# Experience shapes the granularity of social perception: Computational insights into individual and group-based representations

Suraiya Allidina[1], Michael L. Mack[2], & William A. Cunningham[2]

[1] The Ohio State University

[2] University of Toronto

*Authors' Note*. This manuscript is currently in press at the *Journal of Experimental Psychology: General*. We thank members of the Social Cognitive Science lab for feedback on this work. This work was funded by Social Sciences and Humanities Research Council (SSHRC) grant 506547 to WAC and a SSHRC Doctoral Fellowship to SA. Data and analysis scripts for these studies is available at https://osf.io/uj9e7/?view_only=dad674b9e2734eec9758bd911c336585.

Abstract

People are regularly conceptualized at varying levels of resolution, sometimes characterized by their idiosyncratic features while at other times seen as mere tokens of their social groups. Decades of research have sought to understand when perceivers will draw upon each of these types of representations, detailing the perceiver- and target-related features that may decrease reliance on stereotypes in favor of individuated knowledge. However, little work has examined how these representations might be formed in the first place: in order for individuated representations of others to be used, they must first be built through experience. Here, we offer a novel approach to characterizing the formation of social representations through the use of computational models of category learning. Across three experiments, participants learned about members of novel social groups who behaved positively or negatively towards them. Computational modeling of participants' task behavior revealed a critical interaction of perceiver motivations and learning context on representations. Participants who received selective feedback about targets only upon approaching them formed more categorical representations than those who received full feedback. Further, we found tentative evidence that this difference was most pronounced in those who held more racist attitudes, measured in an entirely separate context. Thus, more informative learning contexts could potentially act as a "protective factor" that shields perceivers' representations from their negative attitudes. The results shed light on the psychological underpinnings of prejudice, using a novel approach to reveal how social categorization is selectively employed in a manner that maintains negative stereotypes.

Keywords: social categorization, individuation, approach-contingent feedback, stereotypes, computational modeling

## Public Significance Statement

Little is known about how people form mental representations of others as individuals or as interchangeable members of a group. Using experiments and mathematical models, we found that the availability of information about others shaped these mental representations: when information could only be gained by approaching someone, people were more likely to think of others as interchangeable group members. Further, people with more prejudiced attitudes were potentially more likely to form these group-based representations, but only when information was limited.

People can be characterized across different levels of resolution: we may craft detailed impressions of friends and familiar others while relying on abstracted generalizations to understand strangers. Decades of research have sought to understand when we will draw upon more specific or general knowledge to conceptualize others (Bodenhausen et al., 1999; Brewer et al., 1995; Fiske & Neuberg, 1990; Gawronski et al., 2003; Köpetz & Kruglanski, 2008; Macrae & Bodenhausen, 2000), in search of the features that may decrease reliance on stereotyped knowledge in favor of individual representations. Such research often concludes that the perceiver must have both the motivation and opportunity in the moment of judgment to individuate the target, failing which they will rely on stereotypes. Research that focuses on biased retrieval processes may assume that perceivers have both individuated and stereotyped representations available to draw upon for a given target, and simply must choose which to rely on. However, in order for such representations to be used in understanding others, they must first be built through experience and information search (e.g., Neuberg & Fiske, 1987). This presents another explanation for failures to individuate: a perceiver with both the motivation and cognitive resources to individuate in the moment may simply have no sufficiently individuated representation of the target to draw upon. In this paper, we explore this idea by examining the formation of social representations for novel others based on their group memberships or individuating features. We propose central roles for both the learning environment and the perceiver's attitudes and motivations in creating these basic building blocks of social impression formation that can then be employed for judgment.

Research over the past few decades has sought to understand when people will rely on existing group-based representations at the expense of individuated representations, finding that both individual and environmental factors play a role: some people categorize more than others, and some contexts elicit more categorization than others. Work on perceiver-related factors has found that the individual's motivations, their attentional capacity, and the strength of their stereotypes can shape their reliance on group- or person-related information

(Bodenhausen et al., 1999; Fiske & Neuberg, 1990; Gawronski et al., 2003; Macrae & Bodenhausen, 2000). Complementing this are features of the social information itself, including the alignment of person- and group-related information, the relative accessibility and applicability of the information, and the meaningfulness of the category dimension (Bodenhausen et al., 1999; Brewer et al., 1995; Fiske & Neuberg, 1990; Köpetz & Kruglanski, 2008).

However, implicit in many of these investigations is the idea that the perceiver in the moment of judgment has a variety of potential representations to draw upon, with the core differentiating feature being the choice to employ a representation at a more general level of analysis (i.e., a stereotype about a category) or a more particular one (i.e., an individuating feature). In contrast, in order to draw upon existing representations of a target to form a judgment, the perceiver must first build these representations, often through experience. If the perceiver's experiences have led them to form only shallow representations of a group of targets, motivation and opportunity at the moment of judgment may be insufficient to produce individuation. The initial formation of such representations has been a relatively neglected topic of study, due in part to methodological challenges that can be overcome through the use of computational models of cognition. We employ such models here to explore how social representations are formed through learning, focusing on variation in the creation of group- or person-based representations.

With social learning at the forefront of category development, a key factor determining the formation of representations becomes the availability of social information. Social contexts vary in how they provide people with information about others, with many contexts providing feedback about targets only if the perceiver chooses to approach or interact with them. Learning about someone in these contexts is contingent on approach, engendering some risk for the perceiver if the interaction goes badly, and no learning takes place if the perceiver instead chooses to avoid. This kind of feedback structure perpetuates negative beliefs and stereotypes,

selectively preventing them from being updated (Allidina & Cunningham, 2021; Bai et al., 2022; Denrell & March, 2001; Eiser et al., 2007; Fazio et al., 2004; Konovalova & Le Mens, 2017; Rich & Gureckis, 2018). Previous research, for example, used artificial alien groups to demonstrate that avoidance under conditions of approach-contingent feedback reinforces negative beliefs, both by preventing further information gain and by directly reinforcing negativity (Allidina & Cunningham, 2021). However, it is unclear how processes of social categorization were specifically affected in these studies; approach-contingent feedback may have simply prevented negative group-based beliefs from being updated or may have actually changed the representation structure that participants held for the aliens. In the latter case, participants in the approach-contingent feedback condition may have actually updated their representational structure for the aliens, representing each alien individually (and thus better representing the within-group variation in cooperation rates) rather than only conceptualizing the overall groups. The feedback structure of the social context may thus shape not just people's beliefs about others, but the very structure of their cognitive representations.

In addition to shaping attention to groups or individuals in general, social feedback may serve a further role in "setting the stage" for a perceiver's attitudes, ideologies, or motivations to act. In particular, the structure of feedback in social learning may provide a context of limited information in which individual differences in factors such as prejudice can emerge. As a relatively resource-intensive process, forming person-based representations may require the perceiver to have both the motivation and opportunity to do so (Fazio, 1990). In an attempt to reconcile motivational and cognitive accounts of prejudice, Stangor and Ford (1992) propose that motivations shape intergroup attitudes by pushing perceivers to seek out information that confirms their initial expectations about others. Thus, motivational and attitudinal individual differences should critically interact with the affordances provided by the social environment: contexts in which selectively seeking out expectancy-confirming information is easy should allow for a greater role for these individual differences. Initial evidence for this idea comes from

Ditonto (2019, 2020), who found that people high in racial or gender-based prejudice seek out less information about stigmatized candidates than candidates in dominant groups. Approach-contingent feedback may therefore provide a context in which individual variation in prejudiced attitudes or social motivations play a greater role in the formation of social representations. We explore this as a secondary question in the present studies.

As examining the formation of social representations requires methods beyond simply extrapolating from people's beliefs about a target, we turn to computational models of category formation to more directly assess the representations participants are forming as they learn about others. Models of social learning have often been leveraged within the social domain to understand how attitudes and beliefs about groups form over time, with researchers formally modeling the reinforcement processes by which different groups come to be associated with positive or negative outcomes over time (e.g., Allidina & Cunningham, 2021; Hackel et al., 2015, 2022; Lindström et al., 2014; Schultner et al., 2024; Spiers et al., 2017; Zhou et al., 2022; for a review, see Hackel & Amodio, 2018). Building on this work, we aimed to shift our focus even earlier in the process, asking how people come to actually form the categories that they assign positive or negative value to. Such a question requires an approach that can explicitly model the formation of categorical representations. We therefore draw on models of category formation, which have been widely used within cognitive psychology but have rarely been leveraged to answer questions within the social domain. In particular, we model social behavior using a network model of category learning called SUSTAIN (Love et al., 2004; see Figure 1 for an overview). SUSTAIN represents information as clusters of features along stimulus dimensions, which can be weighted with attention. Critically for our purposes, the model uses a flexible cluster creation algorithm that starts with simple few-cluster solutions and creates new clusters only if existing clusters cannot adequately explain the presented information. It can thus represent a range of category representations including those used by exemplar models (by recruiting one cluster per stimulus), those used by prototype models (by recruiting one cluster

per overall category), and those falling between these two extremes. We can therefore use SUSTAIN to identify whether each participant is building group-based representations, person-based representations, or some combination thereof. SUSTAIN's adaptive clustering mechanism builds on a long history of psychological theories of category learning (Love et al., 2004; Nosofsky et al., 1994; Yamauchi et al., 2002) and there is growing evidence that it predicts neural representations and processes across a variety of regions (Davis et al., 2012; Mack et al., 2016, 2018; Mack et al., 2020; Braunlich & Love, 2019). By providing a formal account of how learning processes lead to organized memory structures, the use of this model allows us to characterize the abstraction of participants' social representations.

In this research, we use SUSTAIN to examine the interaction of contextual and individual factors in producing group-based vs. person-based representations of others. Across two preliminary studies and one focal study, participants play a game adapted from Allidina and Cunningham (2021) in which they learn about alien people who belong to two visually distinctive groups, characterized by their skin color. Participants see one alien at a time and can choose whether or not to interact with each alien. If they say yes, the alien either helps them (by giving money) or hurts them (by taking money away). If they say no, they either do not get any information about them (in approach-contingent feedback conditions) or they find out what the alien would have done if they had approached (in full feedback conditions). Each group consists of 6 aliens who participants interact with multiple times, and cooperation rates vary both within and between groups. Thus, participants can complete the task by relying entirely on group-based representations, by forming representations of each person individually, or by using some combination of the two strategies (such as individuating the more positive group but relying on group-based representations for the more negative group). We determine participants' representations by modeling their task behavior using SUSTAIN. Specifically, we examine 1) whether approach-contingent feedback pushes people towards forming more group-based

rather than person-based clusters, and 2) what individual differences (if any) predict the formation of person- and group-based clusters in each condition.

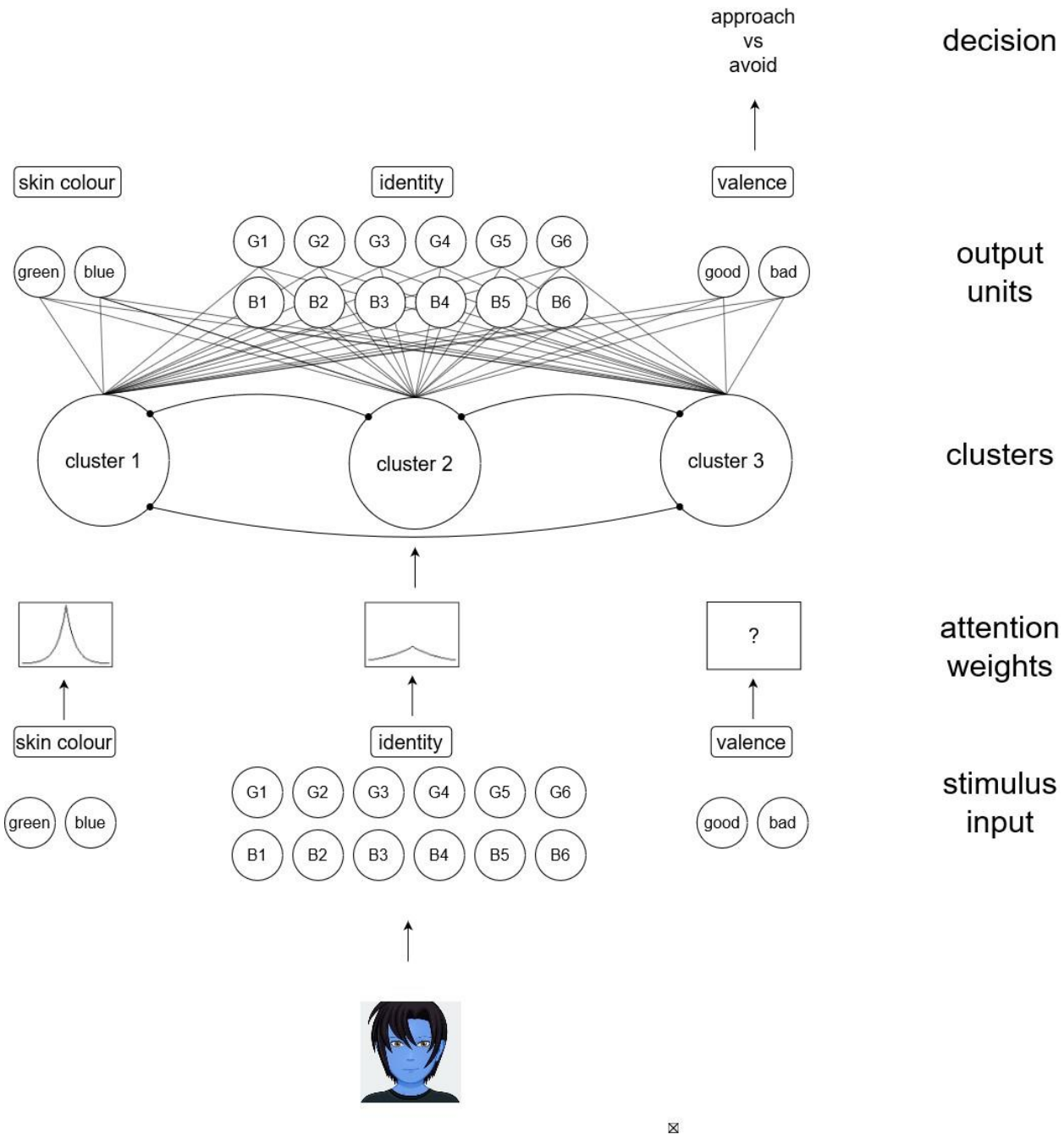


**Figure 1. Overview of the SUSTAIN model.** Stimulus inputs are encoded using the feature dimensions of skin color, identity, and valence (value unknown at stimulus encoding), which are weighted through an

attentional mechanism. The existing clusters compete to respond to the weighted stimulus input, with the cluster that is most active post-competition signaling the output units through a set of connection weights. The output units for valence are then used to drive the decision to either approach or avoid the stimulus.

## Preliminary Studies

We first conducted two preliminary studies to examine the ability of the model to capture the representations of interest in our task under approach-contingent feedback. The two studies were virtually identical, except that monetary outcomes were purely hypothetical in Study 1 but translated into real bonus money for participants in Study 2. Thus, we present results from both studies in the same section.

### Methods

#### *Participants*

*Study 1.* We aimed to recruit 100 participants from Amazon Mechanical Turk; 97 people completed the study and were paid for participating. One participant had participated in a pilot version of a similar study previously and was therefore excluded. To identify participants who were not paying adequate attention and should thus be excluded, we examined the variation in their responses and their response latencies. Specifically, we excluded those who had latencies below 150ms on more than 15% of trials (suggesting they were simply skipping through the task without engaging with it), those who had latencies greater than 5000ms on more than 15% of trials (suggesting that they were distracted by other activities during the task), and those who gave the same response on more than 85% of trials (suggesting they were not engaged with picking up on the variation in the stimuli presented). This left us with a final sample size of 77 participants whose data we analyzed.

*Study 2.* We aimed to collect 150 participants from Amazon Mechanical Turk; 146 people completed the study and were paid for participating. Seven participants had previously participated in study 1 and were thus excluded from analyses. As in study 1, we excluded

participants who had a latency below 150ms on more than 15% of trials, a latency above 5000ms on more than 15% of trials, or the same response given on more than 85% of trials, leaving a final sample size of 119 for analysis.

These studies were approved by the [university name masked for peer review] Ethics Board (protocol #33582)

### *Procedure*

Participants were told that they would play a game that involved making decisions about people from different alien groups. On each trial of this game, they would see a picture of an alien and have to decide whether or not to cooperate with this person. If they chose to cooperate, the person would either give them 1 point or take away 1 point. If they chose not to cooperate, they would not get any information about the other person's actions. Thus, feedback in the game was contingent upon approach. Participants were told that they would see the same people again and again, and that it would be useful to figure out which people were more likely to give or take money and to use that to guide their decisions.

In Study 2 only, participants were informed that they could gain up to an additional $10 depending on their performance in the game. Participants were told that they were starting the game with $2.00 in real money; any points they gained would be added to this total, while any points they lost would be subtracted from this total. At the end of the study, the amount of money they received would be proportional to the number of points they earned in the game.

The aliens that participants saw in the game belonged to two visually distinct groups, characterized by either green or blue skin. Each alien varied randomly in all other features, such as the shape of their face, eyes, mouth, and nose, such that every alien was visually distinctive. In phase 1 of the game, participants saw four different aliens from each group, for a total of eight different people. Each person was seen eight times, producing a total of 64 trials in phase 1. Each alien that participants encountered had a set probability of cooperating (i.e., giving 1

point). On average, one group was more cooperative than the other, but there was variation among the groups such that some aliens in each group cooperated at the same rate.

The four members of the more cooperative group (referred to here as the good group) cooperated with a probability of 0.9, 0.8, 0.6, and 0.5, respectively, producing an average cooperation rate of 0.7. The four members of the less cooperative group (referred to here as the bad group) cooperated with a probability of 0.6, 0.5, 0.3, and 0.2, respectively, producing an average cooperation rate of 0.4. The color of the good group was counterbalanced across participants, so that the green aliens were more cooperative for half of the participants and the blue aliens were more cooperative for the other half.

In phase 2, two of the aliens in each group were replaced with new aliens. The aliens who remained from the first phase were those who had equal probabilities of cooperating (0.6 and 0.5) across the two groups. The new aliens who appeared in phase 2 also had equal probabilities of cooperating (0.6 and 0.5). Thus, in phase 2, members of the two groups were on average equally likely to cooperate with the participant, with an average cooperation rate of 0.55. Each person was again encountered eight times, for a total of 64 trials in phase 2. The presence of initial extreme members who are then replaced by more neutral members was intended to reflect the idea that category-based information is often learned prior to individuating information, since stereotypes can be more easily transmitted indirectly before the perceiver has a chance to meet and learn about an individual person.

After completing both phases of the game, participants were presented with a series of faces and had to rate how likely each person was to cooperate with them, from 0 to 100. They rated each alien that they saw in the game, as well as four new people from each group, for a total of 20 ratings. Analysis of these ratings is presented in the supplementary materials. Participants then completed the Ten-Item Personality Inventory and answered a series of demographic questions.

### *Computational Model*

We modeled participants' responses in this task using SUSTAIN, a network model of category learning that compares incoming stimuli to existing category representations to make a classification decision. Here we present an overview of the model; full mathematical details are presented by Love and colleagues (2004).

Information in SUSTAIN is represented as clusters of stimulus features that are associated with category labels. Specifically, stimuli are encoded by their value on various feature dimensions - in this case, a skin color dimension representing group membership and an identity dimension representing the individual - which combine to form representational clusters. Each stimulus feature is represented by a number of input units equal to the possible values that the feature can take on. Thus, the group dimension is represented by two input units and the identity dimension by twelve input units. Units are set to 1 if the given stimulus holds that feature value and 0 otherwise; thus, a member of the "good" stimulus group might be represented on the group dimension by the units [1 0], and a member of the "bad" group by the units [0 1]. When the model encounters a stimulus, the features of the stimulus are differentially weighted by a series of attention weights (i.e., an attention weight for skin color and an attention weight for identity), with the weighted feature representation then compared to any existing representational clusters. If no clusters exist (as is the case on the first trial before any information has been learned), a new cluster is recruited, centered on the presented stimulus. If clusters do exist, they are activated according to how similar they are to the stimulus, with the weight given to each dimension in this similarity calculation governed by the attention weights. Clusters compete with one another, and the most activated cluster after competition passes its output to the output units of the unknown feature dimension (valence). The output of these units is then used to probabilistically drive a decision to approach or avoid the stimulus, based on the relative output of the positive and negative units.

SUSTAIN starts with zero clusters, and clusters can subsequently be recruited during learning according to a combination of factors. On a given trial, a new cluster will be recruited if the following criteria are met: the model predicts the wrong response and receives error feedback, the maximum cluster output post-competition is less than the threshold parameter, and a highly similar cluster does not already exist. Thus, the model starts with a simple category structure and flexibly recruits new clusters only as needed to represent new stimuli. In this way, it can capture a range of category representations including exemplar-based structures (by recruiting one cluster per stimulus), prototype-based structures (by recruiting one cluster per overall category), and representational structures that fall between these two extremes. Cluster positions and attention weights are updated with experience, with attention weights increasingly tuned towards the features that are most predictive for category membership and clusters updated to better represent the stimuli that activate them.

To illustrate the recruitment of clusters, imagine two participants who have each encountered one blue alien and received a negative outcome upon interacting with them. Each participant then encounters another blue alien, and their behavior towards this alien differs. The first participant decides, based on their prior experience with the initial alien, that this new alien will likely produce a negative outcome and decides to avoid them. For this participant, the existing cluster that represents the first blue alien will be highly activated for the new alien, producing the choice to avoid. The model would thus not recruit a new cluster to represent this new alien specifically, and the existing cluster would instead be updated to include this alien as well. If this happens for every new alien that is encountered, the model would estimate a total of 2 clusters for this participant (one for each group) by the end of the task. The second participant decides that they do not yet know enough about this alien to know what they are like, and decides to approach them. For this participant, their existing cluster is not activated strongly enough by this new alien, and the model would instead recruit a new cluster specifically for this

alien. If this happens for each new alien the participant encounters, the model would estimate a total of 12 clusters (one for each alien encountered) by the end of the task.

This implementation of SUSTAIN contains 6 free parameters that can be fit to the task data. These included the 5 parameters described by Love and colleagues (2004): a parameter governing the degree of attentional weighting ($r$), a parameter controlling the degree of cluster competition ($\beta$), a parameter controlling decisional consistency ($d$), a parameter that sets the threshold for creating a new cluster ($\tau$), and a parameter controlling the learning rate for cluster, attention weight, and connection weight updating ($\eta$). In addition, we included a free parameter representing participants' initial levels of attention to the group vs. identity dimensions, as used in the simulations by Love and colleagues (2004). As our goal was to use SUSTAIN to examine the category representations that each participant formed as they progressed through the task, we fit the model parameters separately to each participant's task responses. For a given participant, stimuli were presented to SUSTAIN in the same order as experienced by the participant, and the model's probability of making the same response as the participant on each trial was calculated and summarised with the log likelihood. A maximum likelihood differential evolution algorithm (Storn & Price, 1997) was used to find the model parameters that best predicted the trial-by-trial response for each participant.

### *Transparency and Openness*

Sample sizes for each study in this paper were determined before any data collection took place, and all exclusions, manipulations, and measures are reported. These studies were not preregistered. Data and analysis code for all studies is available at https://osf.io/uj9e7/?view_only=dad674b9e2734eec9758bd911c336585.

## Results

### *Task Behavior*

Before modeling participants' task behavior using SUSTAIN, we first wanted to ensure they were learning about the differences among the aliens. We therefore ran a multilevel logistic regression model predicting participants' trial-by-trial choices (approach or avoid) from the alien group and a variable indicating the type of alien present on that trial. This variable consisted of three categories: aliens who were present at the beginning of the task and cooperate at group-consistent extreme rates ("extreme aliens"), aliens who were present at the beginning of the task but cooperate at neutral rates ("early neutral aliens"), and aliens who were introduced in phase 2 and cooperate at neutral rates ("late neutral aliens"). Average rates of approach to aliens by group and type for all studies are presented in Supplementary Table 1.

In study 1, this analysis revealed main effects of both the alien group, $X^2 = 49.23$, $p < .001$, and the alien type, $X^2 = 43.59$, $p < .001$. Further, an interaction was found between the two variables, $X^2 = 80.75$, $p < .001$, such that participants always approached the good group more than the bad group, but this difference in approach was less pronounced for aliens who were actually neutral compared to those who were extreme.

Similarly, this analysis on the study 2 data again revealed a main effect of alien group, $X^2 = 79.91$, $p < .001$, a main effect of alien type, $X^2 = 74.78$, $p < .001$, and an interaction between the two variables, $X^2 = 214.84$, $p < .001$. Here, the difference in approach to the good and bad groups was significant for the extreme aliens and early neutral aliens, but not the late neutral aliens.

### *Computational Model*

Participants' behavior in this task was modeled using SUSTAIN, with the 6 free parameters (see Methods) fit individually to each person's trial-by-trial responses using a maximum likelihood differential evolution algorithm (Storn & Price, 1997). The parameters that

minimized the log likelihood for each participant were then used to extract the number of categories, or “clusters”, that each person formed in the task (see Supplementary Table 2 for parameter values). Each cluster represents one grouping that the participant used to represent the aliens, with clusters able to represent any number of individuals. For example, a given participant could allocate each individual alien their own cluster, so that the aliens are fully individuated, while another participant could place all aliens who share a skin color into the same cluster. As seen in Figure 2, there was considerable variation in the number of clusters that participants formed during the task. The histogram displays peaks at two-cluster solutions, in which participants are presumably forming one cluster for each of the two alien groups, and twelve-cluster solutions, in which they are fully individuating the aliens. The color of the bars displays the type of cluster formed. Each cluster was defined by a series of 12 individual dimensions (representing one possible value for each alien) and 2 group dimensions (representing one possible value for each group), as well as 2 outcome dimensions indicating if the cluster was associated with a positive or negative outcome. Clusters were classified as person clusters if the value of any of the 12 individual dimensions (which must sum to 1) was greater than 0.8[1]. Clusters that did not meet this threshold were classified as group clusters if either of the 2 group dimensions (which also must sum to 1) was greater than 0.8. Finally, clusters that did not meet the criteria for either person-based or group-based clusters were classified as “none”. As seen in Figure 2, the two-cluster solutions are almost entirely group-based, whereas solutions above about six clusters are almost entirely person-based. A notable number of participants landed on a one-cluster solution, which does not make sense given the task structure. As seen in the following section, these participants displayed a lack of learning of the basic task contingencies, and likely were not performing the task well.

[1] We also tested a variety of other values for this threshold and confirmed that the results hold under these values; thus, our choice of threshold does not impact the results.

To establish whether the model solutions were accurately mapping onto participants' behavior, we examined each participant's difference in approach behavior to members of the two groups. Specifically, we separated out trials by the alien type (extreme or neutral) and then computed a difference score between the average response to the good group and the bad group for each type. Thus, a value of 0 means that the participant approached the two groups at equal rates, and higher positive values mean that the participant approached the good group far more than the bad group. Critically, the extreme members of each group differ while the neutral members do not; thus, a participant who has fully learned the task should have a positive value for the extreme aliens and a value of 0 for the neutral aliens. As seen in Figure 2, participants with one-cluster solutions do not differentiate the groups much regardless of the alien type, suggesting that they are not even learning that the extreme members of the two groups differ. This supports the notion that these participants simply did not learn the task, either because they were not paying attention or because they could not figure out the task probabilities. In contrast, participants who formed two clusters in the task differentiate the groups both when they are actually different (i.e., the extreme aliens) and also when they are actually the same (i.e., the neutral aliens). Because these participants were relying solely on group-based information and failed to individuate the group members, they are unable to learn that there is variation in the groups and appropriately adjust their behavior to the different aliens. Those who formed more than 2 clusters, on the other hand, correctly differentiated the extreme members of the groups but not the neutral members. By learning about individuals rather than just groups, these participants were able to accurately adjust their behavior based on the aliens' actual rates of cooperation and not just their group membership.

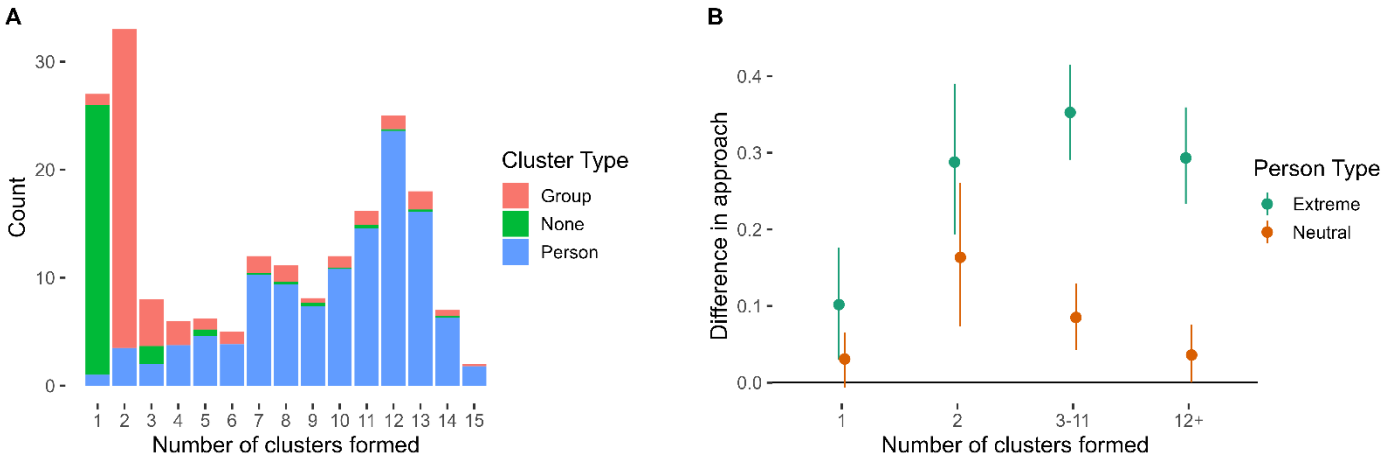


**Figure 2. Distribution of clusters formed by participants in preliminary studies.** Panel A) depicts a histogram of the number of clusters each participant formed during the task. The distribution shows peaks at 2 clusters, representing participants who were simply attending to the group information, and 12 clusters, representing participants who were fully individuating the groups. Bars are colored according to the proportion of clusters formed of each type. Those with only 2 clusters formed primarily group-based clusters, while those with more clusters formed primarily person-based clusters. Panel B) shows how the number of clusters formed relates to participants' task behavior. Error bars represent 95% confidence intervals. The average difference in each participant's behavior to the two groups was calculated; a difference of 0 indicates that the participant approached members of the two groups equally, and positive numbers indicate that the participant approached the good group more than the bad group. Those who formed only one cluster were unable to differentiate the groups even when they were really different, suggesting they were failing to adequately learn the task. Those who formed 2 clusters differentiated the groups both when they were different (extreme aliens) and when they were equal (neutral aliens), suggesting that they could learn the group difference but could not learn the variation within the groups. Finally, those who formed more than 2 clusters were most able to accurately differentiate the extreme members of the two groups while equating the neutral members of the two groups.

## Focal study

After establishing that SUSTAIN can capture meaningful variation in participants' task behavior, we turn to testing our two main questions of interest. Specifically, we examine how the environmental feedback structure and individual differences each shape the representations that participants form in this task.

### Methods

#### *Participants*

399 participants were recruited from Prolific and paid for completing the study. A sensitivity analysis conducted using the simr package in R (Green & MacLeod, 2016) indicated that a sample size of 400 would provide 80% power to detect unstandardized estimates

between 0.05 and 0.1 for the most complex hypothesis test in this study, the interaction of feedback condition and symbolic racism with the alien's group and type. As in the preliminary studies, we examined participants' response variation and response latencies to identify and exclude those who may not have been paying attention. Thus, we excluded participants who had latencies below 150ms on more than 15% of trials (5 participants), participants who had latencies greater than 5000ms on more than 15% of trials (9 participants), and participants who chose the same response on more than 85% of trials (3 participants). This left us with a sample of 382 participants for analysis.

### *Procedure*

Participants completed a task in which they learned to interact with two groups of alien people, characterized by green or blue skin, who could either help them by giving rewards or hurt them by taking rewards away. On each trial, participants were presented with a single alien and asked whether they wanted to interact with the alien or not. If the participant interacted, the alien would either give them one point or take one point away. If the participant chose not to interact, the alien could not give or take away any points. These points translated into real money that participants could earn at the end of the experiment.

Helping rates differed across the two alien groups, but variation also existed within groups. Specifically, each group consisted of 6 individual aliens. Half of these aliens in each group were neutral, cooperating at a rate of 0.5. The other half of the aliens were either very good (cooperating at a rate of 0.8) or very bad (cooperating at a rate of 0.2), depending on their group. Thus, one group on average was a "good" group and the other was a "bad" group, with the color of each group randomized across participants. Each individual alien was encountered a total of 8 times, allowing participants to learn not just about the groups but about specific individuals.

The more extreme aliens were encountered first in the task, with the more neutral aliens encountered later. Specifically, the task was divided into 13 “blocks” (although participants experienced the trials as one continuous task). In each block, a new alien from each group was introduced (first the extreme aliens and then the neutral aliens) until all 6 aliens in the group had been introduced. This was followed by 3 blocks in which participants repeatedly encountered all 6 aliens in each group. Finally, in the remaining blocks, one alien in each group was removed from the task in each block (first the extreme aliens and then the neutral ones). Thus, interactions with the extreme aliens were more likely in the first half of the task, and interactions with the neutral aliens were more likely in the second half. Participants completed a total of 96 trials overall.

Participants were randomly assigned to be in one of two conditions that determined the feedback structure of the experiment. For those in the Partial Feedback condition, information about the aliens was only revealed if participants chose to interact with them, at which time they would either give or take points. If participants chose not to interact, no information was provided about the alien’s actions. In contrast, those in the Full Feedback condition received information about the aliens regardless of whether or not they chose to interact with them. In cases where the participant avoided the alien, they learned what the alien would have done if they had chosen to interact.

Upon completing the task, participants were asked to rate the cooperativeness of various aliens, including all the aliens they saw in the task and 4 new aliens from each group. Each alien’s cooperativeness was rated on a scale of 0 to 100. Finally, participants responded to a series of questionnaires and provided demographic information, before being debriefed about the study.

### *Questionnaires*

Participants completed a series of questionnaires assessing constructs that may predict their behavior in this task, including both socially relevant factors like prejudice and more general cognitive styles like need for closure.

*Social questionnaires.* To assess levels of prejudice, participants completed the Symbolic Racism scale (Henry & Sears, 2002), an 8-item questionnaire designed to assess subtle forms of anti-Black racism such as racial resentment and denial of racial oppression. We chose to assess race-related attitudes as the type of prejudice that might most closely correspond to the tendency to generalize based on skin color. Participants' willingness to generalize within alien groups may also relate to their beliefs about the importance of controlling or managing the expression of prejudiced attitudes. Thus, we also had them complete the Internal and External Motivations to Respond without Prejudice scales (Plant & Devine, 1998), which assess both internally motivated reasons (e.g., fulfilling one's personal values) and externally motivated reasons (e.g., avoiding disapproval from others) for being non-prejudiced. Along the same lines, participants completed the 11-item attitudes subscale of the Social Justice Scale (Torres-Harding et al., 2012), which assessed their endorsement of social justice-related principles.

*Cognitive questionnaires.* Participants also completed the 42-item Need for Closure scale (Kruglanski et al., 1993), designed to assess individual differences in the desire for definite answers or certainty. The scale consists of 5 subscales of 7-10 questions each: Preference for Order, Preference for Predictability, Decisiveness, Discomfort with Ambiguity, and Closed-Mindedness.

### *Computational Model*

Participants' responses were modeled with SUSTAIN as in studies 1 and 2. To aid in parameter recovery and model simplicity for this study, values for the cluster competition

parameter ($\beta$), the threshold parameter ($\tau$), and the attentional parameter ($r$) were set to group values estimated from an initial fit.

## Results

### *Task Behavior*

Before modeling participants' task behavior using SUSTAIN, we first wanted to establish whether they were adequately learning about the two groups. We would expect that participants approach members of the good group more than members of the bad group, and that they potentially show some sensitivity to the variation within each group. Based on previous research showing that interactions with condition emerge only under extreme changes in the groups (Allidina & Cunningham, 2021), we do not expect an overall interaction with feedback condition in this study. However, we model the 4-way interaction of these variables with the trial number, as some difference between conditions may emerge over trials as participants learn even under this moderate change in the groups. Thus, we predicted choice (approach or avoid) from the alien group, the alien type (extreme or neutral), the participant's condition, and the trial number. Random slopes were modeled for each trial-level variable. Full results of this model are presented in Supplementary Table 3, and we report the key effects of interest here. This analysis indicated that participants learned to differentiate the two groups (main effect of group: $X^2 = 463.88$, $p < .001$) and were sensitive to the variation within the groups (interaction of alien type and group: $X^2 = 381.38$, $p < .001$). Further, a 4-way interaction was found between the group, alien type, trial number, and feedback condition, $X^2 = 5.13$, $p = .024$. Decomposing this interaction reveals that by the end of the task, participants in the full feedback condition were better able to learn about the variation within groups. For extreme aliens, by the end of the task people (correctly) differentiated the two groups slightly more in the full feedback condition (good-bad contrast: $b = 2.45$, $SE = 0.160$, $z = 15.33$, $p < .001$) than the partial feedback condition (good-bad contrast: $b = 2.21$, $SE = 0.167$, $z = 13.19$, $p < .001$). For the neutral aliens,

however, the opposite pattern emerged: by the end of the task, people in the full feedback condition (good-bad contrast: $b = 0.44$, $SE = 0.115$, $z = 3.84$, $p < .001$) seemed to differentiate neutral aliens by group slightly less than those in the partial feedback condition (good-bad contrast: $b = 0.67$, $SE = 0.122$, $z = 5.47$, $p < .001$). Thus, even under this more moderate change in the groups, there is some evidence that approach-contingent feedback prevents people from adequately learning about the variation within each group.

### *Computational Model*

**Overall.** Participants' behavior in study 3 was modeled using SUSTAIN, with parameters fit to each person's trial-by-trial choices using a maximum likelihood differential evolution algorithm. To improve parameter recoverability and model simplicity in this study, the $\beta$, $\tau$, and $r$ parameters were fixed to group values that were estimated from an initial fit. The other three parameters ($\eta$, $d$, and initial attention) were fit separately for each person. As in studies 1 and 2, the parameters that minimized the negative log likelihood for each person were then used to extract the number of clusters formed during the task (see Supplementary Table 2 for parameter values). Clusters were classified using the same criteria as in the first two studies, with a threshold of 0.8 used to label a cluster's contents. Overall, participants formed an average of 7.92 clusters, with 0.96 group-based clusters, 6.85 person-based clusters, and 0.12 clusters that fell into neither of these categories.

**Effect of Approach-Contingent vs. Full Feedback Conditions.** A key question in this study is how the feedback structure given to participants changes their category representations. In particular, we aimed to examine whether receiving approach-contingent feedback pushes participants towards more categorical representations rather than representations of individual targets. To answer this question, we compared the person-based

clusters formed by participants in each of the two conditions, asking whether those in the approach-contingent feedback condition are less likely to individuate some aliens[2].

Examining the number of person-based clusters each participant formed provides an indication of how much they were individuating the members of the two groups rather than simply treating them as indivisible wholes. We therefore examined whether people in the two conditions formed different numbers of person-based clusters for the good and bad groups. Specifically, we ran a multilevel model predicting the number of person-based clusters each participant formed from the group, the alien type (extreme or neutral), and the feedback condition, with random slopes for group and alien type. In addition to main effects and lower-order interactions, this analysis revealed a three-way interaction between group, person type, and feedback condition, $b = 0.028$, $SE = 0.012$, $t = 2.27$, $p = .0239$. As seen in Panel A of Figure 3, participants in the full feedback condition formed equal numbers of person-based clusters regardless of the group or type of alien, suggesting they were individuating all aliens approximately equally. In this condition, the number of person-based clusters formed for members of each group does not significantly differ for either the neutral people (good-bad contrast: $b = 0.046$, $SE = 0.086$, $t = 0.54$, $p = .592$) or the extreme people (good-bad contrast: $b = 0.066$, $SE = 0.086$, $t = 0.78$, $p = .439$). In contrast, those in the partial feedback condition formed fewer person clusters for the bad group than the good group, both for the neutral aliens (good-bad contrast: $b = 0.39$, $SE = 0.088$, $t = 4.47$, $p < .001$) and especially for the extreme aliens (good-bad contrast: $b = 0.63$, $SE = 0.088$, $t = 7.22$, $p < .001$). Thus, when feedback is contingent on approach, participants do not fully individuate particularly bad targets, instead treating them as simply representative of the group as a whole. Panel B of Figure 3 provides further insight into the person-based clusters participants are forming in the two conditions. In

[2] Note that when both person- and group-based clusters for the same stimulus are present in the model, the person-cluster will “win” and be activated more strongly than the group-based cluster (since it matches the stimulus on more dimensions). Thus, the number of person-based clusters provides a better measure of individuation than does the number of group-based clusters.

particular, the distribution of person-based clusters differs by group for those in the partial feedback condition (but not the full feedback condition). For participants who receive approach-contingent feedback, the good group is more likely to be fully individuated (i.e., be allocated 6 person-based clusters, one per alien) than the bad group. The bad group, in contrast, is more likely to be partially individuated, receiving 1-5 person-based clusters but not enough to fully individuate the group. Thus, this provides evidence that participants in this condition are treating the bad group as a "group plus exceptions" rather than as individuals.

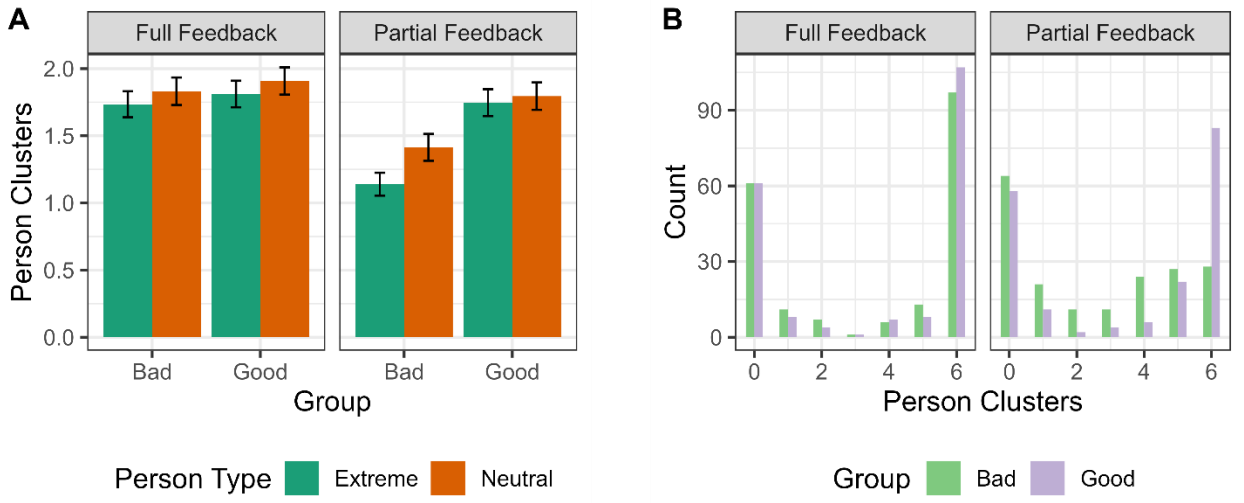


**Figure 3. Person-based clusters extracted from SUSTAIN fits in each condition.** Panel A) depicts the average number of person-based clusters formed by participants in each condition for each group and person type. Those in the full feedback condition formed similar numbers of clusters representing members of the bad and good alien groups who had extreme and neutral cooperation rates. However, those in the approach-contingent feedback condition formed fewer clusters representing the bad group, and especially few clusters for the extreme members of the bad group. Panel B) depicts the distribution of person-based clusters across the two conditions, showing how many participants form each number of clusters. Those in the full feedback condition have similar distributions for the two alien groups, with peaks at 0 and 6 suggesting that most people either fail to individuate targets at all or fully individuate them. A similar pattern is seen for the good group in the approach-contingent feedback condition. However, a different pattern emerges for the bad group in this condition: participants are much more likely to only partially individuate this group, forming 1-5 person-based clusters but not enough to fully individuate the group.

**Effect of Individual Differences**. After establishing that the feedback structure influences the category representations participants hold, we turn to exploring whether any individual differences predict the formation of person-based vs. group-based clusters. For each

of the four individual difference scales (symbolic racism, internal and external motivations to respond without prejudice, social justice orientation, and need for closure), we predict the number of person-based clusters from the group, the alien type, the feedback condition, the individual difference scale, and the interaction of all variables. Random slopes were modeled for the trial-level variables (group and alien type), with trials nested within subjects. We select only participants who created at least one person-based cluster for this analysis ($N$ = 283), as the inclusion of those with zero person-based clusters conflates those who were relying entirely on groups (e.g., who formed two groups) with those who were simply not learning (e.g., who formed only one cluster overall). Thus, these analyses examine which individual differences predict the *number* of person-based clusters a participant forms, given that they have formed at least one. In other words, they differentiate those who may be "subtyping" (Deutsch & Fazio, 2008; Johnston & Hewstone, 1992; Maurer et al., 1995; Richards & Hewstone, 2001), or selectively individuating those who do not fit the group's stereotype (for example, by forming only a single person-based cluster), from those who are fully individuating (by forming many person-based clusters). The main results of these analyses are displayed in Figure 4. When predicting the number of person-based clusters from symbolic racism, we find a significant interaction between racism and feedback condition, $b$ = 0.14, $SE$ = 0.064, $t$ = 2.15, $p$ = .032[3], such that people higher in racism form fewer person-based clusters only if they are in the partial feedback condition. Similarly, we find an interaction between internal motivation to control prejudice and condition, $b$ = -0.077, $SE$ = 0.035, $t$ = -2.20, $p$ = .029, such that those low in internal motivation form fewer person-based clusters if they are in the partial feedback condition. No parallel interaction was found between feedback condition and external motivation to control

[3] While we believe excluding those with no person-based clusters is the appropriate analysis with which to examine this question, results are generally consistent under various other exclusion criteria, though with some variation. For example, keeping all participants in without excluding based on the number of person clusters yields a p-value for this analysis of $p$ = .038. Using an alternate exclusion criterion of only removing those with a single cluster overall instead yields a p-value of .059. Thus, these results should be considered preliminary and replicated in future work.

prejudice, *b* = -0.007, *SE* = 0.033, *t* = -0.21, *p* = .831. When predicting person-based cluster counts from social justice orientation, we find a 4-way interaction between the group, the alien type, the feedback condition, and the participant's score on social justice orientation, *b* = -0.033, *SE* = 0.014, *t* = -2.27, *p* = .024. As seen in Figure 4, social justice orientation predicts the number of person-based clusters specifically for the extreme members of the bad group in the partial feedback condition. Thus, when given limited information about targets, those low in social justice orientation specifically fail to individuate those they deem to have harmed them.

In contrast to these social motivations, need for closure does not significantly interact with the feedback condition, *b* = 0.033, *SE* = 0.08, *t* = 0.40, *p* = .686. We instead find much weaker, suggestive evidence for a main effect of need for closure on the number of person-based clusters formed, *b* = -0.15, *SE* = 0.08, *t* = -1.89, *p* = .06. While this result does not appear to be particularly robust and should be replicated in future work, it may suggest that those high in need for closure form fewer person-based clusters overall, regardless of the feedback structure or target qualities.

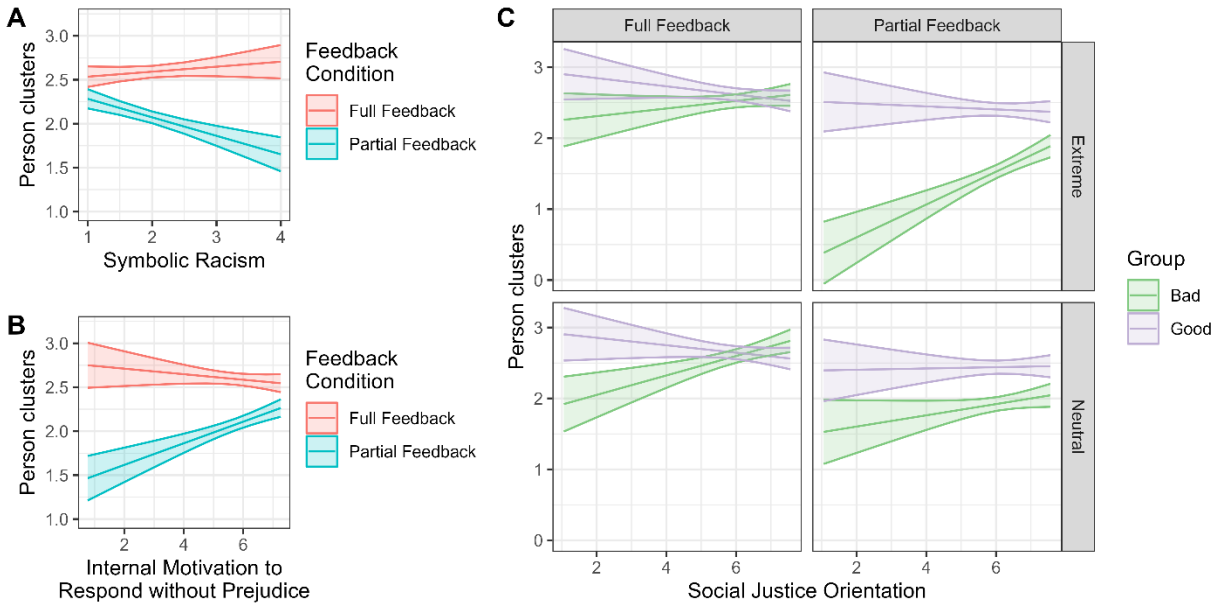


**Figure 4. Interaction of individual differences with feedback condition on the formation of person-based clusters.** Panel A) depicts the interaction of symbolic racism with feedback condition, indicating that those higher in racism were less likely to individuate the group only if they were in a context of limited

feedback. Panel B) shows the parallel interaction of internal motivation to respond without prejudice (IMP) with feedback condition. Those low in IMP who were in the partial feedback condition formed fewer person-specific clusters. Panel C) shows the interaction of social justice orientation, group, person type, and condition. Here, those low in social justice orientation specifically fail to individuate the extreme members of the bad group in the partial feedback condition.

**Predicting Ratings.** Thus far, we have established that the number and type of clusters formed by each participant is shaped by both their motivations and their environment. Finally, we aim to establish how these clusters relate to participants' overall beliefs about each alien. Upon completion of the task, participants were presented with a series of aliens (including aliens seen in the task and novel aliens) and asked to rate how likely each alien was to cooperate with them on a 0-100 scale. We used these ratings as our dependent variable, which we predicted from the alien's group, type (neutral, extreme, or new), and the number of person-based clusters the participant formed for that group. As in previous analyses, trials were nested within subjects and random slopes were modeled for the trial-level variables of group and alien type.

This analysis revealed main effects of group, $X^2 = 208.53$, $p < .001$, and alien type, $X^2 = 6.21$, $p = .045$, as well as two-way interactions of group with alien type, $X^2 = 43.07$, $p < .001$ and number of person-based clusters, $X^2 = 44.10$, $p < .001$. Most critically, a three-way interaction between group, alien type, and number of person-based clusters was also found, $X^2 = 10.08$, $p = .006$. As seen in Figure 5, the influence of the number of person-based clusters on group differentiation depends on the alien type. For extreme aliens, who really do differ by group, even those who fully individuated the aliens still (correctly) rate the groups differently. In contrast, for neutral aliens who did not differ by group, only those who failed to individuate the groups rate members of the two groups differently. Those who fully individuated the groups by forming 6 or more person-based clusters for the group instead show no difference in their estimates of how likely these members of the two groups are to cooperate with them. By treating each alien as an individual rather than simply a member of a group, they were able to accurately learn about the variation within groups and respond accordingly. Finally, new aliens who have never been seen

before fall somewhere in between: those who fully individuated the groups rate the group differences as much smaller than those who relied on categorical representations, but even those who formed individuated representations rate the groups somewhat differently (perhaps reflecting the average cooperation rate across all group members).

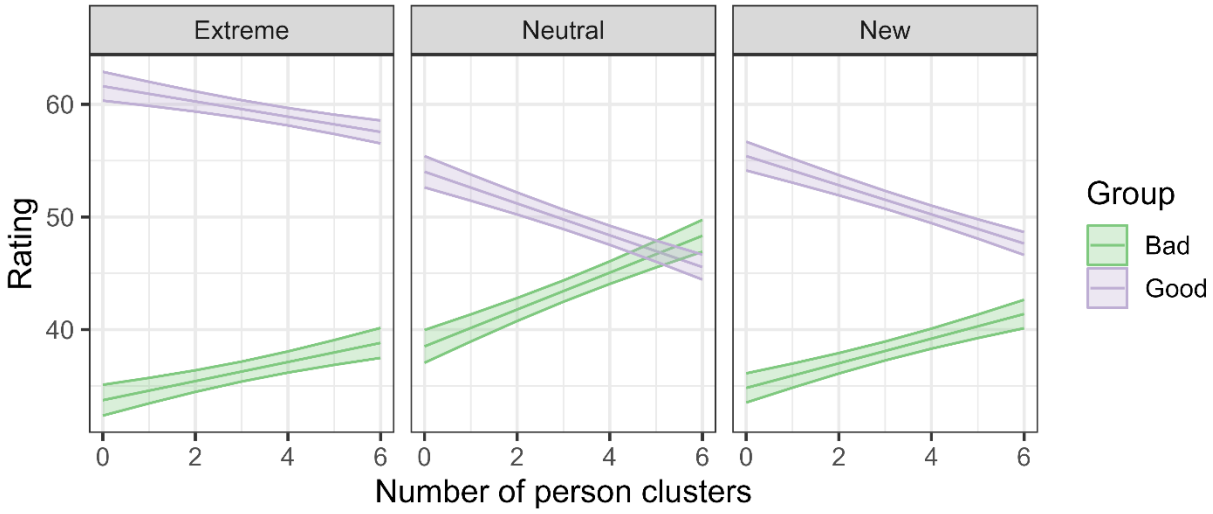


**Figure 5. Relationship of person-based clusters with explicit beliefs about the aliens' cooperation rates.** For all alien types (depicted in the three panels), those who form more person-based clusters report a smaller difference in the cooperation rates of the two groups. For the neutral aliens (whose cooperation rates do not differ by group) and new aliens (whose cooperation rates are unknown), those who fully individuated the targets during the task show no difference in their estimated cooperation rates of the two groups. Those who relied only on group-based representations instead report differences in the cooperation rates of the groups even when none exist.

## Discussion

Faced with the task of simplifying a complex social world, we often end up individuating members of some groups while treating members of other groups as interchangeable. The relative use of such individuating and category-based features to represent others has been of longstanding interest in psychology. Here, through the novel application of a computational model of category learning, we were able to gain insight into not just when these representations are drawn upon, but how they are initially formed. Applying this model to

participants' task behavior revealed a critical role for the structure of learning feedback: when feedback was contingent on approach, people formed more abstracted representations of those deemed to be bad. In doing so, they inadvertently lumped neutral targets in with bad targets who shared their skin color. Further, despite the wholly artificial nature of the task, prejudiced attitudes towards real-world social groups seemed to predict less individuated representations of alien creatures (although results were less robust than the more primary effects of condition). This was only the case under selective feedback, however, with the wider availability of information under full feedback acting as a protective mechanism that shielded people's representations from their negative attitudes.

The use of a formal category learning model was critical to our approach, allowing us to gain layers of insight out of a relatively simple task. Instead of being forced to infer participants' representations from their beliefs, the model allowed us to formalize the mechanisms and cognitive processes that might be creating these beliefs. This approach draws on the idea that individuation and categorization are not necessarily wholly distinct processes (Hamilton & Sherman, 1996) but two ends on a spectrum of cognitive investment (Fiske & Neuberg, 1990), utilizing the same learning mechanisms but with different experiences, information, and motivations as inputs. By formalizing the various processes that push people towards one end of this spectrum or the other, we could examine how attention, learning, and decision-making come together to produce different kinds of representations. Critically, there was intentionally no objectively "optimal" strategy to maximize one's bonus in our focal study: the neutral people cooperated at a rate of 0.5, such that the expected value of approaching such a person is 0. Thus, the choice to approach or avoid these people ultimately did not affect the participant's bonus money. This design choice reflects the fact that there are multiple strategies for person perception that can be equally effective in the moment, but result in drastically different downstream effects (in this case, the post-task beliefs about new aliens). This approach thus allowed to understand who was relying on more group-based or individuated strategies to

represent others, providing a precise way to examine stereotyping – operationalized here as more abstracted representations of targets.

In highlighting the utility of category models for studying social perception, we believe future work can expand these models even further to examine additional aspects of social representations. For example, a generative modelling approach that models not just the dependent variables but also the independent variables can allow for even more precise hypotheses about how motivations and environments may influence representations. We see the present work as an initial step towards modeling these processes, and believe additional modeling approaches will be of value in expanding this work. While we focused specifically on the initial formation of individuated or group-based social representations, a logical next step is to combine this investigation with the selective retrieval of such representations as a function of different motivational and environmental factors. For example, one could modify the cluster activation algorithm to reflect not only stimulus match along each attended dimension, but additional motivational and contextual factors including cognitive resources. This could further allow us to examine how such cluster-based representations may be changing and developing over time; some perceivers, for example, may begin with more group-based representations that they abandon for more individuated understandings as more information is gathered.

Social processing characterized by restricted individuation can have far-reaching implications, as the resolution at which we represent others shapes our subsequent beliefs, interpretations, and behaviors. Our approach focuses on the initial formation of such social representations as a critical process: if individuated representations are not formed during learning, they will not be available to draw upon during retrieval regardless of the perceiver's subsequent motivations. Thus, motivating perceivers to individuate others while making a social decision is not enough, as they may have no individuated representation to draw upon at that point. Instead, motivation early in the process of social impression formation is critical:

perceivers must choose to invest the cognitive resources to form detailed impressions during initial learning.

The number of prototypes a perceiver initially forms during learning is therefore key. While a single prototype leads to full reliance on categories and many prototypes can lead to full individuation, many participants fell between these extremes and instead formed “subtypes” of targets that incorporate schema-inconsistent information while maintaining their overall stereotypes (Deutsch & Fazio, 2008; Johnston & Hewstone, 1992; Maurer et al., 1995; Richards & Hewstone, 2001). These results suggest that such subtyping may be especially likely in contexts of limited information where gaining social feedback requires risk. In such contexts, perceivers will be likely to form shallower representations of others in which negative behaviors are more readily generalized to other group members. Thus, providing people with forms of information gain that reduce risk (e.g., virtual contact, contact in structured environments, or extended contact through members of one’s ingroup; Dovidio et al., 2017) could allow them to more fully individuate others, reducing reliance on categories and stereotype-maintaining subtypes.

The critical interaction of individual and environmental factors in creating prejudice (Akrami et al., 2009; Hodson & Dhont, 2015; Maddux et al., 2005; Meleady et al., 2021; Pettigrew, 1958) was echoed in these findings: we found preliminary evidence that negative attitudes and motivations predicted increased representational abstraction and subtyping only under selective feedback.. As learning about others under approach-contingent feedback involves risk to oneself, these results may speak to the increased salience of self-protective motives in highly prejudiced people. Perceptions of threat posed by an outgroup are closely linked with prejudice (Bahns, 2017; Hugenberg & Bodenhausen, 2003; Quillian, 1995; Riek et al., 2006; Stephan & Stephan, 2000; Velasco González et al., 2008), suggesting that prejudiced people may be more attuned towards the (often imagined) threat posed by others. Thus, more

prejudiced people may have been less willing to take on the risk of approaching potentially negative others under approach-contingent feedback, leading to reduced information and more shallow representations. When approach was not necessary for information gain, on the other hand, they were able to form more nuanced representations of others despite any reluctance to approach. Full feedback may thus have acted as a "protective factor" that shielded participants' representations and behaviors from their negative attitudes: even though some participants in the full feedback condition held negative attitudes, the increased availability of information in the environment nevertheless led them to individuated representations of others. When directly changing attitudes is difficult, these results may suggest that altering the context of information gain could provide an alternate route towards the same short-term results (with the hope that attitudes would eventually follow through increased information).

Future work should further explore the mechanisms through which these prejudice-related individual differences predicted behavior in this task. It is possible that these results reflect greater general category-based processing for those high in prejudice (which could contribute to and/or result from their prejudice). Alternatively, these results could be more specifically race-related, as the differentiating feature in our task was skin color (although the findings on social justice orientation, which does not directly assess race-related attitudes, make this less likely). An important avenue for future research is thus to examine whether this pattern holds in social categories unrelated to skin color, and potentially even in non-social categories. If similar patterns are found in the latter case, this would suggest greater category-based processing in those high in prejudice. If, on the other hand, these results are unique to social categories (and even skin color-related categories), this may suggest that the results instead reflect greater willingness to apply such abstracted processing to the social domain for those high in prejudice. Further, by exploring additional individual differences that more specifically reflect factors such as category-based processing or preference for social hierarchies, we can better understand the relative contribution of different factors that make up social prejudice.

Finally, a valuable avenue for future work will be to directly manipulate motivations that are hypothesized to be of relevance to better isolate their causal effects on representational abstraction.

These results were found in an entirely artificial social setting with none of the implications, histories, or stereotypes that accompany real-world social group interactions. Despite this seemingly arbitrary context, social motivations still shaped representations, suggesting that ideological factors related to racism and social justice orientation may shape very basic processes of person categorization even when there are no real-world implications. These effects may be even further exacerbated in real-world interactions through the confluence of stereotypes, resource implications, and prejudices that work to maintain group divisions and power structures. While we created real differences between the groups in this study, this variation could materialize in the real world through biased media portrayals, negative stereotypes, and gossip that creates perceived group differences even where none exist. When combined with negativity biases that tend to be especially prominent in attitudes towards outgroups (Ratliff & Nosek, 2011), the interactions of motivations and limited information gain could result not only in more shallow representations of outgroups, but in more negative ones.

**Constraints on Generality**

This study was conducted on participants located in the US who participated online through Prolific or Amazon Mechanical Turk. While the use of artificial groups as stimuli increases the likelihood that we are capturing general rather than context-specific processes, generalizability to non-WEIRD (Henrich et al., 2010) populations was not assessed. Future research should examine whether these effects hold across cultures. For example, cultures where skin color is a less relevant feature for differentiating groups across social hierarchies may have different predispositions towards alien groups defined by color.

Further, as discussed above, this work used a controlled paradigm with artificial social groups. This allowed us to assess our effects of interest while minimizing the influence of social desirability and the existing motivational “baggage” that accompanies real-world social groups. However, further work will be needed to ensure that these processes generalize as expected. It is possible that the increased complexity of motivations surrounding real social groups (such as moral considerations around the effects of social categorization) could result in altered category representations than those seen here. Future research should therefore explore whether and how these effects change when applied to social groups defined by dimensions like race and gender. While our specific design choices such as the relative salience of individuating and group-based features and the relative amounts of within- and between-group variation likely influenced mean amounts of individuation in our studies, there is less reason to believe that the effects of interest around individual- and environmental-level predictors of individuation should be specific to these choices. However, future work examining these effects under varying conditions will help to understand the robustness of the results to such design variations.

Finally, this study focused specifically on target behaviors that could help or hurt the perceiver. Given the primacy of morality in impression formation (Goodwin, 2015), we used helping and harming as our starting point to examine these questions. Future work could expand this investigation to examine other types of behaviors, such as those indicating competence or sociability. Critically, a key factor in our studies is the potential for harm: approaching the target must carry some level of risk in order for participants to choose not to approach and gain information. However, this potential for harm does not necessarily need to be intentional harm and could be caused by a well-intentioned but incompetent other, for example. Future work could examine whether these results generalize to situations where other types of potentially harmful behaviors are relevant.

## Conclusion

By using a class of computational models underutilized in social psychology thus far, the current research was able to examine the category representations that participants formed for novel others. We found evidence for the interplay of context and motivations in the very formation of social representations, suggesting that categorization itself is intertwined with the perceiver's worldviews. As a necessary precursor to stereotypes, prejudice, and discrimination, the selective application of social categories provides an important research target in understanding the effects of ideologies and environments on social behavior. This research presents a step in that direction, furthering our understanding of how and when people categorize others.

References

Akrami, N., Ekehammar, B., Bergh, R., Dahlstrand, E., & Malmsten, S. (2009). Prejudice: The Person in the situation. *Journal of Research in Personality*, *43*(5), 890–897. https://doi.org/10.1016/j.jrp.2009.04.007

Allidina, S., & Cunningham, W. A. (2021). Avoidance begets avoidance: A computational account of negative stereotype persistence. *Journal of Experimental Psychology: General*, *150*(10), 2078–2099.

Bahns, A. J. (2017). Threat as justification of prejudice. *Group Processes & Intergroup Relations*, *20*(1), 52–74. https://doi.org/10.1177/1368430215591042

Bai, X., Fiske, S. T., & Griffiths, T. L. (2022). Globally Inaccurate Stereotypes Can Result from Locally Adaptive Exploration. *Psychological Science*.

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

Deutsch, R., & Fazio, R. H. (2008). How subtyping shapes perception: Predictable exceptions to the rule reduce attention to stereotype-associated dimensions. *Journal of Experimental Social Psychology*, *44*(4), 1020–1034. https://doi.org/10.1016/j.jesp.2008.03.001

Ditonto, T. (2019). Direct and indirect effects of prejudice: Sexism, information, and voting behavior in political campaigns. *Politics, Groups, and Identities*, *7*(3), 590–609. https://doi.org/10.1080/21565503.2019.1632065

Ditonto, T. (2020). The Mediating Role of Information Search in the Relationship Between Prejudice and Voting Behavior. *Political Psychology*, *41*(1), 71–88. https://doi.org/10.1111/pops.12599

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

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

Fazio, R. H. (1990). Multiple Processes by which Attitudes Guide Behavior: The Mode Model as an Integrative Framework. In *Advances in Experimental Social Psychology* (Vol. 23, pp. 75–109). Elsevier. https://doi.org/10.1016/S0065-2601(08)60318-4

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

Fiske, S. T., & Neuberg, S. L. (1990). A Continuum of Impression Formation, from Category-Based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation. In *Advances in Experimental Social Psychology* (Vol. 23, pp. 1–74). Elsevier. https://doi.org/10.1016/S0065-2601(08)60317-2

Goodwin, G. P. (2015). Moral Character in Person Perception. *Current Directions in Psychological Science*, *24*(1), 38–44. https://doi.org/10.1177/0963721414550709

Hackel, L. M., & Amodio, D. M. (2018). Computational neuroscience approaches to social cognition. *Current Opinion in Psychology*, *24*, 92–97. https://doi.org/10.1016/j.copsyc.2018.09.001

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

Hackel, L. M., Kogon, D., Amodio, D. M., & Wood, W. (2022). Group value learned through interactions with members: A reinforcement learning account. *Journal of Experimental Social Psychology*, *99*(December 2021), 104267. https://doi.org/10.1016/j.jesp.2021.104267

Hamilton, D., L., & Sherman, S. J. (1996). Perceiving persons and groups. *Psychological Review*, *103*(2), 336–355.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Henry, P. J., & Sears, D. O. (2002). The symbolic racism 2000 scale. *Political Psychology*, *23*(2), 253–283. https://doi.org/10.1111/0162-895X.00281

Hodson, G., & Dhont, K. (2015). The person-based nature of prejudice: Individual difference predictors of intergroup negativity. *European Review of Social Psychology*, *26*(1), 1–42. https://doi.org/10.1080/10463283.2015.1070018

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing Prejudice: Implicit Prejudice and the Perception of Facial Threat. *Psychological Science*, *14*(6), 640–643. https://doi.org/10.1046/j.0956-7976.2003.psci_1478.x

Johnston, L., & Hewstone, M. (1992). Cognitive models of stereotype change. 3. Subtyping and the perceived typicality of disconfirming group members. *Journal of Experimental Social Psychology*, *28*(4), 360–386. https://doi.org/10.1016/0022-1031(92)90051-K

Konovalova, E., & Le Mens, G. (2017). Selective information sampling and the in-group heterogeneity effect. *Proceedings of the Thirty-Ninth Annual Conference of the Cognitive Science Society*, 688–693.

Kruglanski, A. W., Webster, D. M., & Klem, A. (1993). Motivated Resistance and Openness to Persuasion in the Presence or Absence of Prior Information. *Journal of Personality and Social Psychology*, *65*(5), 861–876. https://doi.org/10.1037/0022-3514.65.5.861

Lindström, B., Selbing, I., Molapour, T., & Olsson, A. (2014). Racial Bias Shapes Social Reinforcement Learning. *Psychological Science*, *25*(3), 711–719. https://doi.org/10.1177/0956797613514093

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309

Maddux, W. W., Barden, J., Brewer, M. B., & Petty, R. E. (2005). Saying no to negativity: The effects of context and motivation to control prejudice on automatic evaluative responses. *Journal of Experimental Social Psychology*, *41*(1), 19–35. https://doi.org/10.1016/j.jesp.2004.05.002

Maurer, K. L., Park, B., & Rothbart, M. (1995). Subtyping Versus Subgrouping Processes in Stereotype Representation. *Journal of Personality and Social Psychology*, *69*(5), 812–824. https://doi.org/10.1037/0022-3514.69.5.812

Meleady, R., Hodson, G., & Earle, M. (2021). Person and situation effects in predicting outgroup prejudice and avoidance during the COVID-19 pandemic. *Personality and Individual Differences*, *172*, 110593. https://doi.org/10.1016/j.paid.2020.110593

Neuberg, S. L., & Fiske, S. T. (1987). Motivational influences on impression formation: Outcome dependency, accuracy-driven attention, and individuating processes. *Journal of Personality and Social Psychology*, *53*(3), 431–444. https://doi.org/10.1037/0022-3514.53.3.431

Pettigrew, T. F. (1958). Personality and sociocultural factors in intergroup attitudes: A cross-national comparison. *Journal of Conflict Resolution*, *2*(1), 29–42. https://doi.org/10.1177/002200275800200104

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811–832. https://doi.org/10.1037//0022-3514.75.3.811

Quillian, L. (1995). Prejudice as a Response to Perceived Group Threat: Population Composition and Anti-Immigrant and Racial Prejudice in Europe. *American Sociological Review*, *60*(4), 586. https://doi.org/10.2307/2096296

Ratliff, K. A., & Nosek, B. A. (2011). Negativity and Outgroup Biases in Attitude Formation and Transfer. *Personality and Social Psychology Bulletin*, *37*(12), 1692–1703. https://doi.org/10.1177/0146167211420168

Rich, A. S., & Gureckis, T. M. (2018). The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of Experimental Psychology: General*, *147*(11), 1553–1570. https://doi.org/10.1037/xge0000466

Richards, Z., & Hewstone, M. (2001). Subtyping and subgrouping: Processes for the prevention and promotion of stereotype change. *Personality and Social Psychology Review*, *5*(1), 52–73. https://doi.org/10.1207/S15327957PSPR0501_4

Riek, B. M., Mania, E. W., & Gaertner, S. L. (2006). Intergroup Threat and Outgroup Attitudes: A Meta-Analytic Review. *Personality and Social Psychology Review*, *10*(4), 336–353. https://doi.org/10.1207/s15327957pspr1004_4

Schultner, D. T., Stillerman, B. S., Lindström, B. R., Hackel, L. M., Hagen, D. R., Jostmann, N. B., & Amodio, D. M. (2024). Transmission of societal stereotypes to individual-level prejudice through instrumental learning. *Proceedings of the National Academy of Sciences*, *121*(45), e2414518121. https://doi.org/10.1073/pnas.2414518121

Spiers, H. J., Love, B. C., Le Pelley, M. E., Gibb, C. E., & Murphy, R. A. (2017). Anterior Temporal Lobe Tracks the Formation of Prejudice. *Journal of Cognitive Neuroscience*, *29*(3), 530–544. https://doi.org/10.1162/jocn_a_01056

Stangor, C., & Ford, T. E. (1992). Accuracy and Expectancy-confirming Processing Orientations and the Development of Stereotypes and Prejudice. *European Review of Social Psychology*, *3*(1), 57–89. https://doi.org/10.1080/14792779243000023

Stephan, W. G., & Stephan, C. W. (2000). An integrated threat theory of prejudice. In *Reducing Prejudice and Discrimination*. Psychology Press.

Storn, R., & Price, K. (1997). Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *Journal of Global Optimization*, *11*, 341–359.

Torres-Harding, S. R., Siers, B., & Olson, B. D. (2012). Development and Psychometric Evaluation of the Social Justice Scale (SJS). *American Journal of Community Psychology*, *50*(1), 77–88. https://doi.org/10.1007/s10464-011-9478-2

Velasco González, K., Verkuyten, M., Weesie, J., & Poppe, E. (2008). Prejudice towards Muslims in The Netherlands: Testing integrated threat theory. *British Journal of Social Psychology*, *47*(4), 667–685. https://doi.org/10.1348/014466608X284443

Zhou, Y., Lindström, B., Soutschek, A., Kang, P., Tobler, P. N., & Hein, G. (2022). Learning from Ingroup Experiences Changes Intergroup Impressions. *The Journal of Neuroscience*, *42*(36), 6931–6945. https://doi.org/10.1523/JNEUROSCI.0027-22.2022