_2

Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, & Handelsman J (2012). Science faculty's subtle gender biases favor male students. Proceedings of the National Academy of Sciences, 109(41), 16474–16479. 10.1073/pnas.1211286109

Nickerson RS (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175–220. 10.1037/1089-2680.2.2.175

Oliver W (2003). The structural-cultural perspective: A theory of black male violence. Violent crime: Assessing race and ethnic differences (pp. 280–302).

Palmeri TJ, & Flanery MA (1999). Learning about categories in the absence of training: Profound amnesia and the relationship between perceptual categorization and recognition memory. Psychological Science, 10, 526–530.

Palmeri TJ, & Flanery MA (2002). Memory systems and perceptual categorization. In Ross BH (Ed.), *Vol.* 41. The psychology of learning and motivation: Advances in research and theory (pp. 141–189). San Diego: Academic Press.

Plant EA, & Devine PG (1998). Internal and external motivation to respond without prejudice. Journal of Personality and Social Psychology, 75(3), 811.

Pohl R, Pohl RF, & editors.. (2004). Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory. Psychology Press.

Pratto F, Sidanius J, Stallworth LM, & Malle BF (1994). Social dominance orientation: A personality variable predicting social and political attitudes. Journal of Personality and Social Psychology, 67(4), 741. 10.1037/0022-3514.67.4.741

Reber R, Schwarz N, & Winkielman P (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? Personality and Social Psychology Review, 8(4), 364–382. 10.1207/s15327957pspr0804_3 [PubMed: 15582859]

Reggev N, Chowdhary A, & Mitchell JP (2021). Confirmation of interpersonal expectations is intrinsically rewarding. Social Cognitive and Affective Neuroscience, 16 (12), 1276–1287. [PubMed: 34167150]

Schultz W (2016). Dopamine reward prediction-error signalling: A two-component response. Nature Reviews Neuroscience, 17(3), 183–195. [PubMed: 26865020]

Schwarz N, Newman E, & Leach W (2016). Making the truth stick & the myths fade: Lessons from cognitive psychology. Behavioral Science & Policy, 2(1), 85–95.

Steele CM (2011). Whistling Vivaldi: How stereotypes affect us and what we can do. WW Norton & Company.

Sue DW, Capodilupo CM, Torino GC, Bucceri JM, Holder A, Nadal KL, & Esquilin M (2007). Racial microaggressions in everyday life: Implications for clinical practice. American Psychologist, 62(4), 271. [PubMed: 17516773]

Suldovsky B (2017). The information deficit model and climate change communication. In Oxford Research Encyclopedia of Climate Science.

Tricomi E, Delgado MR, McCandliss BD, McClelland JL, & Fiez JA (2006). Performance feedback drives caudate activation in a phonological learning task. Journal of Cognitive Neuroscience, 18(6), 1029–1043. 10.1162/jocn.2006.18.6.1029 [PubMed: 16839308]

Tricomi E, & Fiez JA (2008). Feedback signals in the caudate reflect goal achievement on a declarative memory task. Neuroimage, 41(3), 1154–1167. 10.1016/j.neuroimage.2008.02.066 [PubMed: 18445531]

Uhlmann EL, & Cohen GL (2007). "I think it, therefore it's true": Effects of self-perceived objectivity on hiring discrimination. Organizational Behavior and Human Decision Processes, 104(2), 207–223. 10.1016/j.obhdp.2007.07.001

Wagenmakers EJ, Marsman M, Jamil T, Ly A, Verhagen J, Love J, … Matzke D (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. Psychonomic Bulletin & Review, 25(1), 35–57. 10.3758/s13423-017-1343-3 [PubMed: 28779455]

Zaki SR, & Nosofsky RM (2007). A high-distortion enhancement effect in the prototype-learning paradigm: Dramatic effects of category learning during test. Memory & Cognition, 35, 2088–2096. [PubMed: 18265623]

Zhu X, & Goldberg AB (2009). Introduction to semi-supervised learning. In, 3(1). Synthesis lectures on artificial intelligence and machine learning (pp. 1–130). Morgan & Claypool.

Zhu X, Rogers T, Qian R, & Kalish C (2007). Humans perform semi-supervised classification too. In Holte RC, & Howe A (Eds.), Proceedings of the 21st conference on artificial intelligence (AAAI-11) (pp. 864–870). Menlo Park, CA: The AAAI Press.

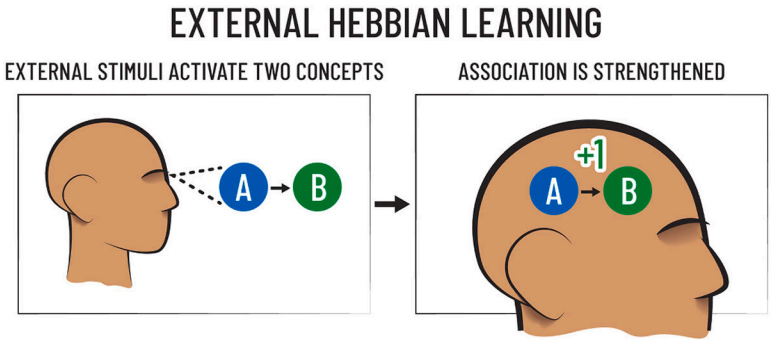

Fig. 1.

External Hebbian Learning. Two concepts co-occur as stimuli (e.g. a Black person on television is portrayed as a criminal), which builds or strengthens the association between those concepts (*Black* → *Criminal*).

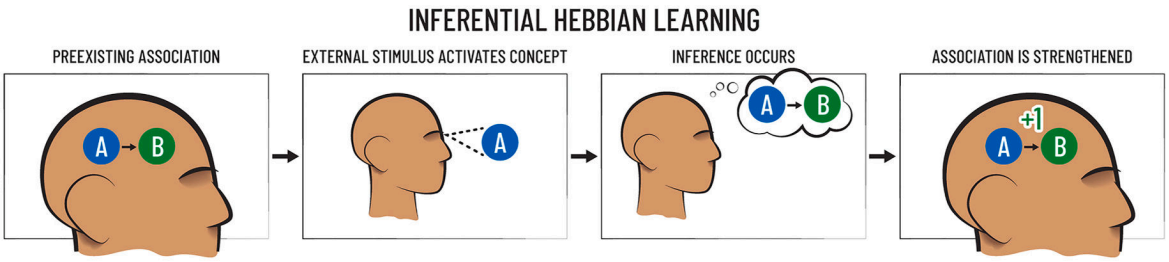

Fig. 2.

Inferential Hebbian Learning. One concept occurs as a stimulus, and activates a pre-existing association, leading to an inference (e.g., someone sees a Black person and makes the stereotypic assumption that they are a criminal). Hebbian learning still occurs, with the internally-generated activation of the concepts strengthening the association between them (*Black* → *Criminal*), making the same inference more likely in the future.

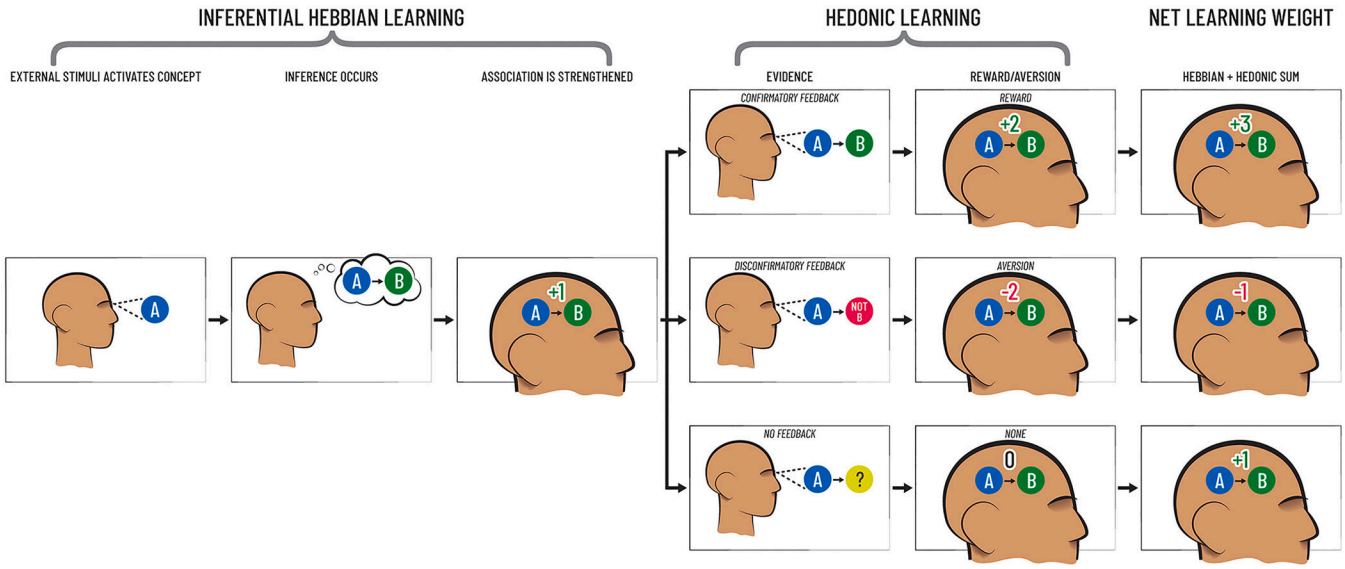

Fig. 3.

Learning Patterns for Three Possible Outcomes of a Stereotypic Inference. If someone makes a stereotypic inference and never learns whether it was correct, there will be no hedonic learning, but the inferential Hebbian learning will have the net effect of strengthening the association — people learn from their untested assumptions. If someone makes a stereotypic inference and then learns that it is correct or incorrect, their initial inference will strengthen the association, and then the positive or negative feedback provides reward or aversion. The aversion learning weight must counteract the inferential Hebbian learning weight, whereas the reward enhances it. This pattern of weighting creates an imbalance between confirmatory and disconfirmatory evidence (e.g., as seen in confirmation bias). *Note:* The numerical values assigned to the learning weights are placeholders to allow comparison across the different hypothesized learning patterns. They are not intended to be precise numerical indicators.

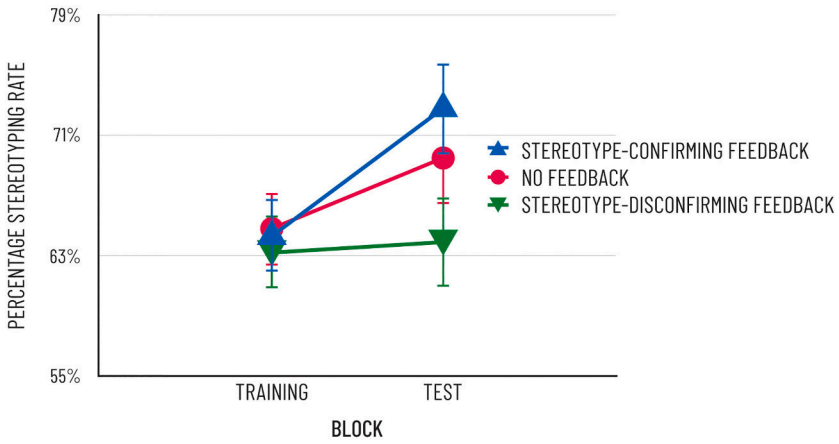

Fig. 4.

Study 1 Stereotyping Rates and Learning Patterns. Error bars represent 95% CIs around the estimated marginal means. As hypothesized, people relied more heavily on confirmatory evidence than disconfirmatory evidence as they learned in the experimental task. When people received no evidence about the accuracy of their judgments, their stereotyping rates nevertheless increased, suggesting that they learned from their untested assumptions, consistent with our theorizing. Stereotyping rates in the Test block of the No Feedback condition did not statistically differ from those of the Stereotype-Confirming condition, $p = 0.126$, but the Stereotype-Disconfirming condition's Test block rates were significantly lower than both the No Feedback condition and the Stereotype-Confirming condition, p's < 0.011.

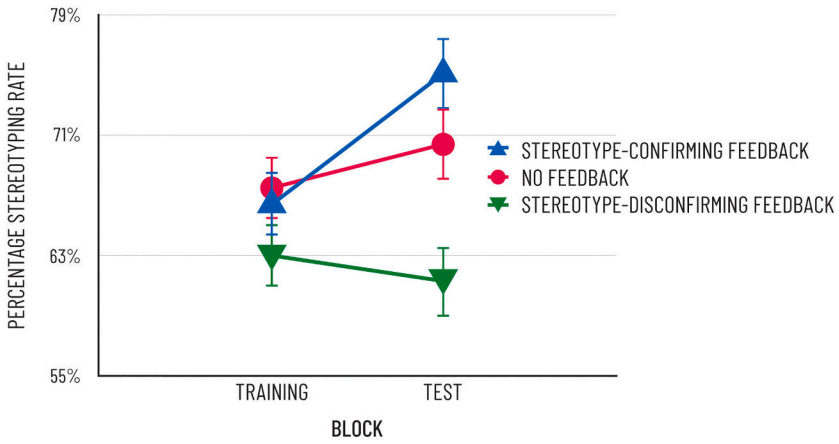

Fig. 5.

Study 2 Stereotyping Rates and Learning Patterns. Error bars represent 95% CIs around the estimated marginal means. Replicating Study 1, confirmatory evidence had a greater influence than disconfirmatory evidence, and participants in the No Feedback condition stereotyped more over time, despite their assumptions remaining untested. Test rates in the Stereotype-Confirming condition were significantly higher than test rates of the No Feedback condition, $p = 0.004$, and the Stereotype-Disconfirming condition, $p < 0.001$. The test rates of the No Feedback condition were significantly higher than those of the Stereotype-Disconfirming condition, $p < 0.001$.

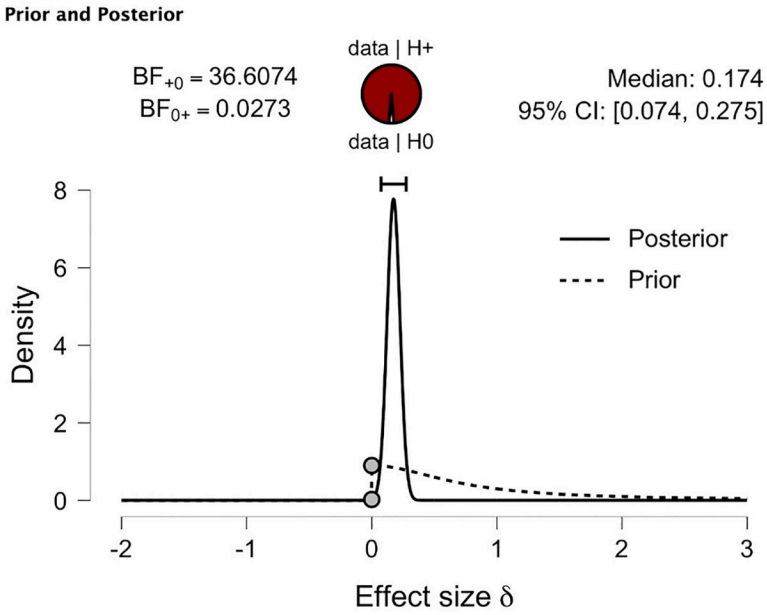

Fig. 6.

Complete Bayes Factor for No Feedback Conditions. The estimated complete Bayes Factor $BF_{+0}(d_{orig}, d_{rep}) = 36.61$, indicates that together, Study 1 and 2's data in the No Feedback condition are predicted by the alternative hypothesis 36.61 times better than the null hypothesis. Following Lee and Wagenmakers' (2013) rule of thumb, this level of evidence is considered *very strong*.

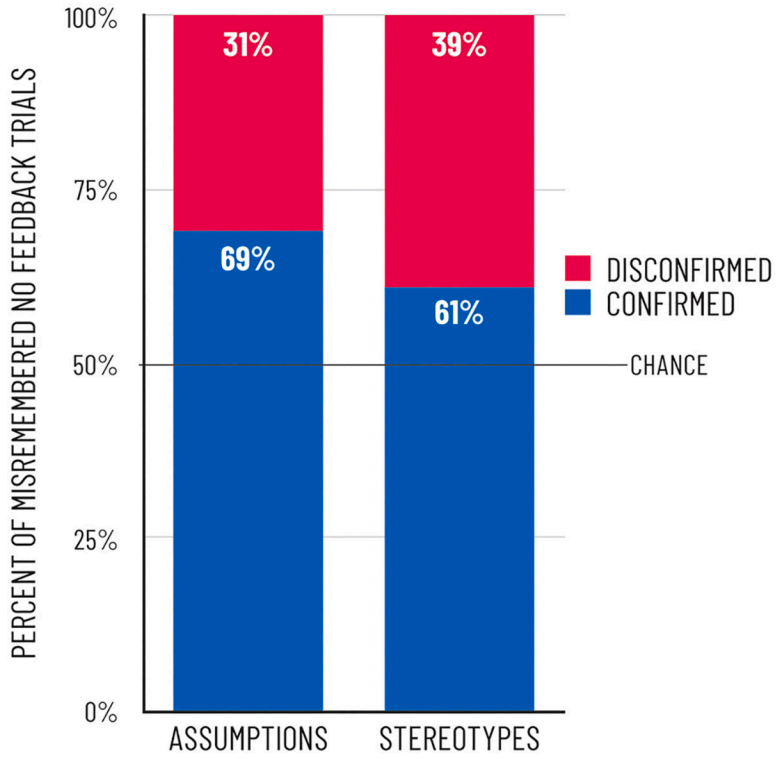

Fig. 7.

False Memories for No Feedback Trials. Participants incorrectly remembered their untested assumptions as being confirmed at a rate higher than chance, and higher than their rate for misremembering the No Feedback trials as confirming stereotypes.

Table 1

Study 1 descriptive statistics.

Feedback Condition		*M*	*sd*	Skew	Kurtosis
Stereotype-Confirming	Training	0.64	(0.145)	0.128	−0.495
	Test	0.73	(0.172)	−0.305	−0.394
No Feedback	Training	0.65	(0.161)	−0.018	−0.606
	Test	0.69	(0.203)	−0.411	−0.580
Stereotype-Disconfirming	Training	0.63	(0.012)	−0.301	−0.211
	Test	0.64	(0.184)	−0.107	−0.548

Stereotyping rates presented as proportions.

Table 2

Study 1 Inferential Statistics.

Effect	*df*	*F*	*p*	η_p^2
Condition	(2,466)	6.00	0.003	0.025
Block	(1,466)	21.67	<0.001	0.044
Condition × Block	(2,466)	5.17	0.006	0.022
Within-Subjects	***df***	***F***	***p***	η_p^2
Stereotype-Confirming	(1,155)	26.97	<0.001	0.148
No Feedback	(1,153)	7.06	0.009	0.044
Stereotype-Disconfirming	(1,158)	0.15	0.702	0.001

The upper portion of the table reports output from the key hypothesis test, which was a 3 (Feedback Condition: Stereotype-Confirming vs. Stereotype-Disconfirming vs. No Feedback) × 2 (Block: Training vs. Test) mixed ANOVA. The lower portion of the table reports output from separate repeated-measures ANOVAs for each condition.

Table 3

Study 2 descriptive statistics.

Feedback Condition		*M*	*sd*	Skew	Kurtosis
Stereotype-Confirming	Training	0.66	0.152	−0.094	−0.667
	Test	0.75	0.178	−0.499	−0.355
No Feedback	Training	0.67	(0.159)	−0.187	−0.445
	Test	0.70	(0.167)	−0.172	−0.400
Stereotype-Disconfirming	Training	0.63	0.149	−0.370	−0.127
	Test	0.61	0.179	−0.450	0.467

Stereotyping rates presented as proportions.

Table 4

Study 2 Inferential Statistics.

Effect	*df*	*F*	*p*	η_p^2
Condition	(2,679)	28.64	<0.001	0.078
Block	(1,679)	17.68	<0.001	0.025
Condition × Block	(2,679)	14.93	<0.001	0.042
Within-Subjects	*df*	*F*	*p*	η_p^2
Stereotype-Confirming	(1,221)	40.27	<0.001	0.154
No Feedback	(1, 227)	4.94	0.027	0.021
Stereotype-Disconfirming	(1,231)	1.65	0.201	0.007

The upper portion of the table reports output from the key hypothesis test, which was a 3 (Condition: Stereotype-Confirming vs. Stereotype-Disconfirming vs. No Feedback) × 2 (Block: Training vs. Test) mixed ANOVA. The lower portion of the table reports output from separate repeated-measures ANOVAs for each condition.

Table 5

Study 3 analyses using count data.

	Count Analyses for Misremembered No Feedback Trials				
	Generalized Linear Model with Poisson Family				Bayesian Paired Samples T-Test with Cauchy Prior of 0.707
Comparison	***Beta***	***b***	***p***	***IRR***	**BF_{10}**
Confirming vs Disconfirming Assumptions	0.17	0.78	<0.001	2.179	>999
Confirming vs Disconfirming Stereotypes	0.09	0.37	<0.001	1.451	>999
Confirming Assumptions vs Confirming Stereotypes	0.11	0.52	<0.001	1.680	>999