Journal of Experimental Psychology: General

Avoidance Begets Avoidance: A Computational Account of Negative Stereotype Persistence

Suraiya Allidina and William A. Cunningham

Online First Publication, August 30, 2021. http://dx.doi.org/10.1037/xge0001037

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

https://doi.org/10.1037/xge0001037

Avoidance Begets Avoidance: A Computational Account of Negative Stereotype Persistence

Suraiya Allidina and William A. Cunningham
Department of Psychology, University of Toronto, St. George Campus

Research on stereotype formation has proposed a variety of reasons for how inaccurate stereotypes arise, focusing largely on accounts of motivation and cognitive efficiency. Here, we instead consider how stereotypes arise from basic processes of approach and avoidance in social learning. Across five studies, we show that initial negative interactions with some members of a group can cause subsequent avoidance of the entire group, and that this avoidance perpetuates stereotypes in two ways. First, when information gain is contingent on approaching the target, avoidance restricts the information available with which to update one's beliefs. Second, computational models that consider the perceiver's full reinforcement history demonstrate that avoidance directly reinforces itself, such that initial avoidance of group members increases the probability of later acts of avoidance toward that group. Finally, we find initial evidence for a potential dissociation between behavior and explicit beliefs, with avoidance reinforcing avoidant behaviors without necessarily affecting self-reported beliefs. Overall, these results suggest that avoidance behaviors toward members of social groups can perpetuate inaccurate negative beliefs and expectations about those groups, such that initial interactions with a group have a compounding effect on overall impressions.

Keywords: avoidance, stereotypes, reinforcement learning, social learning

Supplemental materials: https://doi.org/10.1037/xge0001037.supp

A central question in understanding prejudice and inequality is why well-intentioned people who value egalitarianism still behave in discriminatory ways. Previous work attempting to address this question has focused largely on motivational factors such as drives to uplift one's own group at the expense of others (Brewer, 1999; Kunda & Spencer, 2003; Tajfel & Turner, 1979). Here, we instead suggest that these biases can arise even in the absence of these motivational factors as a result of the basic structure of social learning. In particular, we examine two ways in which basic features of the learning process can give rise to inaccurate stereotypes and biased behaviors. First, individuals are active agents in their own learning, choosing who to approach and gather information from and who to instead avoid. Building on work by Fazio and colleagues (Fazio et al., 2004), we propose that this active sampling of information can shape an individual's experiences in a manner that prevents existing beliefs from being updated (which we refer to as ignorance-based avoidance). Thus, while an individual's beliefs about a group may accurately reflect their experiences, their experiences themselves may be a poor representation of reality, producing biased beliefs and behaviors. Second, drawing from the literature on fear and anxiety learning (e.g., LeDoux et al., 2017), we examine how avoidance behaviors can have self-reinforcing effects, such that the decision to avoid someone now can increase the probability of avoiding that person later on (referred to here as self-reinforced avoidance). Testing such a mechanism requires consideration of a participant's full reinforcement and choice history with a given group, which we accomplish through the use of computational models. Together, ignorance-based avoidance and self-reinforced avoidance may be two mechanisms through which simple processes of approach and avoidance in social information-gathering can lead to the maintenance of inaccurate beliefs about social groups. Thus, echoing early theories on social categorization (Hamilton, 1979), we propose that fundamental principles of cognition and learning can give rise to pervasive social biases when applied in the context of the complex social world.

Ignorance Effects: Avoidance Prevents Information Gain

A key way in which people learn about others is through direct interaction, which can provide information about what future interactions with that person will be like (Behrens et al., 2009; Hackel et

Suraiya Allidina https://orcid.org/0000-0001-6463-8415

We thank Cendri Hutcherson and the members of the Social Cognitive Science lab for helpful comments on earlier versions of this article. We thank Cendri Hutcherson, Russ Fazio, and the members of the Social Cognitive Science lab for helpful comments on earlier versions of this article. We also thank Rakshita Kathuria and Alice Wang for their help with stimuli creation and data collection. This work was funded by Social Sciences and Humanities Research Council (SSHRC) Grant 506547 to William A. Cunningham and a SSHRC Doctoral Fellowship to Suraiya Allidina. Parts of this article were presented at the 2019 meeting of the Society for Personality and Social Psychology and the 2019 meeting of the Canadian Society for Brain, Behavior, and Cognitive Science. We have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Suraiya Allidina or William A. Cunningham, Department of Psychology, University of Toronto, St. George Campus, 100 St. George Street, Toronto, ON M5S 3G3, Canada. Email: suraiya.allidina@mail.utoronto.ca or cunningham@psych.utoronto.ca

al., 2015; Jones et al., 2011; King-Casas, 2005). These expectations are not limited to the specific interaction partner, however, but are generalized to other members of that person's group (Kocsor & Bereczkei, 2017; Ramasubramanian, 2011; Stark et al., 2013; Van Oudenhoven et al., 1996) as a way of minimizing the cognitive resources necessary to form impressions (Macrae et al., 1994). While we focus on direct interaction here as a major mechanism for impression formation, such experience is of course complemented in the real world by more indirect forms of learning such as gossip (E. C. Collins et al., 2011; Smith & Collins, 2009).

Impression formation through direct social interaction is not a passive reception of social information, but an active process in which people select who to interact with (and learn about) and who to avoid. Such approach and avoidance behaviors are largely shaped by attitudes and experiences, with people more likely to continue approaching sources that have produced positive outcomes in the past and avoiding those that have produced negative outcomes (Chambliss, 1965; Chen & Bargh, 1999; Elliot & Covington, 2001; Fazio et al., 2004; Taylor et al., 1969). When combined with group-based generalizations, this can shape patterns of approach and avoidance behaviors toward entire social groups. Indeed, research has found that negative experiences with some members of a group can produce negative expectations about future interactions with that group, which in turn predicts avoidance of the group as a whole (Levin et al., 2003; Plant, 2004; Plant & Devine, 2003). Critically, such early avoidance reduces the opportunity to gain further information about the group, resulting in inaccurate beliefs in cases where initial experiences were unrepresentative of the group. While avoidance can provide information about the effectiveness of the avoidance behavior itself (e.g., an animal that presses a button that halts the presentation of a shock will learn that pressing the button is an effective avoidance mechanism), no information is gained about the avoided *stimulus* (the animal does not learn anything more about whether the shock would have occurred if they had not pressed the button). We focus here on the properties of avoidance in preventing information-gain about a stimulus rather than on the process of learning the effectiveness of avoidance behaviors themselves. We term this mechanism "ignorance-based avoidance" to illustrate the cycle between avoidance and failures to update beliefs: avoidance prevents feedback, leading to ignorance of the true nature of the group, and this ignorance in turn leads to further avoidance even when it is not warranted.

Initial support for the importance of avoidance behaviors in impression formation comes from simulation studies demonstrating that information about initially positive objects is more likely to be updated in the future than information about initially negative objects (Denrell, 2005; Denrell & March, 2001; Konovalova & Le Mens, 2020). Experimental research on attitude formation has also demonstrated the existence of such approach biases based on initial experiences. In particular, Fazio and colleagues have shown that when learning about objects is contingent on approaching them, avoidance based on early experiences can lead to inaccurate negative beliefs about objects that are actually positive (Eiser et al., 2007; Fazio et al., 2004; Shook & Fazio, 2009). Critically, when feedback is instead given on every trial regardless of the decision to approach or avoid, people are better able to accurately update their initial negative beliefs.

Although work examining the formation of attitudes toward objects provides a vital starting point for understanding attitudes about social groups, the formation of social attitudes involves a variety of additional processes that may be unique to the social environment. First, group-based generalizations can be especially harmful in the social domain: whereas it may be rational to avoid beans that look similar to one that made you ill in the past (as a bean's appearance likely relates to its effects), choosing to avoid people who look similar to someone who hurt you in the past involves moral considerations (such as applying potentially harmful stereotypes to someone who you have never interacted with) in addition to pragmatic ones. Second, the social world is dynamic and complex, perhaps even more so than the physical world, and a single interaction is rarely sufficient to determine what a person is like (whereas a single case of food poisoning may indeed be sufficient to determine that a bean is poisonous). Finally, people have a remarkable ability to group others into social categories even when the features they use to form groups are not inherently categorical (Hirschfeld, 1995; Rothbart & Taylor, 1992). As such, rather than solely assessing generalization along continuous gradients, it is also necessary to examine the formation of discrete categories and the implicit assumption that groups are homogeneous. We apply these considerations in the current studies to examine how avoidance may lead to the persistence of existing beliefs about social groups in situations where information is contingent on approach.

Self-Reinforcement Effects: Avoidance Reinforces Itself

Avoidance as a behavioral response may have properties that further contribute to the perpetuation of stereotyping and discrimination. Although initial avoidance of a negative stimulus can be adaptive, work on anxiety and fear learning has demonstrated that avoidance responses can become reinforced until they are largely insensitive to any actual outcomes (LeDoux et al., 2017). The maintenance of avoidance responses over time presents something of a challenge for traditional theories of instrumental learning, which typically assume that a stimulus will become reinforced if it gives rise to positive outcomes (positive reinforcement) or prevents negative outcomes from occurring (negative reinforcement). As avoidance behaviors by nature often prevent the gain of information, someone who avoids typically receives no feedback about the avoided stimulus and, thus, no external reinforcement for the avoidance behavior. It has therefore been a challenge to explain why, absent any external reinforcement, avoidance responses do not extinguish over time (Kim et al., 2006; LeDoux et al., 2017).

To account for this difficulty, theories of avoidance learning have proposed a variety of ways in which avoidance behaviors can become reinforced (Hofmann & Hay, 2018; Kim et al., 2006; LeDoux et al., 2017; Maia, 2010; Palminteri et al., 2015). For instance, using a paradigm in which participants repeatedly chose between two novel conditioned stimuli with different probabilities of causing either monetary loss or monetary reward, Kim and colleagues (Kim et al., 2006) propose that the avoidance of an aversive outcome is itself rewarding, such that avoidance behaviors can reinforce themselves. Critically, when making approach/avoid decisions between two stimuli, such that avoidance of one necessitates approach to the other, every choice provides feedback about the approached stimulus. On every trial, participants know whether their choice allowed them to successfully avoid a loss or not, and this knowledge can be used to update their impressions of the stimuli for future choices. These studies indicate that explicit feedback regarding the successful avoidance of a negative outcome can have reinforcing effects on behavior.

Complementing this line of work, the psychological processes that govern learning in the absence of explicit outcome feedback have also been examined. This work has shown that when individuals receive feedback only selectively based on their behavior (e.g., learning about a company's return only if they decide to invest in it), learning in the absence of external feedback involves the generation of internal feedback, with individuals acting as though they received feedback that confirmed their expectations (Elwin et al., 2007; Henriksson et al., 2010). While not specifically focused on avoidance, these findings may suggest that when individuals avoid a stimulus out of a belief that it is negative and thus receive no feedback, they may encode the expected negative outcome, acting as though they received confirmation that the stimulus was indeed negative. This internally generated or "imagined" feedback may be one way in which avoidance behaviors become reinforced in the absence of external feedback.

Bringing these two lines of work together, we examine here the reinforcement of avoidance responses in situations where no explicit information is gained and people can only infer the appropriateness of a response from an expected or simulated outcome (i.e., if I had approached, something bad *would have* happened). Thus, although no concrete reward or punishment signal is present, an imagined outcome may (perhaps through simulation) function similarly to knowledge of a real negative outcome, serving to reinforce the avoidance response. On this view, not only will avoidance prevent the gain of any actual information about the avoided stimulus, but individuals may further project and strengthen their preexisting beliefs onto the stimulus category after avoidance.

These reinforcing behavioral effects of avoidance may also be paralleled by changes in explicit beliefs about an avoided stimulus. In particular, avoidance may increase one's expectations of the threat posed by the avoided stimulus even in the absence of any actual threat signals (Engelhard et al., 2015; Van Den Hout et al., 2014; van Uijen et al., 2018; van Uijen & Toffolo, 2015; Vervliet & Indekeu, 2015). That is, although avoidance actually removes the opportunity to gain further information about a stimulus, the very act of avoidance can itself sometimes serve to increase perceptions of the stimulus as threatening (perhaps through a form of inferring one's attitudes from one's behavior; Bem, 1972; Fazio et al., 1977). This can apply even to objectively neutral stimuli, with avoidant safety behaviors increasing perceptions of the threat posed by a stimulus even when the stimulus has never actually predicted threat (Engelhard et al., 2015). Thus, in addition to reinforcing itself at the level of stimulus–response associations (reinforcing the implicit behavior of avoiding without changing explicit beliefs), avoidance may also change conscious representations or beliefs about the avoided stimulus.

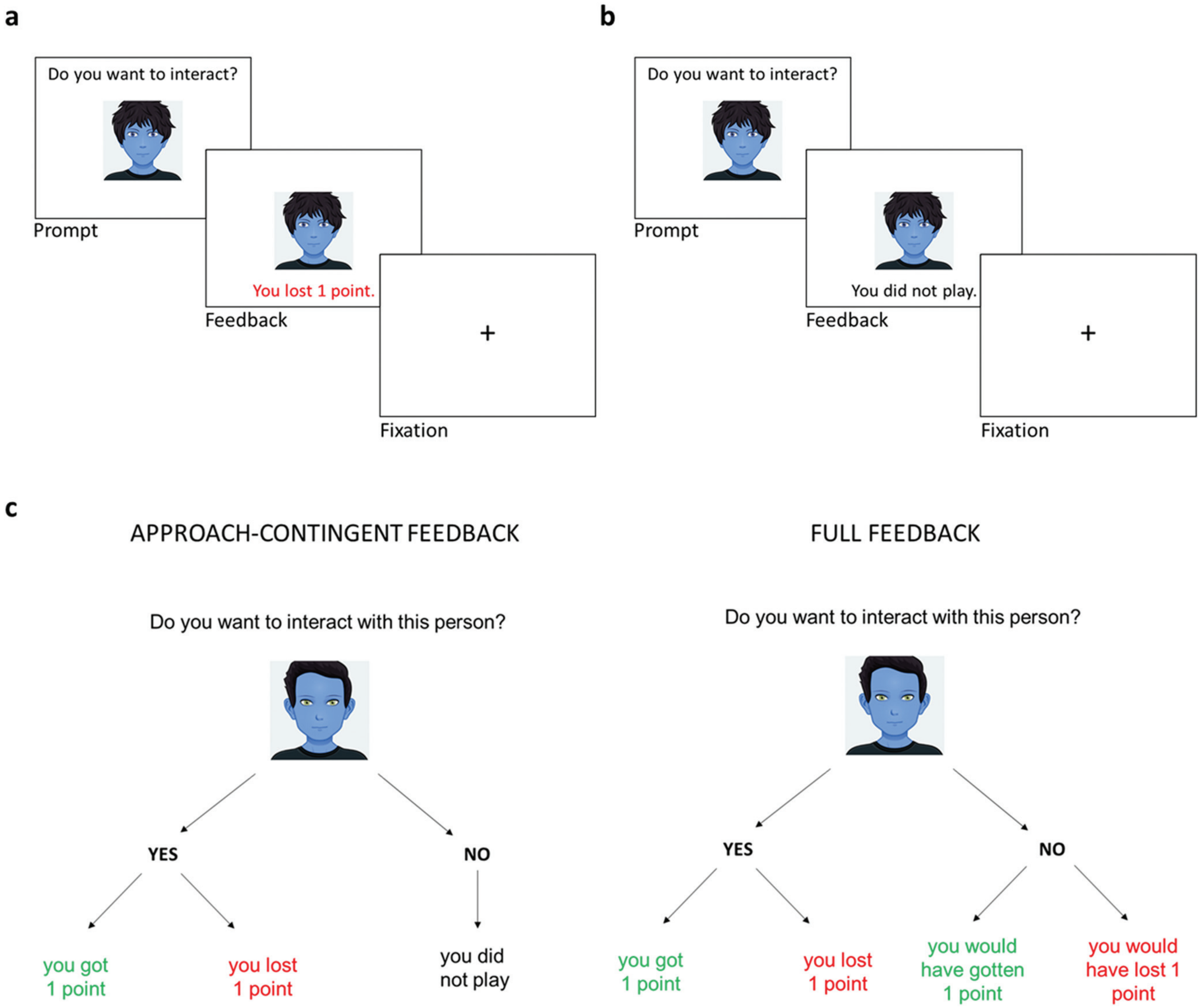
The work reviewed here focuses on relatively automatic or reflexive responses to fear-inducing stimuli; however, we propose that similar principles might apply in the instrumental learning of social rewards, such as in situations where people make decisions about whether or not to interact with someone in a social setting. Thus, in the current work we use computational models to examine whether avoiding a member of a novel social group might have self-reinforcing effects on one's behavior, further increasing the probability of later avoiding that group. To illustrate, take the example of someone who is walking alone late at night and, upon encountering someone she perceives as threatening, decides to cross the street to avoid interacting with them. The sense of relief she feels at having successfully avoided what could have been a negative interaction may serve to reinforce that behavior in the future. This reinforcement might occur at the level of stimulus–response associations in a manner that reinforces only the specific behavior in question, such that the link between seeing someone of a particular group approach and crossing the street to avoid them is strengthened. Alternatively, it might occur at the level of conscious beliefs in a way that extends to other behaviors, such that the general belief that people in that group are threatening is strengthened. The former mechanism would suggest a divergence between explicit beliefs and behavior, with avoidance influencing behavior but not self-reported beliefs, whereas the latter posits that self-reinforced avoidance affects both behavior and explicit beliefs. Although we do not aim to fully distinguish these possibilities in the current work, we explore this question to see if one of these mechanisms is more likely than the other. As addressing this question requires consideration of an individual's full history of choices and feedback with a given group, we combine novel groups with the use of computational models that can isolate any self-reinforcing effects of avoidance.

The Current Research

The current research uses a combination of behavioral experiments and computational models of reinforcement learning to examine the role of avoidance in maintaining inaccurate stereotypes and negative behaviors toward social groups. As outlined above, we propose two separate but related mechanisms through which avoidance may perpetuate negative beliefs about a group. First, the ignorance-based mechanism suggests that in situations where learning about someone requires you to approach them, avoidance removes the opportunity to gain further information about a target, preventing inaccurate beliefs from being corrected. Second, the self-reinforcement mechanism proposes that the "imagined" negative outcome that participants may encode after avoidance can function similarly to a real negative outcome and serve to reinforce the behavior, perhaps by eliciting a sense of relief at having avoided a presumably negative outcome. In this way, avoidance may acquire a positive "value" and directly reinforce itself, increasing the probability of future avoidance behaviors.

To test these two hypotheses, we examine how initial negative experiences with a novel group lead to subsequent avoidance of new group members, even when these new people are much more positive than the initial members. In five studies, we use a learning task in which participants meet two groups of alien people and can choose to approach or avoid each alien that they see (see Figure 1). If approached, an alien can either help the participant (by giving a point) or hurt the participant (by taking a point). One group of aliens (referred to here as the initially cooperative group) initially has a much higher probability of cooperating than the other (the initially uncooperative group); however, as the task goes on, the initial aliens are replaced by new ones with different rates of cooperation, such that by the end of the task the aliens cooperate

Figure 1
Example Trials and Task Design



Note. (a) Example trial where the participant approaches the alien, who does not cooperate with them. (b) Example trial where the participant avoids the alien under approach-contingent feedback. (c) Design of the task in the approach-contingent feedback conditions (Studies 1–5) and the full feedback conditions (Studies 4 and 5). In both conditions, participants choose whether to approach or avoid the alien. In the approach-contingent feedback conditions, participants learn whether the alien gave them 1 point or took 1 point away only if they chose to approach the alien; if they avoided, they do not receive any information about the alien. In the full feedback conditions, participants learn about the alien's behavior regardless of whether or not they approach, although this outcome only affects their point totals if they chose to approach the alien. See the online article for the color version of this figure.

at more similar rates (or, in Study 3, opposite rates). Similar types of games in which participants have repeated interactions with individual targets have been used previously to examine social and moral impression formation (Delgado et al., 2005; FeldmanHall et al., 2018; Siegel et al., 2018). Across studies, we vary the degree of change in the behavior of the groups. In some studies, the groups are neutral by the end of the study, cooperating about half the time. Similar to tasks assessing counterconditioning, in which a stimulus that was originally learned to be positive or negative is reconditioned with the opposite valence (Keller et al., 2020), other studies have the groups end up with a valence opposite to what was initially learned.

This work examines impression formation and group-based generalization when initial experiences with members of a group are unrepresentative of later potential experiences. Generalizing impressions of some members of a group to others can of course be problematic, as even groups that share some similarities have within-group variation. Judging people one has never met based on initial experiences with other members of their groups can lead to inaccurate impressions in many ways. It is possible that one's early experiences with members of the group were simply unrepresentative of the group as a whole, either by chance or through some systematic bias in the way in which they encounter others. Alternatively, the composition or behavior of the groups might actually change over time. In either case, early experiences can build up expectations that are invalid when generalizing to the rest of the group, especially when applied to completely new people. In these studies, the cause of the change in the participant's experience is unknown to them; it is possible that initial experiences were simply unrepresentative or that the group as a whole has changed. Participants meet each alien multiple times and individual aliens do not change in their rate of cooperation across the study; thus, someone who chooses to form individualized impressions of each alien that they meet can accurately learn to approach those who are cooperative and avoid those who are uncooperative. Instead, individual group members are simply replaced by new ones who behave differently from the old members. People who rely on group-based generalizations rather than individual impressions will therefore fail to appropriately approach positive aliens and avoid negative aliens near the end of the task.

We examine whether beliefs based on initial experiences with the two groups persist despite these changes, leading to continued avoidance of the initially negative group even when the group is no longer negative. If avoidance contributes to stereotype maintenance by removing the opportunity to update one's beliefs (ignorance-based avoidance), negative stereotypes should persist for longer in cases where information is contingent on approach. Further, if avoidance causes additional stereotype persistence by directly reinforcing one's negative expectations about the group (self-reinforced avoidance), computational models of reinforcement learning should demonstrate that when information is contingent on approach, avoidance decreases the value associated with the target of the avoidance, thereby further reducing the likelihood of later approach. Whereas standard models of reinforcement learning assume that values of unchosen or avoided options either remain unchanged or decay back to zero (Boorman et al., 2009; Cavanagh, 2015; A. G. E. Collins & Frank, 2016; Niv et al., 2015; Yechiam & Busemeyer, 2005), here we instead test whether negative values actually become more negative after the stimulus is avoided. Note that such self-reinforcement is not necessary for avoidance learning, as even in the absence of such reinforcement participants could learn that the value of approaching an uncooperative alien is less than 0, while avoidance would be ascribed a value of 0 and be the preferred behavioral choice. Thus, the avoidance model tests not simply whether participants are learning that the initially uncooperative group is bad, but whether this learning becomes *stronger* after avoidance. Finally, we examine whether avoidance leads to any dissociations between behaviors and self-reported beliefs, exploring whether self-reinforced avoidance occurs at the level of stimulus–response associations (affecting behavior but not beliefs) or at the level of explicit representations (affecting both behavior and beliefs).

Studies 1 to 3 focus on understanding the mechanisms of belief-updating in conditions where feedback is contingent on approach; Studies 4 and 5 then attempt to isolate the role of approach-contingent feedback by comparing it to conditions where feedback is independent of approach. As individual groups are not encountered in a vacuum but exist within a larger social context with multiple other groups, Study 3 also explores how "local" belief-updating about an individual group might relate to more "global" belief-updating about the world at large or about other groups. In each of these five studies, we model participants' approach behavior using reinforcement learning models to test whether avoidance has self-reinforcing effects. Thus, all studies test the self-reinforced avoidance hypothesis and Studies 4 and 5 directly test the ignorance-based avoidance hypothesis.

Studies 1 and 2: Generalization of Initial Learning Under Approach-Contingent Feedback

To examine the role of avoidance in failures to update beliefs, we first ran an initial study as well as a replication study. As results were highly similar, we report these two initial studies together.

Method

Participants

For Study 1, 98 participants were recruited from Amazon Mechanical Turk (MTurk) and informed that they could earn up to \$5 depending on their performance in the game. Ten participants had participated in a pilot version of a similar study and were excluded, leaving a final sample size of 88 (36 female, 52 male; M_{age} = 36.7). For Study 2, 100 participants (54 female, 46 male; M_{age} = 39.1) were recruited from MTurk and again told that they could earn up to \$5 depending on their performance in the game. For all studies, MTurk workers were eligible to participate if at least 95% of their previous assignments had been approved, they were located in the United States or Canada, and they had not completed an earlier study using a similar task (such that samples were independent across studies). These sample sizes gave us 80% power to detect an effect of $b = -.16$ for the main analysis of interest in Study 1, and of $b = -.15$ for Study 2 (all power analyses were conducted using the simr package in R; Green & Macleod, 2016). For all studies, informed consent was obtained from all participants before beginning the experiment. This research was approved by the University of Toronto Research Ethics Board (protocol 33582), and all relevant ethical regulations were complied with.

Procedure

Participants were told that they would play a game that involved making decisions about people from different alien species. On each trial of this game, they would see a picture of an alien and have to decide whether or not to interact with them. No time limit was imposed for responding, and across all five studies over 95% of response times were less than 3 seconds. If they chose to interact, the alien would either give them 1 point or take away 1 point. If they chose not to interact, they would not get any information about the alien's actions. Thus, feedback in the game was contingent upon approach (see Figure 1 for an example trial). Participants were told that they would see the same aliens again and again, and that it would be useful to figure out which aliens were more likely to give or take money and to use that to guide their decisions. At the end of the task, the total points they had earned would be converted into real bonus money, although participants were not told the scaling factor to convert points to money. Participants started the task with $2 in payment; a positive final point total would lead this amount to increase, while a negative point total would subtract from this amount.

The aliens that participants saw in the game belonged to two visually distinct species, characterized by either green or blue skin in Study 1, and two of green, blue, and yellow skin in Study 2 (only two of these colors were seen during the task, with the third color only seen in the posttask ratings). The behavior of each individual alien was probabilistic and governed by an overall cooperation rate that differed across groups at first. At the beginning of the game, one of the groups (color randomized across participants) had a high rate of cooperation, with members of this group giving the participant money on 80% of trials, while the members of the other group cooperated only 20% of the time. We refer to these groups as the initially cooperative group and the initially uncooperative group, respectively, although these labels were not used with participants. To ensure that skin color was the most salient dimension of categorization, all aliens in our studies appeared male. The aliens differed in all other facial features, such as hair style and color, facial shape, and facial features.

Most aliens were seen multiple times, but throughout the task, the aliens in each group were slowly replaced with new aliens (with each alien within a group again differing along all facial features except skin color). Specifically, the study consisted of nine continuous rounds, with a single alien in each species being replaced each round. The order in which participants encountered aliens within a round was randomized. Eight aliens per group were encountered in each round, for a total of 144 trials in the game. In the first round, all aliens in the initially cooperative group cooperated with a probability of .8, while aliens in the initially uncooperative group cooperated at a rate of .2. Each round, one new alien was introduced into each species, replacing an old alien. For the first five rounds, the cooperation probabilities of the new aliens remained consistent with the rest of their group—that is, new members of the initially cooperative group also cooperated at a rate of .8, and new members of the initially uncooperative group cooperated at a rate of .2. From rounds 6 to 9, however, the new aliens introduced each round were equal to each other, cooperating at rates of .5. Thus, by the final round, half of the aliens in each species were equal to each other, while the other half remained very cooperative or very uncooperative, depending on the species. Individual aliens are consistent in their rates of cooperation over time and do not change; only the participant's experience of the group as a whole changes. For the purposes of analyzing approach behavior, we separate out the aliens who cooperated at their group's initial rates (cooperative or uncooperative) and the aliens who cooperate at equal rates across the groups in the final round.

After completing all nine rounds of the game, participants were presented with a series of faces and had to rate how likely each alien was to cooperate with them, from 0 to 100. They rated each alien that they saw in the game, as well as four new aliens from each group, for a total of 20 ratings. In Study 2, participants additionally rated aliens from a third group of aliens (with a different skin color than the other two groups) whose members were not encountered during the task. Participants then answered a series of demographic questions. In all studies, we also had participants complete the Ten Item Personality Inventory as exploratory data for future research; these data have not been analyzed and are not discussed further here.

Analyses

All analyses of approach behavior were conducted using multilevel logistic regression models (using the "glmer" function from the lme4 package; Bates et al., 2015) with random intercepts for subjects. For all models, we report Type III Wald χ^2 values and two-sided p-values for each effect in the model, which we obtained using the car package (Fox & Weisberg, 2019).[1] R^2 values calculated using the r2glmm package (Jaeger, 2017) are reported for main analyses of interest.

Reinforcement Learning Models

To test for any self-reinforcing effects of avoidance, we applied reinforcement learning models to participants' behavior during the task, comparing models in which avoidance causes belief updating to models in which representations are unchanged after avoidance. In all models, we assume that participants track some value associated with each group as a whole and use this value to decide whether to approach members of each group. On each trial where the participant approaches the alien, the value of the approached group is updated according to the equation

$$V_{t+1} = V_t + \alpha(R_t - V_t)$$

where V_t represents the value at time t, R_t represents the outcome (-1 or 1) at time t, and α represents the learning rate. The probability of approaching the alien on trial t is then governed by the logistic equation, with α as the inverse temperature parameter controlling how deterministic choices are (with higher values indicating more deterministic choices and lower values indicating more random choices):

$$p(approach) = \frac{1}{1 + e^{-\beta V_t}}$$

In addition, all models include an additional parameter that represents participants' initial beliefs about how likely people are to cooperate in general. This parameter influences how likely participants are to interact with aliens at the beginning of the task, when they have no knowledge about the individuals or groups.

We tested three competing models to determine which best characterized participants' behavior during the task. The first model was the null model, which assumed that no updating of representations takes place when the participant chooses to avoid an alien. In other words, after avoidance:

$$V_{t+1} = V_t$$

The two remaining models assume that the values are updated after avoidance, with the magnitude of the best-fit avoidance parameter governing the degree of value-updating. These models assume that if the value for the group presented on the current trial

[1] Note that Wald chi-square values are asymptotically equivalent to F values (Davidson & MacKinnon, 2004)

is negative and the participant chooses to avoid, that value is updated according to the equation

$$V_{t+1} = V_t + \alpha\theta$$

where θ is the avoidance parameter. We test one model in which the same avoidance parameter is fit to both alien groups in the study and another model in which separate avoidance parameters are fit to each group. This value update after avoidance only happens if the current value associated with the group is negative. In other words, even if avoidance reinforcement were found for the good group as well as the bad group, this would only suggest that avoidance toward that group is reinforced during periods of time when expectations of that group are negative.

An avoidance parameter with a positive value would indicate that after avoidance, negative values get slightly less negative. This would fit with models assuming that the values of unchosen options decay back to a value of 0 (Boorman et al., 2009; Cavanagh, 2015; A. G. E. Collins & Frank, 2016; Niv et al., 2015; Yechiam & Busemeyer, 2005) as learning extinguishes over time. In contrast, a negative avoidance parameter would indicate that after avoidance, values are actually becoming more negative, as if people had received negative information about the avoided group. Finally, a value of 0 for the avoidance parameter would indicate that avoidance responses are not reinforcing themselves, but neither are they being extinguished. As described earlier, this would not mean that participants are not learning that the initially uncooperative group is bad, but simply that these avoidance responses do not increase the likelihood of later avoidance responses. Thus, the avoidance parameter tests whether avoidance has self-reinforcing effects over and above basic processes of learning who is good and who is bad.

All models were fit using hierarchical Bayesian parameter estimation, which assumes that subject-level parameter values are drawn from some overall group distribution, estimating values for both the subject-level and group-level parameters. Group-level parameters were given noninformative or weakly informative priors. Model-fitting in all studies was implemented in JAGS (Plummer, 2003) using three chains of 200,000 samples and 20,000 burn-in samples each, saving every second value to reduce model size. Visual inspection of the MCMC chains for all parameters suggested proper mixing; in addition, most parameters had an effective sample size greater than 10,000 and a Gelman-Rubin statistic smaller than 1.1, as recommended by Kruschke (2014). Parameter recovery simulations were conducted to ensure that the model could accurately recover the true parameter values (see Online Supplemental Materials Figure 2).

To test whether a model that includes an avoidance parameter fits the data better than the model without an avoidance parameter, a combination of DIC-based model comparison (Spiegelhalter et al., 2002) and estimation approaches to hypothesis testing (Kruschke, 2014) was used. For the former method, deviance information criterion (DIC) values were calculated for each version of the reinforcement learning model, with lower values indicating better model fit. For the latter method, we test whether the 95% highest density interval (HDI) of credible avoidance parameter values excludes the null value of 0; if it does, this is taken as evidence that the parameter is different from 0.

To explore which model corresponded best to participants' explicit beliefs, the values estimated from each model were used to predict explicit ratings of the aliens' cooperativeness. For each participant, the modal best-fit parameters from each model were used to simulate data 100 times. The average of the final values for each group were then taken across these 100 simulations for each of the three reinforcement learning models, and these average final values were used to predict that participant's average ratings of how cooperative members of the given group were. As ratings of old and new members of the groups were highly similar despite their different rates of cooperation, we averaged across all members within a group to create a single rating per group. Across all studies, the final values produced by these three models were correlated at rates between .46 and .71. We compared these models by examining which model's values produced the highest coefficient when all three values were placed in the same regression model, as well as comparing the Akaike's information criterion (AIC) and Bayesian information criterion (BIC) values when final values from each model are entered into separate regression models to predict ratings. We report results for the simulated values described above here, but all results of explicit ratings replicate if we instead use the final values estimated directly from within the reinforcement learning models. As the models with avoidance parameters produced final values that were negatively skewed, an additional series of models was run for all studies to ensure that the skewed nature of data from two of the models did not account for the results. Overall, the results do not substantially change with these additional checks (see Online Supplemental Materials Table 1).

Results

Initial Learning About Group Differences

In the first five of the nine rounds in the task, participants learn about two groups of aliens, one whose members cooperated 80% of the time and the other whose members cooperated 20% of the time. To establish whether participants accurately learned about the two groups in this initial learning stage, approach behavior to these aliens was predicted from their group membership. Indeed, participants approached aliens in the initially cooperative group at a much higher rate than aliens in the initially uncooperative group, $b = -1.01$, $\chi^2(1) = 3934.82$, $p < .001$, indicating that they accurately learned about the initial members of the two groups.

Generalization of Approach Behavior

By the end of the task, half of the aliens in each group had been replaced by new aliens who cooperated at equal rates to each other (cooperating 50% of the time). The critical hypothesis for this study was that because information is contingent on approach, participants would fail to fully update their beliefs despite the changes in the composition of the groups, and would continue approaching new members of the initially cooperative group more than new members of the initially uncooperative group, despite the fact that they cooperate at equal rates. To test this, we selected only those trials from round 6–9 in which participants could interact with a new alien whose cooperation rate was the same across the two groups, and predicted approach behavior from the alien's group membership. Supporting this hypothesis, we found that participants continued to approach new members of the initially cooperative group more than new members of the initially uncooperative group, $b = -1.07$, $\chi^2(1) = 618.92$, $p < .001$, $R^2 = .145$, suggesting that their initial beliefs persist despite the fact that these individuals cooperated at equal rates (see Online Supplemental Materials Figure 1). Participants' explicit ratings of the aliens' cooperativeness reflect a similar failure to update (see online supplemental material for analyses).

Reinforcement of Avoidance

A key question in this study concerns the role of avoidance in the maintenance of initial beliefs—specifically, the question of whether avoidance behaviors reinforce themselves, such that avoiding a group out of the belief that its members are negative makes further avoidance of that group more likely. If avoidance has self-reinforcing effects, the values that participants associate with a given group should be negatively updated after a member of the group is avoided, so that the group is seen as more negative. If, on the other hand, avoidance simply removes any opportunity to update one's beliefs, the value of the group should remain unchanged after avoidance. Finally, if forgetting occurs, the value of the group may actually become more positive after avoidance, allowing it to decay back to 0. To test this, we fit reinforcement learning models to the data from all trials, comparing a null model in which values remain constant after avoidance to two models in which values are updated after avoidance. We also wanted to allow for the possibility that avoiding an extremely negative group is reinforced differently than avoiding a largely positive group. Thus, the first of these alternative models assumes a single update factor for both groups after avoidance, whereas the second assumes separate update factors for each group after avoidance (see Method). The model with a single avoidance parameter provides the best fit to the data (see Table 1 for DIC values) and the 95% HDI for this avoidance parameter spans negative values that exclude 0 in both the initial study and the replication, providing evidence for the idea that values are negatively updated after avoidance (see Table 2 for parameter values) and that this updating happens similarly for both groups. Critically, although we find evidence of avoidance reinforcement toward both groups, the model assumes this reinforcement happens only if the current expectations of the groups are negative. Thus, the reinforcement of avoidance toward the good group would only occur in cases where a participant erroneously believes that this group is bad (such as through a series of probabilistic negative initial encounters with the otherwise positive group). As such, avoidance reinforcement toward the "good group" should not be overinterpreted, as overall there are few avoidance behaviors toward this group with which to estimate a separate avoidance parameter and, thus, we focus our discussion mainly on avoidance of the negative group.[2]

Together, these models support the idea that avoidance has self-reinforcing effects on behavior. However, it remains unclear whether such avoidance simply strengthens stimulus–response associations, increasing the likelihood of future avoidance without necessarily changing explicit beliefs, or actually affects the underlying representation of the alien in a way that transfers beyond the simple act of avoiding. Although we cannot conclusively distinguish between these possibilities in the current study, we examined whether

Table 1

DIC Values to Compare Reinforcement Learning Models Without Avoidance Reinforcement, With One Avoidance Reinforcement Parameter for Both Groups, and With Separate Avoidance Reinforcement Parameters for Each Group

Study	Model	DIC
1	No avoidance parameter	12,506.29
	One avoidance parameter	12,355.46
	Two avoidance parameters	12,557.70
2	No avoidance parameter	13,678.08
	One avoidance parameter	13,513.78
	Two avoidance parameters	13,585.13
3	No avoidance parameter	47,579.81
	One avoidance parameter	46,825.89
	Two avoidance parameters	46,858.15
4	No avoidance parameter	24,262.77
	One avoidance parameter	24,051.99
	Two avoidance parameters	24,062.92
5	No avoidance parameter	52,126.14
	One avoidance parameter	51,928.32
	Two avoidance parameters	51,790.29
All combined	No avoidance parameter	150246.40
	One avoidance parameter	148867.60
	Two avoidance parameters	148839.90

Note. Lower deviance information criterion (DIC) values indicate a better fit.

participants' subjectively reported ratings of each alien's cooperativeness were better predicted by a model with avoidance reinforcement or the null model. If avoidance serves to reinforce avoidance behaviors themselves without influencing overall representations of the avoided group, the null model should best predict participants' ratings despite providing a worse fit to the behavioral data. In contrast, if avoidance actually updates participants' representations of the group in a way that transfers across behaviors, a model with avoidance reinforcement should provide a better prediction of participants' ratings. Results indicate that the null model without any reinforcement of avoidance best predicts participant's ratings in both studies (see Table 3 for model estimates, AIC values, and BIC values, and Figure 2 for plotted values and ratings). Thus, this potentially suggests an account in which avoidance strengthens stimulus–response associations between the presumably negative group and future avoidance behaviors without necessarily changing explicit beliefs about the group in a way that transfers to other response modalities.

Discussion

Studies 1 and 2 indicate that initial differences between two groups produce different patterns of behavior toward them that apply even to later group members who cooperate at equal rates. Critically, these later members had never been encountered before in the task. The finding that participants nevertheless treated them

[2] We also test whether these effects can instead be explained by choice perseveration (Seymour et al., 2012; Worthy et al., 2013), in which individuals tend to repeat their previous choices regardless of the outcomes they experience. In all five studies, we find that a model with avoidance reinforcement and choice perseveration fits better than a model with only choice perseveration, indicating that this phenomenon is not responsible for our effects.

Table 2

Modal Group Parameter and 95% Highest Density Interval (HDI) Values for Reinforcement Learning Models

Parameter	Study 1	Study 2	Study 3				Study 4	Study 5	All studies
			Both bad	Both good	Stays same	Switch			
Learning rate	0.05 [0.04, 0.06]	0.05 [0.03, 0.06]	0.004 [0.001, 0.04]	0.07 [0.04, 0.10]	0.04 [0.03, 0.06]	0.05 [0.01, 0.08]	0.11 [0.09, 0.12]	0.13 [0.12, 0.15]	0.09 [0.08, 0.09]
Inverse temperature	4.26 [3.64, 5.04]	5.83 [4.97, 6.92]	5.08 [3.99, 6.56]	3.17 [2.59, 3.85]	3.55 [2.92, 4.18]	3.58 [2.93, 4.46]	4.21 [3.85, 4.61]	3.46 [3.26, 3.68]	3.92 [3.75, 4.10]
Starting value	0.31 [0.22, 0.41]	0.16 [0.12, 0.21]		0.26 [0.22, 0.32]			0.19 [0.15, 0.24]	0.26 [0.22, 0.30]	0.24 [0.22, 0.27]
Avoidance reinforcement (approach-contingent feedback)	−0.19 [−0.34, −0.08]	−0.11 [−0.21, −0.03]		−0.25 [−0.32, −0.19]			0.01 [−0.03, 0.07]	Good group −0.50 [−1.04, 0.53]; Bad group −0.06 [−0.09, −0.04]	−0.12 [−0.14, −0.09]
Avoidance update modifier (full feedback)	—	—		—			0.86 [0.16, 1.10]	0.38 [0.001, 0.55]	0.64 [0.04, 0.80]

differently based on their group membership suggests that participants generalized their beliefs about some members of the group to other members, indicating a belief that group membership determined an alien's behavior. In addition to these basic effects, we present evidence that avoidance itself has direct reinforcing effects on behavior, such that the decision to avoid increases the likelihood of later avoidance despite preventing the gain of new information. This reinforcement seems to occur at the level of implicit stimulus–response associations without affecting participants' explicit beliefs about the groups. These results provide direct evidence for the self-reinforcement mechanism and are consistent with the ignorance mechanism, although this mechanism was not directly tested in these studies.

Examining approach behavior to each group over time suggests that participants actually updated their behavior toward the initially uncooperative group more than their behavior toward the initially cooperative group (see Online Supplemental Materials Figure 1). This is unexpected, given that participants are receiving much more feedback about the behavior of the initially cooperative group, who they continually approach and learn from, than that of the initially uncooperative group, who they largely avoid and therefore do not learn from. This asymmetry in updating could potentially reflect differences in the volatility of beliefs about good and bad people, as previous research has found that beliefs about bad people are updated more easily (Siegel et al., 2018). However, while it appears as though people are forming more accurate beliefs about the initially uncooperative group than the initially cooperative group by the end of the task, this is partially confounded by the multiple functions served by approach behavior. In particular, there are two differing reasons participants may have for approaching a given alien: to gain points (under the assumption that the alien will give rather than take points) or to learn about the nature of the alien (when the participant is unsure about them). Thus, when participants began to notice that a change might be occurring in both groups, making their previous knowledge of the groups more uncertain, the logical response may be to gather more information about the groups to reduce uncertainty. As gathering information requires approaching the alien, this results in what, on the surface, seems like differing patterns of updating for the initially cooperative group and the initially uncooperative group: namely, that participants continue to approach the initially cooperative group as they had already been doing, and start approaching the initially uncooperative group even though they had been avoiding them before. In fact, examining participants' ratings of the aliens' cooperativeness at the end of the task reveals that they rate old and new aliens quite similarly, despite the differences in their actual rates of cooperation (see online supplemental materials). This lends support to the idea that participants may not actually be updating their beliefs about the initially uncooperative group more than the initially cooperative group in these studies, but instead approaching both groups when they suspect they may be changing to reduce their uncertainty about their behavior. Of note is that the changes in the two groups in Studies 1 and 2 were fairly small and happened quite late in the task, meaning that participants may not have had time to get past the "information-gathering" stage of adjusting their behavior after the groups changed. Thus, to more accurately examine how participants update their behavior in response to the groups' changes, in later studies the behavior of the two groups

Table 3
Predicting Ratings of the Aliens' Cooperativeness from the Final Values Produced by Each Participant's Reinforcement Learning Models

Study	Model	Estimate (same model)	AIC	BIC
1	No avoidance parameter	25.73	1,494.26	1,510.11
	One avoidance parameter	−17.05	1,510.92	1,526.77
	Two avoidance parameters	25.62	1,508.57	1,524.42
2	No avoidance parameter	46.46	1,690.17	1,706.66
	One avoidance parameter	4.03	1,713.32	1,729.81
	Two avoidance parameters	−6.68	1,716.42	1,732.91
3	No avoidance parameter	30.55	5,115.21	5,137.18
	One avoidance parameter	−1.27	5,233.88	5,255.85
	Two avoidance parameters	3.81	5,214.32	5,236.28
4	No avoidance parameter	46.17	2,971.59	2,990.83
	One avoidance parameter	−2.33	3,019.59	3,038.82
	Two avoidance parameters	4.62	3,007.15	3,026.38
5	No avoidance parameter	22.03	7,012.65	7,036.14
	One avoidance parameter	−2.68	7,054.51	7,078.00
	Two avoidance parameters	9.1	7,052.22	7,075.71

Note. For each study, we compare a null model without avoidance reinforcement to a model with a single avoidance parameter for both groups (assuming the same degree of reinforcement after avoiding the initially cooperative group and the initially uncooperative group) as well as a model with separate avoidance parameters for each group (allowing different amounts of reinforcement after avoiding each group). We compare models by examining how well each model's values predict ratings when entered into the same regression model (examining which accounts for the most unique variance, with higher estimates better), as well as comparing the Akaike's information criterion (AIC) and Bayesian information criterion (BIC) values when entering each model's values into separate regressions (with lower AIC and BIC values indicating better fit).

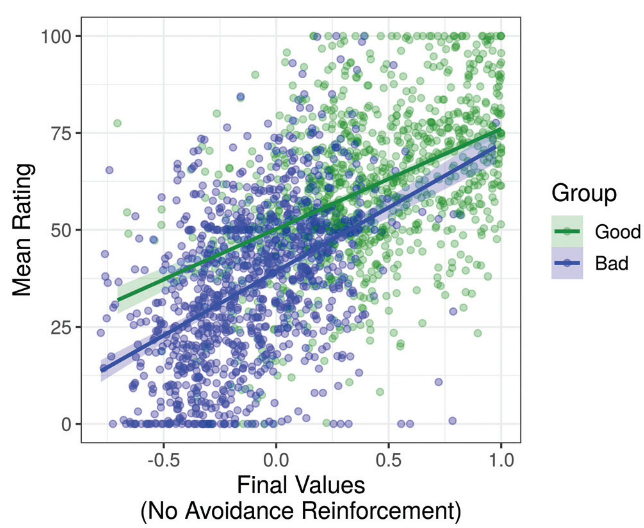
changes more drastically and the change occurs earlier in the task.

Study 3: Higher-Order Beliefs About Group Dynamics

Social decisions do not take place in a vacuum; rather, each social group that is encountered exists alongside many other groups. Thus, in addition to their beliefs about specific groups, people have higher-order beliefs about the homogeneity or consistency of groups in general. When a change in the behavior of one group is apparent, people may either update only the specific, local belief about the group in question (expecting members of that group to behave differently but all other groups to continue behaving as previously) or update their more global belief about how consistent groups are in general (taking the change in the previously predictable group's behavior as evidence that the environment has changed, and so expect other groups to change as well). For example, someone who believes that race is highly diagnostic of behavior may come to realize that a racial group they thought was wholly positive is actually more mixed. In addition to updating their beliefs about the group in question, they may also revise their higher-level beliefs about the diagnosticity of race and begin interacting more with members of racial groups they had previously avoided. In line with this idea, the aim of Study 3 was to examine how generalizations of initial group-based beliefs under approach-contingent feedback are affected by the larger context of the interaction, asking whether changes in the behavior of one group might serve as a signal about the potential behavior of another group.

Whereas the composition of the two groups changed in sync in Studies 1 and 2, Study 3 independently varied whether each group changes or not to explore whether a change in one group's behavior serves as a signal that the other group should also be re-explored. One group in this study started as cooperative (with its members cooperating 80% of the time) and the other started as uncooperative (with its members cooperating 20% of the time).

Figure 2
Participants' Average Ratings of Each Group as a Function of the Final Values Produced by the Best-Fit Reinforcement Learning Model



Note. Lines and dots are colored by the group the alien belonged to. Each dot represents one participant's rating for the given group, with a total N = 1,065 across all five studies. See the online article for the color version of this figure.

We then manipulated whether new members of each group were cooperative or uncooperative, creating four conditions: both groups end as cooperative, both groups end as uncooperative, the two groups switch in rates of cooperation, or the groups keep their initial rates of cooperation. If participants use a change in one group to update their global beliefs about the stability of groups in general, a change in the behavior or composition of one group may cause changes in participants' approach behavior to the other group, even when the other group has not changed. If, on the other hand, participants only update their local beliefs about the group that has actually changed, a change in one group should have no effect on their behavior toward the other group.

Method

Participants

Two hundred ninety-nine participants (149 female, 150 male; $M_{\text{age}} = 36.8$) were recruited from MTurk and told that they could earn up to $5 depending on their performance in the game. This sample size provided 80% power to detect effects of $b = -.14$ or larger for the two-way interactions of interest.

Procedure

The study consisted of a 2 (initially uncooperative group: stays uncooperative vs. becomes cooperative) $\times$ 2 (initially cooperative group: becomes uncooperative vs. stays cooperative) between-subjects design. Participants completed a task similar to that in Studies 1 and 2. In the first round of the game, aliens in the initially cooperative group cooperated at a rate of .8 and those in the initially uncooperative group cooperated at a rate of .2. Starting in the second round, the cooperation rate of each new alien that was introduced varied according to condition. Specifically, we varied whether later members of the initially cooperative group remained cooperative (cooperating at a rate of .8) or became uncooperative (cooperating at a rate of .2), as well as whether later members of the initially uncooperative group remained uncooperative or became cooperative. This created four between-subjects conditions: the two groups both ended as cooperative, both ended as uncooperative, switched in rates of cooperation, or kept their initial rates of cooperation. Again, as in Studies 1 and 2, individual aliens do not change in cooperation rates; rather, initial aliens are replaced by new ones who may have different rates of cooperation. In conditions where the groups change, the first new aliens that were introduced after round 1 had the new probability of cooperating (in contrast to Studies 1 and 2, where the first new aliens that were introduced cooperated at the same rate as the initial members of their groups). As a result, by the end of the game, all aliens had the new probability of cooperating (unlike in Studies 1 and 2). The game consisted of nine rounds with 16 aliens per round, producing a total of 144 trials.

After completing the game, participants were shown aliens from the game and new aliens and asked to rate how likely each one was to cooperate. They then completed the TIPI (for exploratory purposes as in the previous studies) and demographics.

Reinforcement Learning Models

Reinforcement learning models were then applied to the data exactly as in Study 1, except that we assumed separate group learning rate and temperature parameters for each condition.

Results

Initial Learning About Group Differences

To determine whether participants accurately learned about members of the two groups who actually did differ by group, approach behavior to these aliens was predicted from the alien's group membership. This analysis indicated that participants approached members of the initially cooperative group much more than those in the initially uncooperative group, $b = -.78$, $\chi^2(1) = 2350.67$, $p < .001$.

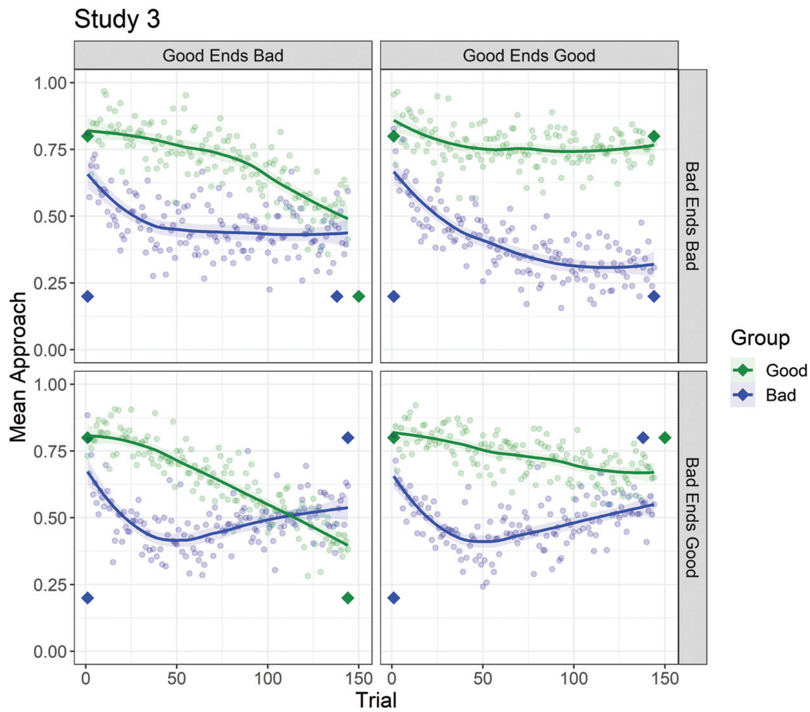
Generalization of Approach Behavior

Although we expect beliefs about initial group members to generalize somewhat to approach behavior toward later group members, the differences between earlier and later members are quite drastic in some conditions and participants may therefore still be able to update their beliefs somewhat, leading to differences among conditions. Thus, to examine how this learning generalized to new members of the two groups as a function of the participant's condition, approach behavior was predicted from the alien's group membership, whether the initially cooperative group stayed cooperative or became uncooperative, and whether the initially uncooperative group stayed uncooperative or became cooperative. Looking only at the final 16 aliens with the new rates of cooperation by condition, participants approached members of the initially cooperative group more than members of the initially uncooperative group, $b = -.52$, $\chi^2(1) = 121.63$, $p < .001$, model $R^2 = .074$. They do somewhat pick up on changes in the groups, as demonstrated by the two-way interactions between group and conditions, approaching new members of the initially uncooperative group more when they have become cooperative, $b = -.25$, $\chi^2(1) = 28.79$, $p < .001$, and approaching new members of the initially cooperative group more when they remain cooperative, $b = .29$, $\chi^2(1) = 37.63$, $p < .001$. Thus, although participants still do not update their beliefs entirely to reflect the character of new group members, they do somewhat notice changes in the groups and adjust their behavior accordingly. Figure 3 shows rates of approach to members of the two groups over trials to allow for visualization of how learning progressed over time. Only initial aliens are present for early trials and only new aliens are present for later trials, with both types of aliens present in the middle of the task. Collapsing across old and new aliens allows for better visualization of how participants' behavior to the entire group changed over time; to instead see the rates of approach to each individual alien with their own probability of cooperating for all studies, see Online Supplemental Materials Figure 1.

Effect of Changes in One Group on Approach to the Other

A critical question in Study 3 was whether changes in the composition of one group might signal to participants that the other group will also change, even in the absence of any actual changes in the other group. To test this, we examine how approach behavior toward new members of a group that has not changed varies

Figure 3
Approach Behavior Over Trials in Study 3



Note. Dots and smoothed lines represent average participant response, colored by group, and diamonds represent each group's actual average probability of cooperating at the beginning and end of the experiment in each of the four conditions. See the online article for the color version of this figure.

across trials as a function of whether the other group has changed. When predicting approach to the initially cooperative group, a significant interaction was found between trial number and whether or not the other group has changed, $b = .13$, $\chi^2(1) = 14.82$, $p < .001$, indicating that when new members of the initially uncooperative group are more cooperative than the old members, participants reduce their rates of approach to the initially cooperative group over time (see Online Supplemental Materials Figure 3). Similarly, when predicting approach to the initially uncooperative group members, a significant interaction was found between trial number and whether or not new members of the initially cooperative group are less cooperative, $b = .108$, $\chi^2(1) = 9.58$, $p = .002$, such that participants do not avoid members of the initially uncooperative group as much when new members of the other group are also uncooperative (see Online Supplemental Materials Figure 3). Together, these results suggest that changes in the composition of one group serve as a signal to participants that the other group is also going to change, causing them to adjust their behavior toward the other group even in the absence of any actual change in that group's behavior. These findings support the idea that global belief-updating about the social environment plays a role in learning and generalization about individual groups.

Reinforcement of Avoidance

As in the previous studies, we investigated the self-reinforcing effects of avoidance on approach behavior toward the two groups by comparing the fit of three reinforcement learning models (a model with separate avoidance parameters for each group, a model with a single avoidance parameter for both groups, and a null model with the avoidance parameter effectively fixed to 0). There were no theoretical reasons to expect differences in the reinforcement of avoidance across conditions in this study and thus we collapse across conditions when estimating the avoidance parameters (although learning rate and inverse temperature parameters were modeled separately by condition to reflect differences in learning). Replicating the previous studies, the model with a single avoidance parameter provided the best fit to the data (see Table 1) and the 95% HDI of values for the avoidance parameter in this model did not include 0 (see Table 2 for parameter values). This provides further evidence for the idea that values associated with the alien groups are negatively updated after avoidance, making future avoidance even more likely.

As in the previous studies, we find that final values from the model without any reinforcement of avoidance provide the best prediction of participants' ratings, as indicated both by the largest estimate when the different values are placed in the same regression model, and the lowest AIC and BIC values when each value is used to predict ratings in separate models (see Table 3). This provides further tentative support for the idea that avoidance reinforces stimulus–response associations but does not affect explicit beliefs about the aliens' cooperativeness.

Discussion

Study 3 demonstrates that people are able to pick up on large changes in the behavior of the groups, but that even their updated beliefs do not fully reflect the new group members' probability of cooperating (as seen in Figure 3). In addition, examining the slopes of change in approach to the initially cooperative group and the initially uncooperative group when both groups change (lower left quadrant of Figure 3) suggests that the slope of change in approach behavior to the initially cooperative group is steeper than that to the initially uncooperative group. This supports the idea that avoidance prevents belief-updating about presumably negative targets by limiting information-gain, whereas beliefs about presumably positive targets can be more easily updated through approach behavior (Denrell & March, 2001; Fazio et al., 2004). Under approach-contingent feedback, individuals in this condition who start out by avoiding members of a group they believe to be negative do not have enough of an opportunity to learn that they have overgeneralized, and that new members of this group are actually much more positive. Those who start out by approaching members of the presumably positive group, on the other hand, can better update their beliefs with the information they receive and realize that their generalized beliefs should not apply to the later members of the groups.

Further, we find evidence that people look to the broader environment of the interaction when making approach/avoid decisions toward members of these groups. Participants who are faced with a drastic change in a previously predictable group seem to take this as a signal that the environment has changed more broadly, and therefore come to expect changes in the other group even though they do not exist. These results fit with perspectives arguing that a change in reinforcement probabilities leads to increased uncertainty in one's beliefs about the underlying rules of the task, leading to faster learning (Courville et al., 2006). Here, it seems that this change resulted in increased uncertainty around beliefs about both groups, not simply the group that changed. Such an increase in uncertainty about the initially positive group could lead to two possible responses: participants could increase their rates of approach to this group to learn whether they have changed (at the cost of increased perceived risk to the self through approaching) or they could decrease their rates of approach to protect themselves from the increased perceived risk of loss (at the cost of reducing the information they get about this group). In these studies, it seems that participants on average responded in the latter way, perhaps because they were simultaneously trying to gather information about the previously bad group (and did not have the capacity to actively track and update both groups at once).

Finally, we replicate our earlier findings on the self-reinforcing effects of avoidance, providing further support for the idea that avoidance behaviorally reinforces itself but that these effects do not extend to participants' later beliefs about the groups. Thus, these results again provide direct evidence for the self-reinforcement mechanism and are consistent with the ignorance-based mechanism.

Study 4: Manipulating Approach-Contingent Feedback

Studies 1 to 3 provide direct evidence for self-reinforced avoidance, but we have not yet directly tested for effects of ignorance-based avoidance. Thus, Studies 4 and 5 aim to more directly test this mechanism by isolating the role of approach-contingent feedback in belief-updating. Whereas all participants in previous studies received feedback only if they approached an alien, here we manipulated whether participants received feedback only if they approached an alien (approach-contingent feedback condition) or feedback on every trial regardless of whether they approached (full feedback condition). If approach-contingent feedback hinders belief-updating, group biases at the end of the task should be larger for participants in the approach-contingent feedback condition than those in the full feedback condition.

At the beginning of the task in Study 4, the initially cooperative group cooperates 90% of the time and the initially uncooperative group 10% of the time. As in the previous studies, the groups converge over time, such that by the end of the task both groups cooperate at a rate of 60%. Note here that the change in the initially uncooperative group is larger than the change in the initially cooperative group. As such, the aim of this study is not to compare the initially uncooperative group to the initially cooperative group, but to compare approach to each group across the two conditions.

Method

Participants

One hundred seventy-three participants (78 female, 94 male, one did not specify gender; M_{age} = 36.1) were recruited from MTurk to participate in the study. Participants could earn up to $5 depending on how well they did in the task. This sample size provided 80% power to detect effects of at least $b = -.23$ for the two-way interaction of interest for this study.

Procedure

Participants completed a game similar to that in the previous studies, divided into three stages (although the stages were not explicit to participants). In the first stage, the two groups of aliens had extremely different cooperation probabilities—one group cooperated at a rate of .9, while the other cooperated at a rate of .1. This round consisted of 60 trials, in which 20 different aliens were encountered between one and five times each.

In the second stage, each new alien that appeared in the game was slightly more cooperative (in the initially uncooperative group) or less cooperative (in the initially cooperative group), such that by the end of the stage the remaining aliens in the two groups cooperated at the same rate of .6. This stage consisted of 80 trials in which 24 different aliens (eight from the first round and 16 new) were encountered. In the final stage of the game, all aliens that participants encountered had the same rate of cooperation (.6), regardless of group. This stage consisted of 20 trials and 12 different aliens. Thus, the game consisted of 160 trials and 20 aliens in each group. The study consisted of two between-subjects conditions: in the full feedback condition, participants received feedback about the alien's actions regardless of whether they approached the alien or not, while participants in the approach-contingent feedback condition received feedback only if they approached the alien (see Figure 1). As in previous studies, participants then rated each alien's likelihood of cooperating, completed the TIPI (for exploratory purposes as before), and filled out a demographics questionnaire.

Reinforcement Learning Models

As in the previous studies, reinforcement learning models were applied to the data from Study 4 to examine the reinforcement of avoidance. Models for the approach-contingent feedback condition were identical to the models described in Study 1. For the full feedback condition, participants get feedback about the alien's actions even if they choose to avoid. As such, it would not make theoretical sense to model the reinforcement of avoidance in this condition (as there is no imagined negative outcome to reinforce the avoidance behavior, only the real feedback about the outcome that was avoided). Instead, we model the degree to which participants update their beliefs about the group after getting feedback about the outcome they avoided. The rationale for this is that although participants likely update their beliefs according to the feedback they get when avoiding, they may not update to the same degree as when they approach the alien and actually get the outcome (Camerer & Ho, 1999; Yechiam & Busemeyer, 2005, 2006). Therefore, we model value-updating after avoidance in the full feedback condition as:

$$V_{t+1} = V_t + \alpha\lambda(R_t - V_t)$$

where λgoverns the degree to which values are updated after avoidance (with $\lambda = 1$ being updating equal to that after approach).

Results

Initial Learning About Group Differences

We first examined whether participants accurately learned about differences in the members of the two groups encountered in the first half of the task and whether this initial learning differs by condition. Participants approached members of the initially cooperative group more than the initially uncooperative group, $b = -1.84$, $\chi^2(1) = 555.44$, $p < .001$, indicating that they accurately learned about the differences between the two groups. Participants in the approach-contingent feedback condition also approached more overall than those in the full feedback condition, $b = -.19$, $\chi^2(1) = 16.99$, $p < .001$, as would be expected since approach is necessary for learning in the approach-contingent feedback condition but not the full feedback condition. Somewhat unexpectedly, those in the full feedback condition differentiated the groups marginally more than those in the approach-contingent feedback condition as seen in the two-way interaction, $b = -.15$, $\chi^2(1) = 3.69$, $p = .055$. This may result from differential reinforcement of negativity in the two conditions; we return to this point further below.

Generalization of Approach Behavior

To test generalization of this initial learning to later members of the groups in the two conditions, approach behavior to the new members of the two groups (who cooperated at equal rates) was predicted from the alien's group membership, the participant's condition, and the interaction between these two variables. Looking only at the final twelve aliens who cooperated equally across groups, participants approached members of the initially cooperative group more than the initially uncooperative group, $b = -1.09$, $\chi^2(1) = 169.87$, $p < .001$, model $R^2 = .159$. However, the interaction between group membership and condition was not significant, $b = -.14$, $\chi^2(1) = 2.75$, $p = .097$, and in fact trended in the opposite direction to what we expected, with participants in the full feedback condition differentiating the groups in their approach behavior slightly more than participants in the approach-contingent feedback condition.

One potential explanation for this finding lies in the differential reinforcement of negativity experienced by participants in the two conditions. Even after participants in the full feedback condition learn to avoid members of the initially uncooperative group, they still receive feedback trial by trial about the group's behavior. On approximately 80% of initial trials, this feedback was negative, with the initially uncooperative group behaving uncooperatively. Thus, the negativity of the initially uncooperative group for participants in this condition is continually reinforced. In contrast, participants in the approach-contingent feedback condition stop getting feedback about the initially uncooperative group once they decide to avoid the members of this group. Even though the values associated with the initially uncooperative group in this condition are updated slightly after avoidance, this updating is much smaller than it would be if these participants had actually received negative feedback about the group. Thus, participants in the full feedback condition may have had a larger "barrier" of initial negativity to overcome to change their beliefs about the group. We test this idea below using values extracted from the reinforcement learning model.

Reinforcement of Avoidance

As in the previous studies, the reinforcement of avoidance in the approach-contingent feedback condition was examined by comparing models in which values are updated after avoidance to a null model in which values remain constant after avoidance. Reinforcement of avoidance is not modeled for the full feedback condition, as here participants actually receive feedback after avoiding and there is therefore no imagined negative outcome to reinforce the avoidance behavior (see Method). The results of the DIC-based model comparison for participants in the approach-contingent feedback condition indicate that a model with a single avoidance parameter for both groups provides the best fit to the data (see Table 1 for DIC values). However, unlike in the previous studies, the 95% HDI for this avoidance parameter includes 0, suggesting that the parameter is not credibly different from 0 (see Table 2 for parameter values). Together, this provides mixed support for whether values are being negatively updated after avoidance in this study.

Paralleling the previous studies, the model without avoidance reinforcement provides the best prediction of participants' ratings of each alien's cooperativeness at the end of the task (see Table 3), suggesting that the reinforcement of avoidance may be in the form of stimulus–response associations and not generalizable representations of the avoided groups.

Value Comparison Across Conditions

To potentially help explain why those in the full feedback condition differentiated the groups later in the task more than those in the approach-contingent feedback condition, we extracted the values associated with the initially uncooperative group from the best-fitting reinforcement learning model before any of the more cooperative aliens have appeared. If people who received feedback on every trial really experienced more negative reinforcement regarding the initially uncooperative group than those who received feedback only when they approached, this value should be lower for those in the full feedback condition than those in the approach-contingent feedback, indicating that they have a larger barrier of initial negativity to overcome when the group changes. This is indeed what we find: the value of the initially uncooperative group before the groups change is lower in the full feedback condition than the approach-contingent feedback condition, $b = -.11$, $t(171) = -4.20$, $p < .001$ (see Online Supplemental Materials Figure 4), indicating that those in the full feedback condition did indeed have more negative expectations at this point in the task than those in the approach-contingent feedback condition. This may explain why those in the full feedback condition show slightly less updating than those in the approach-contingent feedback condition: the change in the groups simply was not large enough to overcome their greater initial negative beliefs.

Discussion

Study 4 replicates the finding that participants fail to update their initial beliefs once the groups have changed and provides somewhat positive but mixed evidence for the self-reinforcing effects of avoidance on behavior. However, contrary to what would be expected under ignorance-based avoidance, participants did not update their beliefs more in the full feedback condition than the approach-contingent feedback condition. This finding can potentially be explained by the differential reinforcement of negativity experienced by participants in each condition: reinforcement learning models show that even before any change in the groups occurs, those in the full feedback condition have more negative expectations of the initially uncooperative group than those in the partial feedback condition. Thus, it seems that the change in the groups simply was not large enough to overcome the greater negativity that was present in the full feedback condition. Further, some researchers have suggested that receiving selective feedback about outcomes rather than full feedback may actually facilitate learning by increasing the efficiency of the information stored in working memory, as under selective feedback only examples with high value are stored in memory, rather than all examples (Griffiths & Newell, 2007). Such a mechanism may also help to counter the ignorance-based effects of approach-contingent feedback, potentially providing an additional reason for the lack of an advantage for the full feedback condition in this study.

Study 5: Manipulating Approach-Contingent Feedback Under Extreme Group Changes

The results of Study 4 suggest that the change in the groups was too subtle to overcome the greater reinforcement of negativity in the full feedback condition, potentially masking any effects of ignorance-based avoidance. Thus, in addition to manipulating whether feedback was contingent on approach, Study 5 also manipulated whether the initially uncooperative group underwent a moderate change (becoming neutral by the end of the task) or an extreme change (becoming cooperative by the end of the task). Because we are mainly interested in changes in behavior toward initially negative groups, the behavior of the initially cooperative group in this study was kept constant. The study consisted of a 2 (switch condition: moderate change vs. extreme change) × 2 (feedback condition: approach-contingent feedback vs. full feedback) between-subjects design. We expected approach-contingent feedback to hinder belief-updating relative to the full feedback condition when the initially uncooperative group has undergone an extreme change (and not when they have undergone only a moderate change, based on the results of Study 4).

Method

Participants

Four hundred five participants (200 female, 201 male, one other gender, three did not specify gender; M_{age} 35.2) were recruited from MTurk and paid up to $5 for completing the study, based on their performance in the game. This sample size provided 80% power to detect effects of $b = .21$ for the three-way interaction of interest in this study.

Procedure

The study consisted of a 2 (switch condition: extreme change vs. moderate change) × 2 (feedback condition: full feedback vs. approach-contingent feedback) between-subjects design. The task was similar to that of Study 4, with the game divided into three stages characterized by the aliens' cooperation probabilities. In all conditions, aliens in the first stage cooperated at a very high rate (.9) in the initially cooperative group and a very low rate (.1) in the initially uncooperative group. For all participants, the initially cooperative group's cooperation probabilities did not change across the game—members of this group always cooperated at a rate of .9. We manipulated whether the initially uncooperative group ended up as extremely cooperative (cooperation rate of .8; extreme change condition) or more neutral (cooperation rate of .45; moderate change condition). The game consisted of 150 trials in which 38 different aliens were encountered. As in Study 4, we also manipulated whether participants received feedback on every trial (full feedback) or feedback contingent on approach (approach-contingent feedback; see Figure 1). After the game, participants rated how likely each alien was to cooperate and completed the TIPI (for exploratory purposes as in the previous studies) and a demographics questionnaire.

Reinforcement Learning Models

Reinforcement learning models were then applied to the data as in Study 4.

Results

Initial Learning About Group Differences

We first examined whether participants accurately learn about members of the two groups encountered in the first half of the task, whose behavior differs by group membership. As expected, participants approached members of the initially cooperative group more than the initially uncooperative group, $b = -1.57$, $\chi^2(1) = 1124.52$, $p < .001$, and approached more overall in the approach-contingent feedback condition than the full feedback condition, $b = -.16$, $\chi^2(1) = 22.70$, $p < .001$. There was no interaction between group and feedback condition, $b = .007$, $\chi^2(1) =$

.022, p = .88, suggesting that participants in all conditions accurately learned about the differences between the two groups at the beginning of the task.

Generalization of Approach Behavior

The critical question in this study was whether participants in the full feedback condition would update their beliefs more than participants in the approach-contingent feedback condition when the initially uncooperative group changes drastically. Supporting the critical hypothesis of this study, a three-way interaction was found between alien group, feedback condition, and switch condition when looking at approach to the 12 aliens who were encountered last in the task, b = .17, $\chi^2(1)$ = 5.43, p = .020, model R^2 = .232, indicating that participants in the full feedback condition update their beliefs to a greater degree than those in the approach-contingent feedback condition, but only when the initially uncooperative group has undergone an extreme change and is now cooperative (see Figure 4). We also find main effects of the alien's group membership, $b = -1.66$, $\chi^2(1)$ = 514.54, $p < .001$, and two-way interactions of group membership with switch condition, b = .16, $\chi^2(1)$ = 4.740, p = .029, and with feedback condition, b = .23, $\chi^2(1)$ = 10.25, p = .001. Overall, these results provide evidence for the ignorance-based avoidance mechanism, suggesting that when the initially uncooperative group has undergone an extreme change in behavior, approach-contingent feedback hinders participants' abilities to fully update their beliefs about this group.

Reinforcement of Avoidance

To test the self-reinforced avoidance mechanism of stereotype maintenance, reinforcement learning models in which the probability of approaching a group decreases after avoidance were compared with models in which this probability remains unchanged after avoidance, as in the previous studies.[3] As in the previous studies, DIC-based model comparison suggests that a model that assumes reinforcing effects of avoidance provides a better fit to the data than a model that assumes values remain unchanged after avoidance (see Table 1). Unlike the previous studies, a model with separate avoidance parameters for the initially cooperative group and the initially uncooperative group fit the data better than a model with a single avoidance parameter for both groups. Examining the values of these avoidance parameters reveals that values associated with the initially uncooperative group are negatively updated after avoidance (95% HDI of avoidance parameter excludes 0), whereas estimates of the initially cooperative group's avoidance parameter spanned a very large range that includes 0 (see Table 2). This is likely because the initially cooperative group remains very cooperative throughout the entire study, unlike in previous studies where the initially cooperative group changes to become less cooperative. The lack of negative actions from the initially cooperative group (and avoidance behaviors toward this group) may explain why we only find evidence of avoidance reinforcement for the initially uncooperative group in this study. This lack of avoidance behaviors toward the initially cooperative group also results in low effective sample sizes for parameters describing the mean and variance of this group avoidance parameter as well as very large credible intervals; thus, we cannot draw many conclusions about reinforcement of avoidance toward the initially cooperative group in this study. Overall, the results of this analysis replicate our earlier findings on the reinforcement of avoidance for the initially uncooperative group, indicating that at least for aliens in the initially uncooperative group who are avoided often, avoidance makes later avoidance behaviors more likely.

Replicating the previous studies, final values from the model without any reinforcement of avoidance best predicted participants' ratings of how cooperative each alien was (see Table 3), providing tentative support for the idea that avoidance reinforces stimulus–response associations between the presumably negative group and the avoidance behavior without necessarily influencing people's explicit beliefs.

Discussion

Study 5 demonstrates that approach-contingent feedback hinders belief-updating when an initially negative group undergoes a drastic change in behavior: under these conditions, those who receive feedback on every trial start approaching this group more than those who receive feedback only if they approach. As expected based on the results of Study 4, when changes in the group were only moderate, participants did not update their beliefs regardless of the type of feedback they received and initial learning persisted. These results also provide further evidence for the proposal that avoidance has self-reinforcing effects on behavior, with earlier avoidance of the initially uncooperative group increasing the probability of avoiding them later on. Thus, we find evidence for both ignorance-based avoidance and self-reinforced avoidance in maintaining initial behavior toward the groups.

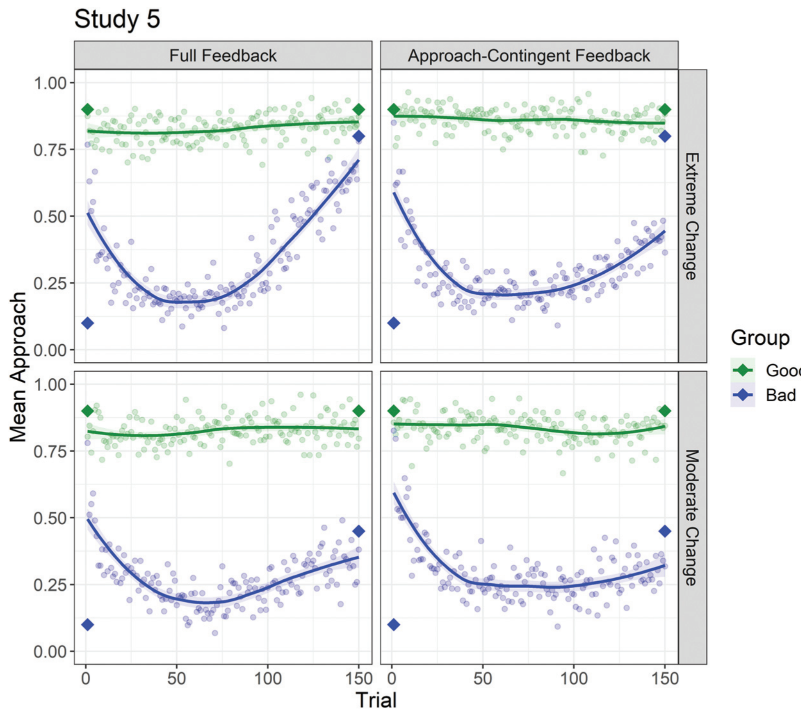
General Discussion

In a world where much of our behavior is social in nature, accurate impressions of other people are vital to making good decisions. The existence of inaccurate stereotypes therefore presents a puzzle: not only can these stereotypes be harmful for the targets, they can in fact decrease the perceiver's ability to make good social decisions. Why, then, do even people who report egalitarian motivations still sometimes possess these inaccurate beliefs? The findings presented here suggest that asymmetries in information-sampling provide one reason for the persistence of these inaccurate stereotypes.

Overall, the results of five studies provide evidence that people often fail to adequately update their initially formed beliefs about social groups, even when these initial beliefs are no longer accurate. Across five studies, participants rely heavily on their initial experiences when changes in the groups are small, exploiting their existing beliefs about the two groups rather than exploring new individuals. When the groups change more drastically (Studies 3 and 5), participants do update their beliefs more, though behavior is still largely biased by their initial experiences. In addition to finding further support for the role of approach-contingent feedback in preventing belief-updating through ignorance-based avoidance (Fazio et al., 2004), we present evidence that avoidance has

[3] Models in which the learning rate and temperature group parameters were separated by the two switch conditions (extreme change vs. moderate change) were also run, but no differences in these parameters were found between the conditions. Thus, we report results of models with a single group parameter for all conditions.

Figure 4
Approach Behavior Over Trials in Study 5



Note. Dots and smoothed lines represent average participant response, colored by group, and diamonds represent each group's actual average probability of cooperating at the beginning and end of the experiment in each of the four conditions. See the online article for the color version of this figure.

direct self-reinforcing effects on behavior, leading to the perpetuation of avoidance behaviors toward groups when negative members are encountered first.

Building on previous work demonstrating the learning asymmetries that arise from approach-contingent feedback (Eiser et al., 2007; Fazio et al., 2004, 2015), the results of Studies 4 and 5 help to clarify the role of this type of feedback in stereotype maintenance. Our findings suggest that approach-contingent feedback plays the largest role in preventing belief-updating when initial experiences are highly unrepresentative of later group members (when the group's behavior has changed drastically). When initial experiences are only moderately unrepresentative (the group has changed only moderately), on the other hand, people fail to update their beliefs regardless of the type of feedback they receive. Under these more moderate changes, the stronger reinforcement of negativity in the full feedback conditions may counteract any advantages that this type of feedback provides in belief-updating, producing similar failures to update as approach-contingent feedback. This failure to update in the face of moderately discrepant information mirrors work on motivated reasoning, which suggests that prior beliefs will shape the interpretation of new information as long as this new information is not too highly discrepant (Braman & Nelson, 2007; Druckman, 2012; Kunda, 1990). Indeed, while people do generally update their beliefs somewhat after the groups change, some persistence of initial behavior is apparent for both the initially cooperative group and the initially uncooperative group, reflecting the more general power of initial information in shaping subsequent learning. This bias toward initial behaviors may be especially prevalent when the groups change only moderately, as no strong signals have been provided that initial beliefs are no longer relevant. Thus, approach-contingent feedback seems to play the largest role when changes in the information source are large enough to overcome any biases toward prior beliefs, highlighting the fact that ignorance-based avoidance is one of multiple mechanisms preventing people from updating their initial beliefs about social groups. When changes in a group's composition or behavior are great enough to overcome these other barriers to belief-updating, avoidance still facilitates the persistence of initial beliefs by removing the opportunity for feedback.

In addition to preventing the gain of new information, these findings support the proposal that avoidance has direct self-reinforcing effects on an individual's behavior toward group members. Applying reinforcement learning models to participants' behavioral data indicates that avoiding a supposedly negative person increases the probability of future avoidance behaviors toward that person's group, perhaps by eliciting relief at having evaded a negative outcome. Further, these effects of avoidance reinforcement persisted despite the inclusion of a general choice perseveration mechanism in the model, suggesting that this more general mechanism does not account for our effects. These findings parallel other research demonstrating the potentially powerful effects of initial behavioral responses on later choices (Salomon et al., 2018;

Schonberg et al., 2014) and point to the potential for our behaviors in intergroup contexts to be self-reinforcing even without our awareness. Further, given the potential for social transmission of behaviors toward members of other groups (Fazio et al., 2004), deciding to avoid someone may have compounding effects on intergroup behaviors both for the initial perceiver and for close others. Thus, an initial decision to avoid someone, while seemingly innocuous, may in fact begin a cycle of avoidance that reinforces itself.

The self-reinforcing effects of avoidance on behavior could be accounted for by two potential mechanisms: avoidance may update people's representations of the groups in question or it may simply reinforce the stimulus–response associations between the group and the avoidance behavior without affecting explicit beliefs. Although the current studies cannot fully dissociate these possibilities behaviorally, the data suggest that the reinforcement may occur at the level of stimulus–response associations and not extend to participants' explicitly held beliefs about the groups. When asked at the end of the task to rate each alien's cooperativeness from 0 to 100, participants' explicit ratings are best predicted by models which posit no updating of values after avoidance. Although explicit negative beliefs were clearly present in our findings, the fact that avoidance reinforcement did not seem to contribute to these beliefs despite contributing to behavior may point to a potential distinction between factors contributing to behavioral discrimination and explicit attitudes or beliefs about a group. Such a distinction would fit with models incorporating the influence of multiple automatic and controlled processes on behavior (Sherman et al., 2008), as well as with more recent perspectives defining implicit bias as a primarily behavioral phenomenon (De Houwer, 2019) in which behavior is influenced by cues about others' social group membership without necessarily requiring a role for biased cognitive or affective representations. In fact, processes that affect behavior without affecting explicitly endorsed attitudes may be some of the most difficult barriers to overcome in reducing discrimination, as one must first draw attention to the biased behavior and convince the perceiver that their actions may not reflect their consciously reported beliefs. Future research should more directly test the differential effects of self-reinforced avoidance on behavior and beliefs and explore potential mechanisms underlying this dissociation. For example, expressing an explicit belief may engage corrective processes that a simple behavioral approach/avoidance decision does not, especially since the behavior has direct consequences for participants' monetary outcomes whereas the ratings do not.

This research outlines how basic learning processes of approach and avoidance may contribute to the perpetuation of stereotyping and discrimination. Critically, we do not argue that the processes described here are unique to social learning; rather, that the same processes that are at play in learning about nonsocial objects may function in the social domain, contributing to the widespread phenomena of stereotyping and discrimination. Of course, the social world is very complex and our behavioral paradigm is necessarily simple to isolate the processes of interest. To better generalize these processes to the real world, it will be necessary to consider a variety of more complex factors such as different numbers and types of groups, varied behaviors that go beyond a simple approach/avoid or positive/negative dichotomy, and historical and systemic power differentials between groups. In addition to these factors, another critical way in which these processes may differ in the social and nonsocial domains is the reciprocal nature of impression formation. In particular, when forming impressions of another person, that person is often simultaneously forming an impression of you too. The impression they form and their resulting behavior toward you depends critically on how you behave toward them, potentially creating feedback cycles of avoidance and negative impression formation.

Although the groups in our studies initially have real differences in behavior, these supposed differences in the perceiver's experience can arise even in the absence of any actual differences between the groups. One way in which this may occur in the social world is through biases in one's sources of information about groups. For example, biases both in the media (e.g., Hassan et al., 2017; Mahony, 2010) and in individual cognitive processes (e.g., Howard & Rothbart, 1980; Sherman et al., 1998; Ybarra et al., 2000) mean that we are often presented with disproportionately negative examples of marginalized groups or outgroups, whereas the same negative behavior in dominant groups or ingroups is downplayed or excused. These biases in information sources may create inaccurate stereotypes and perceptions of marginalized groups that are then maintained through the avoidance processes described here. While we simulate such biases in this work by having participants first meet only negative members of one of the groups, future work should incorporate more complex forms of information biases that better encapsulate the nature of information transmission about social groups. Unrepresentative initial experiences with a group may also arise simply through random variation in encounters. In particular, social groups in the real world are heterogeneous, such that no one individual is representative of the entire group. As perceivers do not meet an entire group simultaneously but must instead extrapolate from their knowledge of individual members to form impressions of the group, this within-group variability ensures that no two perceivers will have the same experience of the group. Someone who initially meets a few negative group members just by chance may decide that the whole group is negative and should be avoided in the future. Someone who initially meets members who are more positive, on the other hand, may be more likely to continue interacting with new members of the group who are encountered. Even if this person later encounters negative members of the group, they can draw on their initial positive experiences to ensure that the negative encounters are not generalized to the group as a whole. These findings therefore suggest that random variation in the order in which one encounters members of a new group may have a lasting influence on the impressions that are formed of the group even later on. This is not simply because of cognitive overweighting of initial information at the expense of later experiences (although such overweighting likely plays a role as well; Freund et al., 1985; Webster et al., 1996), but because that initial information actually shapes the experiences that one has later on. A few unrepresentative negative encounters with a new group may be enough to foster future avoidance of the group as a whole, which in turn prevents the incorrect negative impression from ever being updated. Such processes may play a role in maintaining inaccurate negative stereotypes about outgroup members, especially when the perceiver has not had much contact with the group in the past.

If unrepresentative initial experiences can create negative stereotypes that are then maintained through avoidance, one avenue for reducing such stereotypes may be to focus on the ways in which people acquire information about their social worlds. In particular, it may be useful to focus on forms of information gain that are not directly contingent on approaching a target. The literature on indirect forms of intergroup contact has demonstrated this principle, suggesting that prejudice can be reduced through observational learning through mass media, virtual contact over the Internet, and extended contact through members of one's ingroup (see Dovidio et al., 2017 for a review). Given the current findings, these methods of indirect contact may be particularly successful in situations where direct contact poses a perceived risk to the self (regardless of the accuracy of this belief). In such situations, the opportunity to learn about someone despite misplaced avoidance behaviors can help perceivers break out of the feedback cycle of self-perpetuating avoidance, allowing them to receive information that challenges their beliefs. By targeting the mechanisms through which people learn about others, these approaches may successfully reduce biased behaviors and beliefs without actually targeting beliefs directly.

Furthermore, the findings of Study 3 suggest that people's higher-order contextual beliefs about the stability or homogeneity of the groups in the current environment will influence their exploration-exploitation trade-offs in social contexts. When one group's behavior in this study began to change, participants seemed to take this as a signal that the other group would also soon change, and adjusted their behavior accordingly. Rather than simply updating their local beliefs about the group that actually changed, participants seemed to at least slightly update their more global beliefs about the environment of the social interactions in general, taking the change in one group as a signal that the other group may start behaving differently as well. We find evidence for this mechanism in novel groups, but to ensure that these effects generalize beyond these learning tasks it will be important to extend this research to our understanding of group processes in real-world settings, where these social learning processes are embedded within cultural stereotypes and motivated group cognitions. Supporting the potential for generalizability, these findings parallel work showing that highlighting the malleable nature of traits through "incremental" theories of personality can reduce negative attitudes and increase motivations to interact with outgroups (Halperin et al., 2012; Levontin et al., 2013). Thus, targeting these higher-order beliefs about groups (e.g., by showing that a presumed positive group also has some negative members or can change over time) may be another fruitful avenue for reducing these stereotypes.

The role of avoidance in maintaining potentially inaccurate stereotypes, both indirectly by limiting further information-gain and directly by reinforcing future avoidance behaviors, may provide an explanation for how even well-intentioned perceivers can develop inaccurate beliefs. In particular, even if perceivers develop rational beliefs given their experiences, their experiences themselves may be biased, leading to inaccurate perceptions. That is, an individual's beliefs about a group may accurately represent the experiences they have had with that group while still failing to accurately represent the nature of the group more generally. Because attitudes are formed from experience, a perceiver who wishes to form accurate beliefs must be aware that their experiences may be biased even without their awareness. To ensure they are truly treating others fairly, the perceiver must be aware not only of whether their behaviors accurately reflect their experiences, but also of how their behaviors actively shape the experiences that they have.

Context of the Research

The current work is part of a general program of research that aims to examine how basic cognitive processes might combine with features of the social environment to produce pervasive biases and prejudices. Our general aim is to examine how interactions between low-level cognitions and the higher-level societal context give rise to commonly seen beliefs and behaviors. In the current work, this takes the form of examining how basic processes of approach and avoidance learning can combine with contextual features like the structure of information-gain and biases in initial information (such as those that might arise through the media) to produce lasting negative stereotypes. In future work, we hope to bring additional social and interpersonal aspects to this research by examining the reciprocal nature of impression formation. In particular, when people are simultaneously learning about one another and updating their beliefs in response to the other's behavior, feedback cycles of avoidance may occur not only within individuals, but between groups.

References

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Behrens, T. E. J., Hunt, L. T., & Rushworth, M. F. S. (2009). The computation of social behavior. *Science*, *324*(5931), 1160–1164. https://doi.org/10.1126/science.1169694

Bem, D. (1972). Self-perception theory. *Advances in Experimental Social Psychology*, *6*, 1–62. https://doi.org/10.1016/S0065-2601(08)60024-6

Boorman, E. D., Behrens, T. E. J., Woolrich, M. W., & Rushworth, M. F. S. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron*, *62*(5), 733–743. https://doi.org/10.1016/j.neuron.2009.05.014

Braman, E., & Nelson, T. E. (2007). Mechanism of motivated reasoning? Analogical perception in discrimination disputes. *American Journal of Political Science*, *51*(4), 940–956. https://doi.org/10.1111/j.1540-5907.2007.00290.x

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love or outgroup hate? *Journal of Social Issues*, *55*(3), 429–444. https://doi.org/10.1111/0022-4537.00126

Camerer, C., & Ho, T.-H. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, *67*(4), 827–874. https://doi.org/10.1111/1468-0262.00054

Cavanagh, J. F. (2015). Cortical delta activity reflects reward prediction error and related behavioral adjustments, but at different times. *NeuroImage*, *110*, 205–216. https://doi.org/10.1016/j.neuroimage.2015.02.007

Chambliss, W. J. (1965). The selection of friends. *Social Forces*, *43*(3), 370–380. https://doi.org/10.2307/2574767

Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin*, *25*(2), 215–224. https://doi.org/10.1177/0146167299025002007

Collins, A. G. E., & Frank, M. J. (2016). Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. *Cognition*, *152*, 160–169. https://doi.org/10.1016/j.cognition.2016.04.002

Collins, E. C., Percy, E. J., Smith, E. R., & Kruschke, J. K. (2011). Integrating advice and experience: Learning and decision making with social and nonsocial cues. *Journal of Personality and Social Psychology*, *100*(6), 967–982. https://doi.org/10.1037/a0022982

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294–300. https://doi.org/10.1016/j.tics.2006.05.004

Davidson, R., & MacKinnon, J. G. (2004). *Econometric theory and methods* (5th ed.). Oxford University Press.

De Houwer, J. (2019). Implicit bias is behavior: A functional-cognitive perspective on implicit bias. *Perspectives on Psychological Science*, *14*(5), 835–840. https://doi.org/10.1177/1745691619855638

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, *8*(11), 1611–1618. https://doi.org/10.1038/nn1575

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review*, *112*(4), 951–978. https://doi.org/10.1037/0033-295X.112.4.951

Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science*, *12*(5), 523–538. https://doi.org/10.1287/orsc.12.5.523.10092

Dovidio, J. F., Love, A., Schellhaas, F. M. H., & Hewstone, M. (2017). Reducing intergroup bias through intergroup contact: Twenty years of progress and future directions. *Group Processes & Intergroup Relations*, *20*(5), 606–620. https://doi.org/10.1177/1368430217712052

Druckman, J. N. (2012). The politics of motivation. *Critical Review*, *24*(2), 199–216. https://doi.org/10.1080/08913811.2012.711022

Eiser, J. R., Shook, N. J., & Fazio, R. H. (2007). Attitude learning through exploration: Advice and strategy appraisals. *European Journal of Social Psychology*, *37*(5), 1046–1056. https://doi.org/10.1002/ejsp.419

Elliot, A. J., & Covington, M. V. (2001). Approach and avoidance motivation. *Educational Psychology Review*, *13*(2), 73–92. https://doi.org/10.1023/A:1009009018235

Elwin, E., Juslin, P., Olsson, H., & Enkvist, T. (2007). Constructivist coding: Learning from selective feedback. *Psychological Science*, *18*(2), 105–110. https://doi.org/10.1111/j.1467-9280.2007.01856.x

Engelhard, I. M., van Uijen, S. L., van Seters, N., & Velu, N. (2015). The effects of safety behavior directed towards a safety cue on perceptions of threat. *Behavior Therapy*, *46*(5), 604–610. https://doi.org/10.1016/j.beth.2014.12.006

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology*, *87*(3), 293–311. https://doi.org/10.1037/0022-3514.87.3.293

Fazio, R. H., Pietri, E. S., Rocklage, M. D., & Shook, N. J. (2015). Positive versus negative valence: Asymmetries in attitude formation and generalization as fundamental individual differences. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 51, pp. 97–146). Academic Press. https://doi.org/10.1016/bs.aesp.2014.09.002

Fazio, R. H., Zanna, M. P., & Cooper, J. (1977). Dissonance and self-perception: An integrative view of each theory's proper domain of application. *Journal of Experimental Social Psychology*, *13*(5), 464–479. https://doi.org/10.1016/0022-1031(77)90031-2

FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(7), E1690–E1697. https://doi.org/10.1073/pnas.1715227115

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage. https://doi.org/10.1177/0049124105277200

Freund, T., Kruglanski, A., & Shpitzajzen, A. (1985). The freezing and unfreezing of impressional primacy: Effects of the need for structure and the fear of invalidity. *Personality and Social Psychology Bulletin*, *11*(4), 479–487. https://doi.org/10.1177/0146167285114013

Green, P., & Macleod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Griffiths, O., & Newell, B. R. (2007). The impact of complete and selective feedback in static and dynamic multiple-cue judgment tasks. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (pp. 2884–2889). Cognitive Science Society.

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

Halperin, E., Crisp, R. J., Husnu, S., Trzesniewski, K. H., Dweck, C. S., & Gross, J. J. (2012). Promoting intergroup contact by changing beliefs: Group malleability, intergroup anxiety, and contact motivation. *Emotion*, *12*(6), 1192–1195. https://doi.org/10.1037/a0028620

Hamilton, D. L. (1979). A cognitive-attributional analysis of stereotyping. *Advances in Experimental Social Psychology*, *12*(C), 53–84. https://doi.org/10.1016/S0065-2601(08)60259-2

Hassan, I., Azmi, M. N. L., & Abubakar, U. I. (2017). The use of terminology in reporting Islam: A comparative analysis. *International Journal of English Linguistics*, *7*(6), 236. https://doi.org/10.5539/ijel.v7n6p236

Henriksson, M. P., Elwin, E., & Juslin, P. (2010). What is coded into memory in the absence of outcome feedback? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 1–16. https://doi.org/10.1037/a0017893

Hirschfeld, L. A. (1995). Do children have a theory of race? *Cognition*, *54*(2), 209–252. https://doi.org/10.1016/0010-0277(95)91425-R

Hofmann, S. G., & Hay, A. C. (2018). Rethinking avoidance: Toward a balanced approach to avoidance in treating anxiety disorders. *Journal of Anxiety Disorders*, *55*, 14–21. https://doi.org/10.1016/j.janxdis.2018.03.004

Howard, J. W., & Rothbart, M. (1980). Social categorization and memory for in-group and out-group behavior. *Journal of Personality and Social Psychology*, *38*(2), 301–310. https://doi.org/10.1037/0022-3514.38.2.301

Jaeger, B. (2017). r2glmm: Computes R squared for mixed (multilevel) models (R package version 0.1.2) [Computer software]. https://cran.r-project.org/package=r2glmm

Jones, R. M., Somerville, L. H., Li, J., Ruberry, E. J., Libby, V., Glover, G., Voss, H. U., Ballon, D. J., & Casey, B. J. (2011). Behavioral and neural properties of social reinforcement learning. *The Journal of Neuroscience*, *31*(37), 13039–13045. https://doi.org/10.1523/JNEUROSCI.2972-11.2011

Keller, N. E., Hennings, A. C., & Dunsmoor, J. E. (2020). Behavioral and neural processes in counterconditioning: Past and future directions. *Behaviour Research and Therapy*, *125*, 103532. https://doi.org/10.1016/j.brat.2019.103532

Kim, H., Shimojo, S., & O'Doherty, J. P. (2006). Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. *PLoS Biology*, *4*(8), e233–1461. https://doi.org/10.1371/journal.pbio.0040233

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science*, *308*(5718), 78–83. https://doi.org/10.1126/science.1108062

Kocsor, F., & Bereczkei, T. (2017). First impressions of strangers rely on generalization of behavioral traits associated with previously seen facial features. *Current Psychology*, *36*(3), 385–391. https://doi.org/10.1007/s12144-016-9427-1

Konovalova, E., & Le Mens, G. (2020). An information sampling explanation for the in-group heterogeneity effect. *Psychological Review*, *127*(1), 47–73. https://doi.org/10.1037/rev0000160

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Kunda, Z., & Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin*, *129*(4), 522–544. https://doi.org/10.1037/0033-2909.129.4.522

LeDoux, J. E., Moscarello, J., Sears, R., & Campese, V. (2017). The birth, death and resurrection of avoidance: A reconceptualization of a troubled paradigm. *Molecular Psychiatry*, *22*(1), 24–36. https://doi.org/10.1038/mp.2016.166

Levin, S., van Laar, C., & Sidanius, J. (2003). The effects of ingroup and outgroup friendships on ethnic attitudes in college: A longitudinal study. *Group Processes & Intergroup Relations*, *6*(1), 76–92. https://doi.org/10.1177/1368430203006001013

Levontin, L., Halperin, E., & Dweck, C. S. (2013). Implicit theories block negative attributions about a longstanding adversary: The case of Israelis and Arabs. *Journal of Experimental Social Psychology*, *49*(4), 670–675. https://doi.org/10.1016/j.jesp.2013.02.002

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*(1), 37–47. https://doi.org/10.1037/0022-3514.66.1.37

Mahony, I. (2010). Diverging frames: A comparison of Indonesian and Australian press portrayals of terrorism and Islamic groups in Indonesia. *The International Communication Gazette*, *72*(8), 739–758. https://doi.org/10.1177/1748048510380813

Maia, T. V. (2010). Two-factor theory, the actor-critic model, and conditioned avoidance. *Learning & Behavior*, *38*(1), 50–67. https://doi.org/10.3758/LB.38.1.50

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *The Journal of Neuroscience*, *35*(21), 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, *6*, (1), 8096. https://doi.org/10.1038/ncomms9096

Plant, E. A. (2004). Responses to interracial interactions over time. *Personality and Social Psychology Bulletin*, *30*(11), 1458–1471. https://doi.org/10.1177/0146167204264244

Plant, E. A., & Devine, P. G. (2003). The antecedents and implications of interracial anxiety. *Personality and Social Psychology Bulletin*, *29*(6), 790–801. https://doi.org/10.1177/0146167203029006011

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, *124*(125), 1–8. http://www.ci.tuwien.ac.at/Conferences/DSC-2003/

Ramasubramanian, S. (2011). The impact of stereotypical versus counter-stereotypical media exemplars on racial attitudes, causal attributions, and support for affirmative action. *Communication Research*, *38*(4), 497–516. https://doi.org/10.1177/0093650210384854

Rothbart, M., & Taylor, M. (1992). Category labels and social reality: Do we view social categories as natural kinds? In G. R. Semin & K. Fiedler (Eds.), *Language, interaction and social cognition* (pp. 11–36). Sage Publications, Inc. https://psycnet.apa.org/record/1992-97980-001

Salomon, T., Botvinik-Nezer, R., Gutentag, T., Gera, R., Iwanir, R., Tamir, M., & Schonberg, T. (2018). The cue-approach task as a general mechanism for long-term non-reinforced behavioral change. *Scientific Reports*, *8*(1), 1–13. https://doi.org/10.1038/s41598-018-21774-3

Schonberg, T., Bakkour, A., Hover, A. M., Mumford, J. A., Nagar, L., Perez, J., & Poldrack, R. A. (2014). Changing value through cued approach. *Nature Neuroscience*, *17*(4), 625–630. https://doi.org/10.1038/nn.3673

Seymour, B., Daw, N. D., Roiser, J. P., Dayan, P., & Dolan, R. (2012). Serotonin selectively modulates reward value in human decision-making. *The Journal of Neuroscience*, *32*(17), 5833–5842. https://doi.org/10.1523/JNEUROSCI.0053-12.2012

Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & Groom, C. J. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review*, *115*(2), 314–335. https://doi.org/10.1037/0033-295X.115.2.314

Sherman, J. W., Klein, S. B., Laskey, A., & Wyer, N. A. (1998). Intergroup bias in group judgment processes: The role of behavioral memories. *Journal of Experimental Social Psychology*, *34*(1), 51–65. https://doi.org/10.1006/jesp.1997.1342

Shook, N. J., & Fazio, R. H. (2009). Political ideology, exploration of novel stimuli, and attitude formation. *Journal of Experimental Social Psychology*, *45*(4), 995–998. https://doi.org/10.1016/j.jesp.2009.04.003

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, *2*(10), 750–756. https://doi.org/10.1038/s41562-018-0425-1

Smith, E. R., & Collins, E. C. (2009). Contextualizing person perception: Distributed social cognition. *Psychological Review*, *116*(2), 343–364. https://doi.org/10.1037/a0015072

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. Statistical Methodology*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Stark, T. H., Flache, A., & Veenstra, R. (2013). Generalization of positive and negative attitudes toward individuals to outgroup attitudes. *Personality and Social Psychology Bulletin*, *39*(5), 608–622. https://doi.org/10.1177/0146167213480890

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Brooks/Cole.

Taylor, D., Altman, I., & Sorrentino, R. (1969). Interpersonal exchange as a function of rewards and costs and situational factors: Expectancy confirmation-disconfirmation. *Journal of Experimental Social Psychology*, *5*(3), 324–339. https://doi.org/10.1016/0022-1031(69)90057-2

van den Hout, M., Gangemi, A., Mancini, F., Engelhard, I. M., Rijkeboer, M. M., van Dam, M., van Dams, M., & Klugkist, I. (2014). Behavior as information about threat in anxiety disorders: A comparison of patients with anxiety disorders and non-anxious controls. *Journal of Behavior Therapy and Experimental Psychiatry*, *45*(4), 489–495. https://doi.org/10.1016/j.jbtep.2014.07.002

Van Oudenhoven, J. P., Groenewoud, J. T., & Hewstone, M. (1996). Cooperation, ethnic salience and generalization of interethnic attitudes. *European Journal of Social Psychology*, *26*(4), 649–661. https://doi.org/10.1002/(SICI)1099-0992(199607)26:4<649::AID-EJSP780>3.0.CO;2-T

van Uijen, S. L., & Toffolo, M. B. J. (2015). Safety behavior increases obsession-related cognitions about the severity of threat. *Behavior Therapy*, *46*(4), 521–531. https://doi.org/10.1016/j.beth.2015.04.001

van Uijen, S. L., Leer, A., & Engelhard, I. M. (2018). Safety behavior after extinction triggers a return of threat expectancy. *Behavior Therapy*, *49*(3), 450–458. https://doi.org/10.1016/j.beth.2017.08.005

Vervliet, B., & Indekeu, E. (2015). Low-cost avoidance behaviors are resistant to fear extinction in humans. *Frontiers in Behavioral Neuroscience*, *9*, 351–312. https://doi.org/10.3389/fnbeh.2015.00351

Webster, D. M., Richter, L., & Kruglanski, A. W. (1996). On leaping to conclusions when feeling tired: Mental fatigue effects on impressional primacy. *Journal of Experimental Social Psychology*, *32*(2), 181–195. https://doi.org/10.1006/jesp.1996.0009

Worthy, D. A., Pang, B., & Byrne, K. A. (2013). Decomposing the roles of perseveration and expected value representation in models of the Iowa gambling task. *Frontiers in Psychology*, *4*, 640–649. https://doi.org/10.3389/fpsyg.2013.00640

Ybarra, O., Stephan, W. G., & Schaberg, L. (2000). Misanthropic memory for the behavior of group members. *Personality and Social Psychology Bulletin*, *26*(12), 1515–1525. https://doi.org/10.1177/01461672002612006

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin & Review*, *12*(3), 387–402. https://doi.org/10.3758/BF03193783

Yechiam, E., & Busemeyer, J. R. (2006). The effect of foregone payoffs on underweighting small probability events. *Journal of Behavioral Decision Making*, *19*(1), 1–16. https://doi.org/10.1002/bdm.509

Received August 12, 2020
Revision received December 9, 2020
Accepted December 11, 2020 ■