Stereotypes as Byproducts of Hierarchical Bayesian Inference

Junsong Lu[1*], Chujun Lin[2]

[1] Department of Psychology, University of California San Diego, La Jolla, U.S.

[2] Department of Psychology, Columbia University, New York, U.S.

Author Note

Junsong Lu https://orcid.org/0000-0001-6987-6228

Chujun Lin https://orcid.org/0000-0002-7605-6508

Data and code are available at https://osf.io/suycv/overview?view_only=62132642749542879a52c347f3960ba9.

The authors have no conflict of interest to disclose.

*Corresponding author *E-mail address*: jul140@ucsd.edu

Abstract

Stereotypes are ubiquitous despite being inaccurate. Existing theories primarily offer psychological explanations, suggesting that stereotypes arise as mental shortcuts (e.g., categorization). However, understanding why stereotypes emerge requires identifying the computational problem people solve when forming impressions. We propose a novel computational account: stereotypes arise as a byproduct of maximizing perceptual accuracy in judgments of individual targets. We formalize this process using hierarchical Bayesian inference (HBI), whereby perceivers adaptively learn group-level stereotypes as priors that prevent inferences from becoming overly sensitive to noisy social cues. Although computationally adaptive, this process can facilitate biased stereotypes. Across five simulations and seven empirical samples ($n = 1{,}273$), we show that people form and maintain group stereotypes consistent with the computational principles of HBI. Stereotypes emerged when perceivers were asymmetrically exposed to different groups (Study 1) and, even under sufficient exposure, because HBI facilitates faster acquisition of group-level impressions (Study 2). Established stereotypes persisted because counterevidence was attributed either to individual idiosyncrasies (Study 3) or to a target's secondary group membership (Study 4). Together, these findings identify HBI as a unifying account of stereotyping, advancing a generalizable computational framework for social inference at the computational level.

Keywords: stereotype, Bayesian cognition, person perception, social cognition, computational modeling

Introduction

To navigate a complex social world, humans display a striking capacity for rapid inference of others' intentions, traits, and behaviors from sparse or noisy information[1–4]. For instance, people can infer traits such as competence, or even guess someone's occupation, from a face alone [5,6]. A key mechanism underlying these efficient social inferences is the use of social categories[5,7–12]. Perceivers quickly identify a target individual's group membership (e.g., gender, race, or occupation) and draw on stored group-level knowledge to form judgments about the individual. However, these group knowledge, commonly referred to as stereotypes[13,14], are often inaccurate and biased[3,8,15–17]. This raises a fundamental question: if stereotypes are systematically biased, why do they persist across cultures and societies[18]?

A long-standing account, adaptive categorization, explains stereotypes as a byproduct of minimizing mental cost. For example, when a restaurant employer hires a chef from an unfamiliar group, knowledge of that group can be useful: it reduces the need to track detailed information about each individual and helps guide future hiring decisions. From this perspective, stereotypes function as cognitive shortcuts in social inferences[13,15]. However, it remains unclear in what sense stereotyping is adaptive—since whether a cognitive process is adaptive or maladaptive necessarily depends on the problem the perceiver is trying to solve.

A recent account, biased-sampling, instead explains stereotypes as a byproduct of maximizing rewards[16,17,19,20]. As in the example above, a restaurant employer who forms a positive early impression of a chef from an unfamiliar group may be more likely to hire others from that group, while avoiding groups that initially seemed less promising. Over time, these choices produce a biased sample of experiences, reinforcing the original stereotypes [16,19–22]. Despite this theoretical progress, people in real life cannot always control whom they interact with, making exposure to counterevidence inevitable. Yet stereotypes still persist, a pattern that remains puzzling.

Despite a long history of diverse theories, most have focused on explaining specific mechanisms (e.g., inadequate exposure)[16] or particular phenomena of stereotyping (e.g., illusory correlations)[23]. In terms of level of analysis[24], current accounts largely operate at the algorithmic level, emphasizing the psychological processes underlying stereotype formation. However, different psychological processes may in fact reflect different solutions to a shared computational problem—namely, the goal people are trying to achieve when forming impressions. Thus, identifying this computational problem and how people solve it provides a route to addressing the fundamental "why" question of stereotyping, with the potential to yield a unifying account.

Moving beyond the adaptive-categorization and biased-sampling accounts, we propose a new account of stereotypes, hierarchical-inference account, explaining how stereotypes are biased outcomes from adaptive functions. Although some stereotypes may carry a kernel of truth[25], many do not [15,16,20,26–29]; the current research focused on the latter. Specifically, we proposed that stereotypes are byproducts of maximizing perceptual accuracy of individual targets. Under this argument, human minds aim to produce accurate individual-level inferences rather than forming biased inferences of groups, as accurate individual-level inferences are important for guiding daily social activities (e.g., hiring the best chef).

Computationally, maximizing individual-level perceptual accuracy can be formalized as performing Bayesian inference [30,31], in which beliefs about an individual's latent traits are computed by combining prior expectations with observed evidence. This procedure can facilitate estimates that minimize expected prediction error [31,32]. In social contexts, for example, one might judge a chef's cooking ability based on both the quality of the current dishes they produced and prior beliefs about their culinary competence. Crucially, integrating prior beliefs improves accuracy because the prior acts as a form of statistical regularization: the prior constraints the impact of current observed evidence on the final judgment [31,33,34]. Although this process may appear biased because judgments are not based solely on the observed evidence, it prevents inferences from being overly driven by limited or noisy observations, thereby supporting more reliable social judgments and resulting in lower prediction error.

However, simple Bayesian inference poses two challenges for perceivers. First, it is unclear where the prior for a specific individual comes from. In the absence of regularization from appropriate priors, Bayesian inference collapses into an observation-only model, losing its optimality [35]. Second, Bayesian inference can be computationally demanding [36], especially in real life where people must track multiple individuals and specify different priors. We propose that a more efficient, cost-effective solution is hierarchical Bayesian inference[33,37], in which individual-level and group-level inferences are computed simultaneously and constrain one another, and the latter serves as the prior that can be applied to all individuals of the group and that gets updated upon observations of any individuals in the group.

This hierarchical-inference account explains why biased group-stereotypes are produced and maintained when the mind aims to maximize individual-level perceptual accuracy. Hierarchical Bayesian inference optimizes accuracy at the level of individuals rather than the level of groups. Because stereotypes arise as a byproduct rather than a direct optimization target, they need not contain a kernel of truth. More importantly, hierarchical Bayesian inferences yield optimal individual-level inferences only when they accurately capture the real-world data-generating process; otherwise, *model mismatch* happens[31]. For example, if a chef's performance is not actually influenced by group membership, an employer who bases their judgments on group-level knowledge will form inaccurate judgments about individual chefs sharing that group membership. This mismatch is likely in social contexts because groups are often culturally constructed without necessary differences in individuals' characteristics (e.g., the competence of individuals in different racial groups)[38].

We argue that stereotypes are byproducts of hierarchical Bayesian inference in social context: while trying to optimize accurate representations of individuals, biased inferences of the individuals' group are produced. To test whether people spontaneously perform hierarchical Bayesian inferences when observing unfamiliar individuals, we tested whether four specific predictions of the hierarchical Bayesian inferences were supported. First, hierarchical Bayesian inferences predict that asymmetric exposure to different groups induces stereotypes, a phenomenon connecting to existing literature on contact theory[15,39]. Second, hierarchical Bayesian inferences predict that stereotypes emerge even without such asymmetries, because hierarchical inference enables faster impression formation at the group level. Third, hierarchical Bayesian inferences predict that stereotypes are maintained even with counterevidence because it could be attributed to individual-level differences, linking to existing literature on individuation[40,41]. Finally, hierarchical Bayesian inferences predict that stereotypes are maintained even with counterevidence because it could also be attributed to individuals' alternative group memberships—a process that complements subtyping accounts[42].

Here, across four studies, we provided evidence supporting all four predictions of hierarchical Bayesian inferences. In each study, we first conducted simulations to illustrate our prediction and established the normative benchmark for hierarchical Bayesian inference. We then present data from human experiments to provide empirical support of these predictions from hierarchical Bayesian inference. We operationalized inferences of individuals and groups as probabilistic representations (e.g., posterior distributions) over traits. We defined biased group-stereotypes as differences in the mean (or representational certainty) of group-level inferences in the absence of true differences between groups (i.e., when no hierarchical structure exists in the data-generating process).

Results

Support for hierarchical inference 1: asymmetric exposure drove stereotype formation

The first prediction of hierarchical inference suggests that stereotypes emerge in contexts characterized by *asymmetric exposure* to members of different groups. This prediction builds on the Bayesian account of illusory correlation proposed by Gershman and Cikara[43]. Illusory correlation refers to the tendency to perceive a spurious association between groups and traits under imbalanced base rates of contact[26,27]. For example, a perceiver may interact with members of Group A twelve times and Group B four times, observing 9 and 3 positive interactions, respectively. Although the positivity rates are identical (9/12 = 3/4 = 75%), people tend to form more favorable impressions of the more frequently encountered Group A. Gershman and Cikara explain this pattern within a simple Bayesian framework: starting from neutral priors, greater exposure to positive outcomes shifts posterior beliefs further above neutrality, producing a bias toward the more frequently sampled group.

We extended this account within the hierarchical Bayesian framework: while group-level impressions function as prior for inferences about individuals, they are themselves learned from observed individuals and begin from a neutral prior (e.g., a hyperprior). As a result, shifting a group impression in a more positive direction requires a larger number of positive social interactions to move it away from neutrality. Building on Gershman and Cikara[43], we further link this mechanism to environmental structure to explain stereotype formation. In particular, any environmental factor that creates asymmetric exposure across groups can facilitate stereotype formation when group impressions are learned in a hierarchical Bayesian manner. Also, hierarchical inference predicts similar stereotype strengths across groups under similar environments. While not exhaustive, we focused on two common sources of such asymmetry: social segregation and group size imbalance, which both result in a perceiver having more contacts with individuals from one group than another group.

We first used agent-based simulations in Studies 1a–1b to show what group-level inferences an agent performing hierarchical Bayesian inference would produce when faced with asymmetric exposure to different groups. We then used empirical data from human participants in Studies 1c–1d to show that human perceptions were consistent with those from hierarchical-Bayesian agents.

Simulation study 1a-1b. In Study 1a, we examined the first type of asymmetric exposure—social segregation—examining its impacts on stereotype formation in hierarchical Bayesian agents. Social segregation suggests that individuals from the same group are spatially close to each other. We simulated a segregated society of three groups where agents preferentially assort with ingroup members, leading to increasing spatial segregation over time (Fig. 1). Social segregation was calculated as the average proportion of outgroup members within each group member's immediate surroundings.

A perceiver performing hierarchical Bayesian inference entered the society after segregation stabilized and explored the environment by interacting with members of all three groups. Based on binary social outcomes (positive vs. negative), the perceiver inferred a latent trait (kindness) of both the individuals and the group. Social outcomes were sampled for all individuals without real underlying group differences (model-mismatch condition): this is consistent with the notion that social groups are often culturally constructed rather than reflecting the group-to-individual structure assumed by hierarchical inference, hence a model-mismatch. Therefore, we sampled outcomes from a single distribution across individuals and groups (e.g., everyone was equally nice).

We varied the segregation level across 900 simulated societies. Stereotypes were measured as the standard deviation of the posterior means of group estimation (μ) across the three groups, capturing the extent to which the perceiver differentiated among groups. A linear mixed-effects model showed that, across 900 societies, social segregation significantly increased stereotypes ($\beta = 0.32, p < .001$). Greater social contact (e.g., earlier entry into society, 200 minus entry time) decreased stereotypes ($\beta = -0.22, p < .001$). Shifting to a model-match assumption where social outcomes were generated from a hierarchical process that matched the hierarchical mental model did not change the main conclusion (social segregation: $\beta = 0.25, p < .001$; social contact: $\beta = -0.22, p < .001$). These findings together show that when perceivers perform hierarchical Bayesian inference, they form group stereotypes when they are exposed to members from different groups asymmetrically due to social segregation, regardless of whether the true structure of individuals' traits matches the perceiver's hierarchical-inference assumption or not.

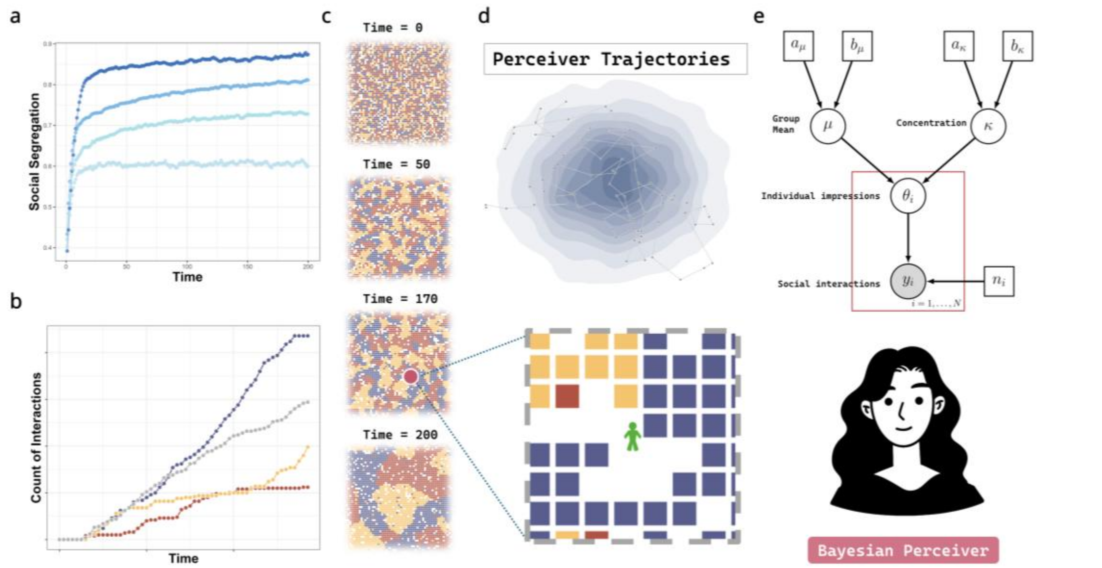

Figure 1. Simulating stereotype formation in segregated societies using hierarchical Bayesian agents. a, Four sample societies with varying levels of social segregation, indicated with different colors. **b,** Segregation induces asymmetric exposure: a new perceiver entering the society interacts with the three different groups (indicated with red, yellow, blue) with increasingly different accumulated counts over time. The grey line indicates no interaction with any group member at a given iteration. **c,** Snapshots of the simulated society at different time points, showing increasing segregation over time. The three colors (red, yellow, and blue) denote individuals from the three groups. **d**, Density contour of the perceiver's movement trajectory, approximating a bivariate normal distribution with a randomly generated mean. **e,** The Bayesian perceiver's mental model, formalized as hierarchical inference equivalent to a hierarchical beta–binomial model.

In Study 1b, we examined how a second type of asymmetric exposure, group size imbalance, affects stereotype formation in hierarchical Bayesian agents. We again simulated a three-group society (Fig. 2) and defined imbalanced group size as the sum of squared group shares in the population: $\sum_{k=1}^{K} S_k^2$, where S denotes the group share in the population. We generated 30 sets of imbalanced group sizes and crossed them with additional factors including the number of targets encountered and the number of interactions per target, yielding different conditions.

In each condition, the hierarchical Bayesian perceiver encountered individuals from each group with probabilities proportional to their population sizes and updated beliefs based on observed social outcomes. Stereotypes were measured as the standard deviation of the posterior means of group estimation (μ) across the three groups, as in Study 1a.

Linear mixed-effects analyses across all conditions showed that greater group size imbalance significantly increased stereotypes (model-mismatch condition: $\beta = 0.007, p < .001$; model-match condition: $\beta = 0.006, p < .001$;). These results showed that when perceivers perform hierarchical Bayesian inference, they form group stereotypes when they are exposed to members from different groups asymmetrically due to imbalanced group sizes, regardless of whether the true structure of individuals' traits matches the perceiver's hierarchical-inference assumption or not.

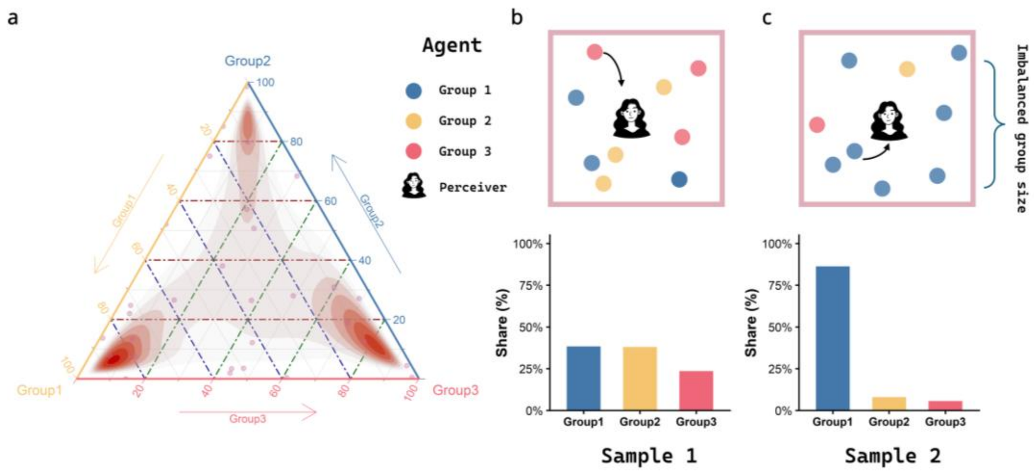

Figure 2. Simulating stereotype formation under imbalanced group sizes using hierarchical Bayesian agents. a, Thirty samples of imbalanced group sizes drawn from a sparse Dirichlet distribution; each sample consists of shares of three groups that sum to one. Points closer to the vertices indicate greater group size imbalance. Shares of the three groups summing to one. **b,** Illustration (upper panel) and corresponding bar graph (lower panel) of a relatively balanced composition of group sizes, in which the perceiver has approximately equal probabilities of interacting with members from each group. **c,** Illustration (upper panel) and corresponding bar graph (lower panel) of a relatively imbalanced group size composition, in which the perceiver predominantly interacts with members of the blue group.

Empirical study 1c-1d. In Studies 1c–1d, we tested whether humans form group-level inferences similar to hierarchical Bayesian agents under asymmetric exposure to different groups.

In Study 1c, we used previously collected human subject data[44]. Specifically, we quantified stereotypes based on perceived trait differences among ethnic groups reported by participants across the 50 U.S. states, with higher values indicating stronger stereotyping. We then quantified social segregation and group size imbalance using data from the U.S. Census Bureau, which provided detailed demographic information for all 50 states. Social segregation was derived separately for each ethnicity. Correlational analyses showed that social segregation was positively associated with stereotypes (Fig. 3a), with correlations ranging from 0.28 to 0.53 across different ethnicity-specific social segregation (all ps < .049). For example, segregation index based on the White population was significantly correlated with stereotypes toward ethnic groups (Spearman's $\rho = .41$, $p = .004$). The average segregation index across ethnic groups was also associated with stereotypes (Spearman's $\rho = .51$, $p < .001$). For group size imbalance, we replicated prior work's findings that more imbalanced group sizes were associated with stronger stereotypes[44] (Spearman's $\rho = .40$, $p = .004$; Fig. 3b). These results remained robust when we conducted a geospatial analysis accounting for spatial autocorrelation between states (see Supplementary Material S1a). Together, these findings suggest that human perceptions are consistent with those of hierarchical-Bayesian agents, with stronger group stereotypes emerging as exposure becomes more asymmetric due to social segregation and group size imbalance.

While Study 1c relied on cross-sectional data, Study 1d provided a complementary test using panel data (Fig. 3d-f). We analyzed data from the U.S. General Social Survey (GSS) [45]. We extracted human participants' ratings of White and Black Americans on three traits (intelligence, work ethic, and wealth) across 15 to 17 years from 1990 to 2024. We also used the GSS item, "liveblks", which assessed the racial composition of respondents' neighborhoods, to quantify social segregation and group size imbalance. The data were aggregated by year and the nine U.S. Census divisions (e.g., Pacific).

Using a fixed effects regression, we found that both social segregation and group size imbalance significantly predicted greater gaps in perceived traits (stereotypes) for Black versus White Americans. Specifically, higher social segregation ($b = 0.81$, $t = 3.83$, $p < .001$) and greater group size imbalance ($b = 0.87$, $t = 5.73$, $p < .001$) were each associated with larger Euclidean distances between trait ratings of Black and White Americans across Census divisions. The linear time trend was not significant ($b = -0.001$, $t = -0.17$, $p = .866$), suggesting that the association between asymmetric exposure and group stereotypes was stable over time after accounting for Census division fixed effects. Consistent with the cross-sectional results, the panel data similarly suggested that human perceptions aligned with those of hierarchical-Bayesian agents, with stronger group stereotypes emerging under greater asymmetric exposure.

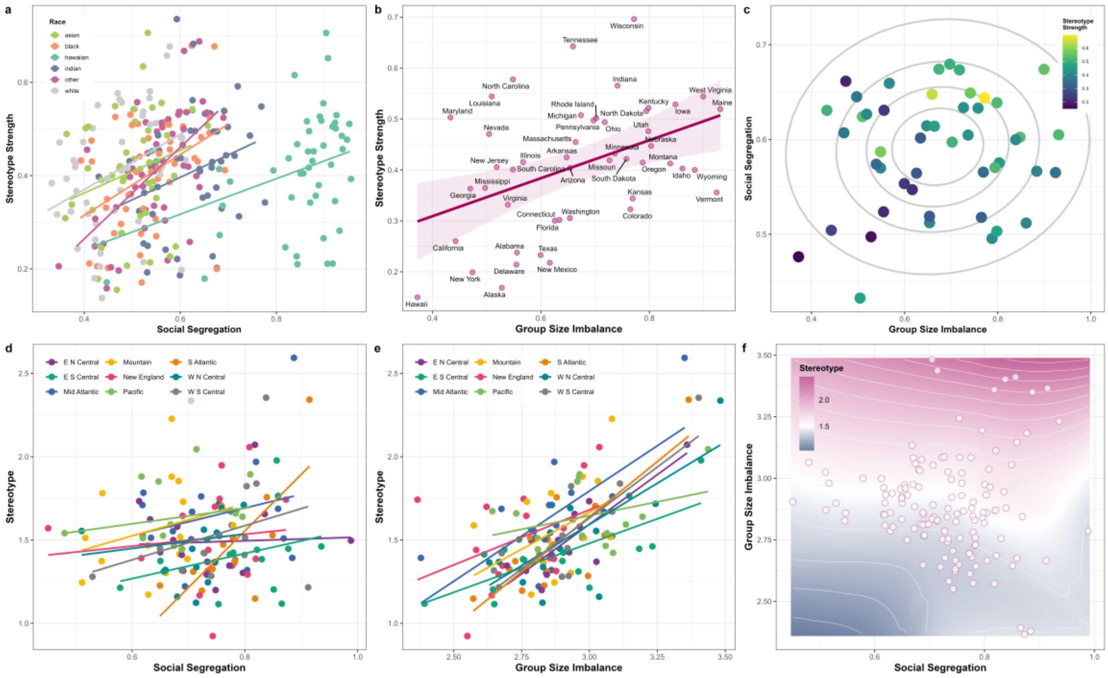

Figure 3. Social segregation and group size imbalance associated with group stereotypes in cross-sectional and panel data (Study 1c-1d). a, social segregation positively correlated with stereotype across the 50 U.S. states, shown separately for each ethnicity-based segregation measure (i.e., segregation based on Hawaiian was correlated with the overall ethnic stereotype). **b,** group size imbalance positively correlated with stereotype across the 50 U.S. states. **c,** Joint distribution (density contour) of social segregation and group size imbalance across states, indicating a low correlation between the two predictors, suggesting that they capture distinct types of asymmetric exposure. **d,** Social segregation positively correlated with stereotype across Census divisions, with each point representing one census division in a given year, spanning 1990 to 2024. **e**, group size imbalance positively correlated with stereotype across Census divisions, with each point representing one census division in a given year, for years ranging from 1990 to 2024. **f**. The contour plot indicating a LOESS-smoothed surface of Black–White stereotype dispersion (color gradient, blue = low, red = high) as a joint function of social segregation (x-axis) and group size imbalance (y-axis). Stereotypes were strongest when both factors were high (top-right corner, dark red).

Support for hierarchical inference 2: Shallow interactions drove stereotype formation

While asymmetric exposure is common in daily life, with increasing diversity in societies, individuals are likely to interact with different groups relatively equally (e.g., male versus female colleagues). Next, we examined how hierarchical Bayesian inference facilitated stereotype formation without asymmetric exposure, and how human participants' data supported this prediction.

A key feature of hierarchical Bayesian inference is that learning occurs simultaneously at multiple levels of abstraction, enabling faster learning of group-level knowledge than individual-level knowledge in certain contexts [33,34,46]. Specifically, when perceivers have an equal total number of social interactions across different groups, interacting with more individuals within a group and thus observing each individual less often (shallow interactions) leads hierarchical inferences to form more certain group-level inferences over individual-level inferences (Fig. 4). Such shallow interactions thus facilitate stereotype formation and can occur when total social time is limited but many individuals are available for interaction. For instance, at a networking party, the duration is typically fixed and short while people try to meet as many new individuals as possible. Consequently, perceivers can learn more quickly about the group than the different individuals within the group.

Computationally, differences in learning speed across levels of abstraction (e.g., group versus individual) arise from partial pooling in hierarchical Bayesian models [33,46]. Partial pooling allows information learned from one cluster (e.g., an individual group member) to generalize to others. In the context of stereotype formation, this occurs by aggregating information across individuals to estimate a group-level parameter, which then serves as a prior for subsequent optimal individual-level learning. As a result, the group-level estimate typically accumulates more evidence than any single individual and becomes more certain. When only limited information is available for each individual target and that information varies across individuals, the perceiver lacks sufficient evidence that targets are different. Consequently, partial pooling would induce certain group-level impressions.

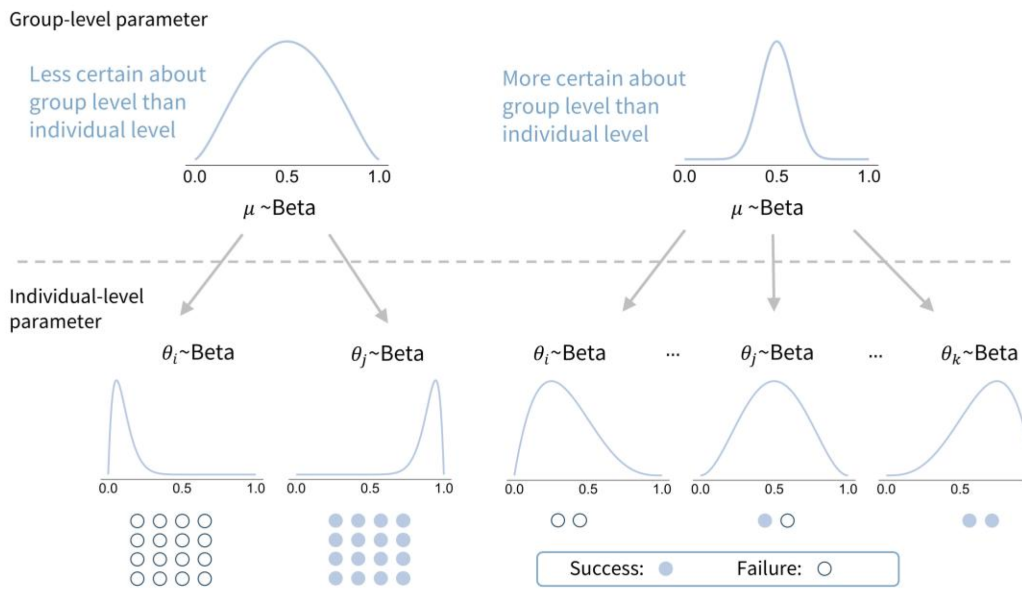

Figure 4. Shallow interactions facilitate group stereotype formation in hierarchical Bayesian Inference. Left: Deep interactions with highly variable individuals (i.e., high dispersion, e.g., one individual succeeds 0/16 times while another succeeds 16/16) yields a weak group-level impression μ (low group-level certainty). Right: Shallow interactions with many individuals (i.e., obtaining just a few observations per individual) leads to uncertainty about individual-level impressions; however, because there is no strong evidence that individuals are highly distinctive**,** pooling across individuals produces a certain group-level impression.

Simulation study 2a. We first used agent-based simulations to demonstrate the behavior of perceivers who performed hierarchical Bayesian inferences in situations where interactions with group members were shallow.

In our simulation, the hierarchical Bayesian perceiver estimated the competence of an unfamiliar alien group at performing a job. Aliens attempted the job multiple times, each resulting in either success or failure. From these observations, the perceiver drew inferences about each alien's competence and the group's competence. We modeled the perceiver's inferences of individual aliens and the group using a specific model of hierarchical Bayesian inferences, hierarchical beta-binomial model, and manipulated the distribution of the observations across 177 simulation conditions that described various situations of shallow interactions. Specifically, we varied the number of encountered group members and the number of observations per individual, while holding constant the total number of observations and overall positivity (i.e., the number of successes). For each condition, stereotype strength was measured as the standard deviation of the posterior distribution of perceived group-level competence (μ), where smaller values indicated greater certainty and stronger stereotypes. Because individual observations constitute the primary input for group-level inference, we also computed a dispersion score, defined as the standard deviation of success counts across individual aliens, to examine how variability in individual interactions (i.e., dispersion) influences group-level inferences.

Simulations showed that shallow interactions promoted stereotype formation in hierarchical Bayesian perceivers. Specifically, increasing the number of observed targets (thus, reducing the number of observations per target) significantly reduced posterior uncertainty in group impressions while controlling for total observations and overall number of successes, $\beta = -5.19 \times 10^{-4}, p < .001$. There was also a significant interaction effect between the number of targets and the dispersion across targets, $\beta = 2.69 \times 10^{-4}, p = .01$, such that encountering more targets further weakened stereotype formation when those targets were heterogeneous. These findings showed that when a hierarchical Bayesian perceiver had shallow interactions with more group members, they held more certain group-level impressions, even when information about each individual was limited and insufficient for accurate individual-level inference. Furthermore, the more heterogeneous the observed group members were, the stronger the group stereotypes.

Experiment 2b. We next recruited human participants (n = 365) to test whether human perceptions were consistent with those generated through hierarchical Bayesian inference. We induced shallow interactions by varying the number of targets (2 vs. 16) and dispersion across the targets (low vs. high) while holding constant the total number of observations (32) and the number of successes (16) across targets. Each participant was randomly assigned to only one of the conditions. For example, in the high-dispersion, two-target condition, participants observed two aliens, each with 16 attempts of mining (32 ÷ 2), with one alien succeeding in just a few attempts (4) and the other in most (12). Participants read a cover story in which they investigated aliens on a distant planet (Fig. 5). Aliens on the planet extracted blue minerals. Participants observed outcomes of mining attempts by individual aliens from one group defined by eye color (e.g., red). Each trial showed outcomes from multiple attempts (Fig. 5b, lower panel), with each attempt yielding either one mineral (success) or none (failure). After viewing outcomes from all mining attempts, participants rated both the competence and their confidence of such inference, for the last alien they observed and for the alien group as a whole using 7-point Likert scales.

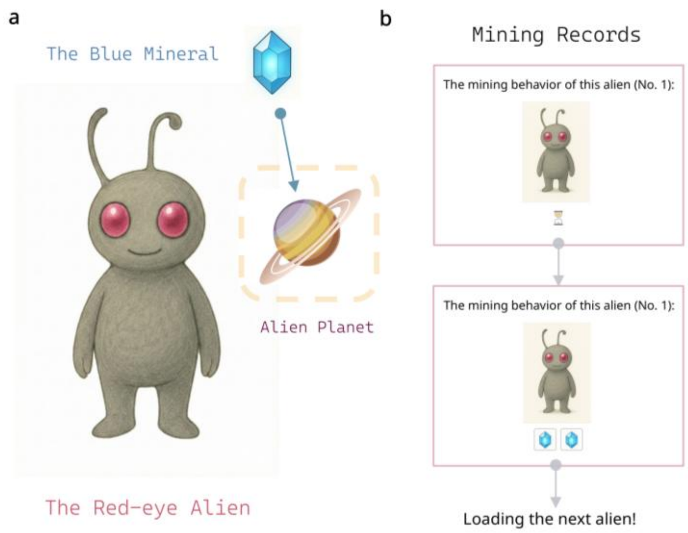

Figure 5. Experiment design for learning mining capability of aliens in Exp 2b. a, Participants acted as explorers observing aliens mining blue minerals. Participants were told that the aliens' eye color indicated their group membership, and each participant only observed aliens from one group (e.g., red eye aliens). **b,** Procedures: first, participants viewed an alien; then the outcomes of multiple mining attempts of that alien were shown. In the two-alien condition (16 trials per alien), outcomes from two attempts were shown per screen, with eight screens in total per alien sequentially revealing outcomes from its all attempts. This design was to match the sixteen-alien condition (2 trials per alien), where each screen also showed outcomes from two attempts of the alien, with one screen per alien followed by the next alien.

We first analyzed participants' confidence ratings of their inferences of the alien group's competence using a robust two-way between-subjects ANOVA. As preregistered, we accounted for the potential differences in how different participants used the Likert scales by computing the confidence difference score for each participant (group-level minus individual-level confidence), where positive values indicated more certain group-level impressions (i.e., stronger stereotypes). We found that, as expected, observing more targets increased the gap between group- and individual-level certainty ($Q = 32.87$, $p = .001$); there was no significant relation between observing greater dispersion across targets and confidence difference for group- versus individual-level impressions ($Q = 3.62$, $p = .06$). These effects were buried when individual differences in scale use was not accounted for (all *ps* > .24).

These main effects were qualified by a significant interaction effect between the number of targets and the dispersion across targets on the confidence difference between group- and individual-level impressions ($Q = 6.59$, $p = .012$). As expected, when participants observed only two heterogeneous targets (i.e., deep interactions), they expressed greater confidence in individual- than group-level impressions ($W = 251$, $p < .001$); when participants observed many targets (i.e., shallow interactions), regardless of how different those targets were (i.e., high or low dispersion), group-level impressions were more certain than individual-level impressions (Fig. 6; 16-low: $W = 876$, $p = .043$; 16-high: $W = 1158$, $p = .034$). Together, these results suggest that shallow interactions with many group members facilitate group stereotype formation in human participants, consistent with those generated by hierarchical-Bayesian agents.

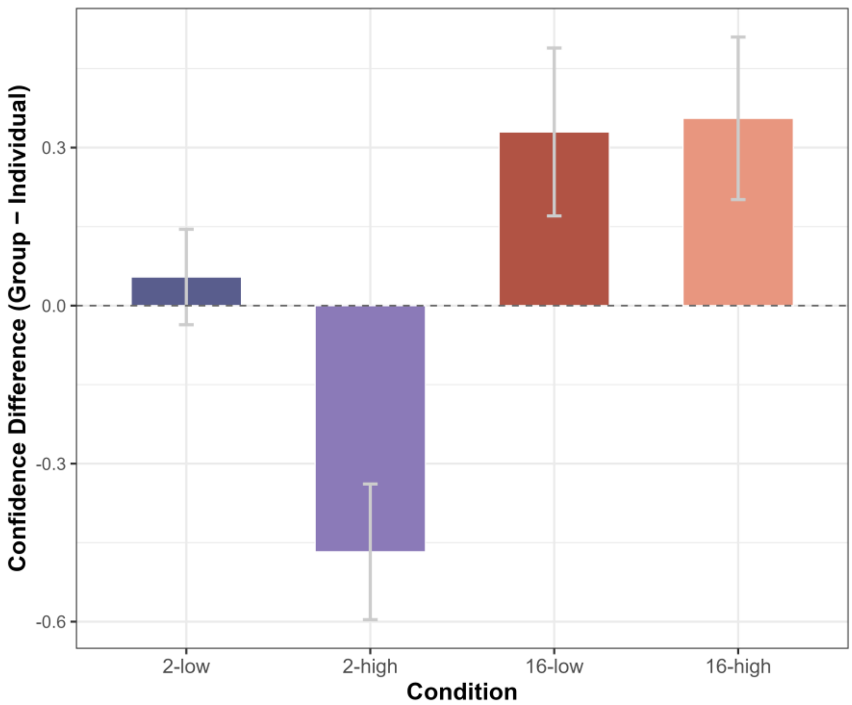

Figure 6. Shallow interactions elicited stronger group stereotypes in human participants. Y-axis indicates the confidence rating difference between group-level impressions and individual-level impressions. Each bar indicates the finding in each condition, where the number of targets the participants observed (2 vs. 16) and the variability across targets (low vs. high) were manipulated. Confidence in group-level impressions was significantly greater than that for individual-level impressions when interactions were shallow.

Support for hierarchical inference 3: Stereotypes persisted when stereotype-inconsistent group members behaved heterogeneously

So far, we have shown that stereotypes are readily formed in a hierarchical Bayesian manner in human participants when exposures to different groups are asymmetrical or to different group members are shallow. Next, we show how hierarchical Bayesian inferences also explain why learned stereotypes (e.g., via social transmission such as media exposure) are hard to change.

Hierarchical Bayesian inference predicts that existing stereotypes are preserved when the perceived contingency between the group membership and individual members is attenuated, characterized by the extent to which individual members' traits are perceived to depend on the group average. When perceivers observe multiple stereotype-inconsistent group members whose behaviors are heterogeneous, the perceived link between individuals and the group is less salient, leading stereotype-inconsistent evidence to be attributed to the individuals rather than used to update beliefs about the group. For example, if aliens on Mars have a reputation of being emotionally neutral, but one encounters two aliens there that are both very different from this stereotype, one being overjoy while the other depressed. This may lead the perceiver to attribute their individual deviations from the group stereotype to the two individuals (e.g., exception to the group) rather than using the individual deviations to update beliefs about aliens on Mars.

In the hierarchical Bayesian framework, the process of attenuated contingency is governed by the concentration parameter κ in the hierarchical beta-binomial model. A greater dispersion (e.g., variability) among individual impressions θ_i around the group mean μ yields lower κ and weaker perceived group-individual contingency (e.g., how strong individuals' traits θ_i are dependent on the group average μ). When κ is low and both individual-level impressions θ_i and the group mean μ are uncertain, new observations update both levels. However, when the group-level stereotype is strong (e.g., a tight distribution of μ), stereotype-inconsistent evidence preferentially updates the more uncertain individual-level θ_i, leaving the group mean μ largely unchanged (Fig. 7). Computationally, high heterogeneity among individuals induces a low κ, creating a paradox: given a strong learned group stereotype, the more heterogeneous the observed group members are, the harder for the group stereotype to be updated.

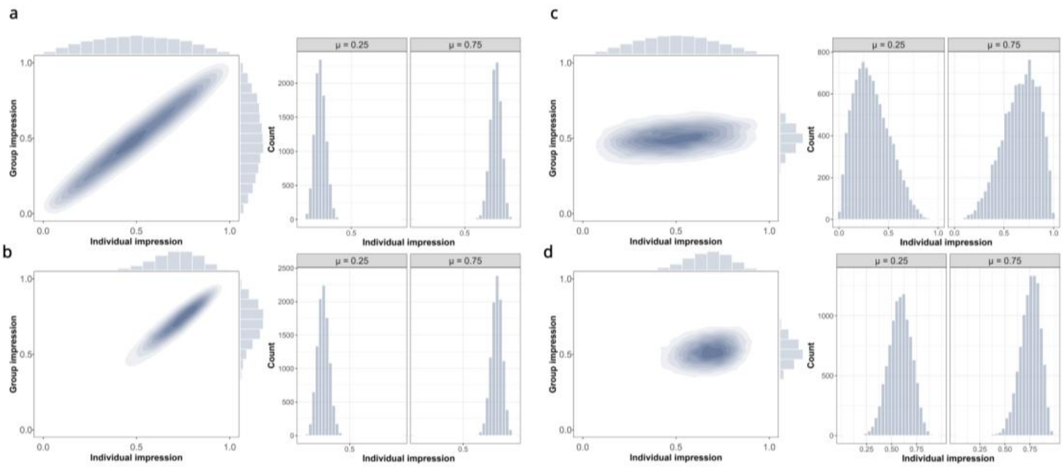

Figure 7. Attenuated group-individual contingency impedes stereotypes updating in hierarchical Bayesian inference. a-b, Under a high κ (strong perceived group–individual contingency), initially uncertain group-level and individual-level impressions (a, left) are jointly updated when a group member shows strong competence (b, left), as reflected by tighter marginal distributions concentrated at the high end. Paired histograms on the right ($\mu = 0.25$ and $\mu = 0.75$) show the conditional distributions of individual-level impressions given the group average (i.e., the perceived group-individual-trait contingency). **c-d,** Under a low κ and a certain group-level impression (c, left; as indicated by the tight group-level marginal distributions), the same evidence (i.e., when observing a group member with strong competence) primarily updates individual-level impressions while the group-level impression remains stable (d, left). Paired histograms on the right show the conditional distributions of individual-level impressions given the group average, which are wider than those in a and b, reflecting a weaker perceived group–individual-trait contingency.

Simulation study 3a. Using the same hierarchical beta-binomial model as in Study 2, we simulated the belief updating process of hierarchical Bayesian perceivers in response to stereotype-inconsistent evidence. The stimulation can again be understood as estimating the mining capability of the alien group. To induce a prior stereotype toward the group mimicking a learned group stereotype via social transmission, we assigned the group average μ a relatively strong prior distribution, reflecting a confident belief of low competence (mean ≈ 0.33). We also placed a weak prior on κ so that the perceived contingency can be influenced by social information. The hierarchical-Bayesian perceiver then observed new mining records from multiple individual aliens and updated both individual- and group-level impressions. To test generalizability, we manipulated the number of observed group members, observations per target, and the distribution of ground-truth competence, resulting in 11,539 conditions. We quantified stereotype persistence using an update score, defined as the difference between the posterior and prior mean of μ, with smaller values indicating stronger stereotype persistence.

As predicted, simulated results showed that dispersion of observations consistently discouraged stereotype updating ($\beta = -0.020, p < .001$) after controlling for the number of targets encountered ($\beta = 0.008, p = .003$) and observations per target ($\beta = 0.025, p < .001$). There was also a significant interaction between the number of group members observed and dispersion ($\beta = -0.002, p = .01$), suggesting that when more heterogeneous targets were observed, stereotype updating was further attenuated. This slower updating was associated with a lower concentration parameter κ under high dispersion ($\beta = -13.02, p < .001$), reflecting the computational process of perceived low group-individual-trait contingency that made stereotypes less responsive to stereotype-inconsistent evidence. Together, these results show that when a hierarchical-Bayesian perceiver holds a relatively strong stereotype about a group that they have learned from social transmission, observing more stereotype-inconsistent group members whose behaviors are heterogeneous induces stronger barriers to updating beliefs about the group.

Experiment 3b-3c. We next recruited human participants (n = 333 in Experiment 3b, n = 160 in Experiment 3c) to test whether humans, like hierarchical Bayesian perceivers, are reluctant to update stereotypes under attenuated group-individual contingency.

In Experiment 3b, we induced attenuated perceived group-individual contingency by manipulating the distribution of stereotype-inconsistent evidence across targets (concentrated vs. dispersed) and the number of targets observed (5 vs. 10), resulting in four conditions. Participants were randomly assigned to one of the four conditions. They read a cover story similar to Experiment 2b and estimated the mining capability of red-eyed aliens, defined as the average number of minerals collected in 10 attempts. To establish an initial group stereotype paralleling the simulation, participants first completed a social learning task: they observed and predicted the responses of four prior "participants" who had viewed multiple aliens' mining records. Participants were told that prior "participants" made decisions about which of two given mining records better fitted a red-eyed alien's mining outcome. Unbeknownst to participants, these decisions were generated by an algorithm. For each prior "participant", participants completed 10 trials, where they indicated which of the two given mining records that the prior "participant" selected as better fitted the red-eyed alien's mining outcome, and then they received feedback on the prior "participant's" actual selection.

After all of the 40 social learning trials, participants observed the aliens themselves. They viewed mining records of 10 attempts (e.g., extracted 5 minerals out of 10 attempts) by multiple aliens from the group and then provided their perceived group-level competence by completing a Markov Chain Monte Carlo with People (MCMCP) task[47]. The task consisted of 20 binary choices (e.g., 3 vs. 5 minerals out of

10 attempts) in which participants selected the option they believed better described the red-eyed alien groups' competence. To compare stereotype updating across conditions, we used the mean of participants' 20 choices as the dependent variable, with smaller values indicating slower stereotype updating.

Results showed no significant effects of target numbers, $Q = 1.32$, $p = .254$, distribution of stereotype-inconsistent evidence, $Q = 1.25$, $p = .265$, or their interaction, on posterior stereotypes, $Q = 0.01$, $p = .934$. Post-experiment interviews with eight participants suggested that the social learning task was confusing: the choices of 'prior participants' were generated probabilistically and produced inconsistent patterns across trials (e.g., preferring 2.5 minerals out of 10 attempts over 2.2 in one trial but 2.0 over 2.4 in another), making it difficult to infer a stable underlying value. These inconsistencies appeared to weaken the induced prior and increase uncertainty.

In Experiment 3c ($n = 160$), we replaced the social learning task with an unambiguous manipulation for establishing an initial group stereotype. Participants were randomly assigned to either a concentrated or dispersed counter-stereotype evidence condition. To induce an initial stereotype, participants first viewed a single group-level competence estimate from each of three prior "participants", who had observed multiple aliens. These estimates (e.g., 2.6, 2.8, and 3.0 minerals out of 10) provided a mildly negative impression with room for updating stereotypes. Each estimate was shown for 10 seconds. After viewing these estimates from prior "participants", they completed the MCMCP task for 20 trials, based on which we computed their prior estimate of group-level competence.

After establishing an initial group stereotype, participants themselves observed mining records of seven aliens. To test whether perceived group-individual contingency was successfully manipulated, we asked participants the extent to which individual aliens' mining performance was driven by their group membership (eye color) and by individual differences, both rated on a 7-point Likert scale. Finally, participants completed the MCMCP task again, based on which we computed their updated estimate of group-level competence.

We confirmed that our manipulation was successful: Participants in the concentrated condition (more homogenous mining records) rated group membership as more influential on the individual aliens' mining capability than those in the dispersed condition, $t(158) = 5.77$, $p < .001$, Cohen's $d = 0.91$; while participants in the dispersed condition (more heterogenous mining records) rated individuals' own traits as more influential on the individual aliens' mining capability, $t(158) = -6.27$, $p < .001$, Cohen's $d = -0.99$.

Next, we examined the causal effect of attenuated perceived group-individual contingency on stereotype updating. We calculated each participant's capability estimate as the mean of their 20 MCMCP choices. A robust mixed ANOVA with the condition (concentrated vs. dispersed) as a between-subjects factor and the measurement time (prior vs. posterior) as a within-subjects factor revealed a significant main effect of measurement time, $F(1, 70.69) = 2786.46$, $p < .001$, and a significant condition by measurement time interaction, $F(1, 70.69) = 4.25$, $p = .043$. This interaction (see Fig. 8) indicated that the posterior estimate of the alien groups' mining capability was lower in the dispersed condition ($M = 4.77$, $SD = 0.68$) than in the concentrated condition ($M = 4.93$, $SD = 0.53$). Together, these results suggest that when people hold a relatively strong group stereotype learned from social transmission, observing more stereotype-inconsistent group members whose behaviors are heterogeneous hinder group stereotype from updating, consistent with stereotype updating in hierarchical-Bayesian agents.

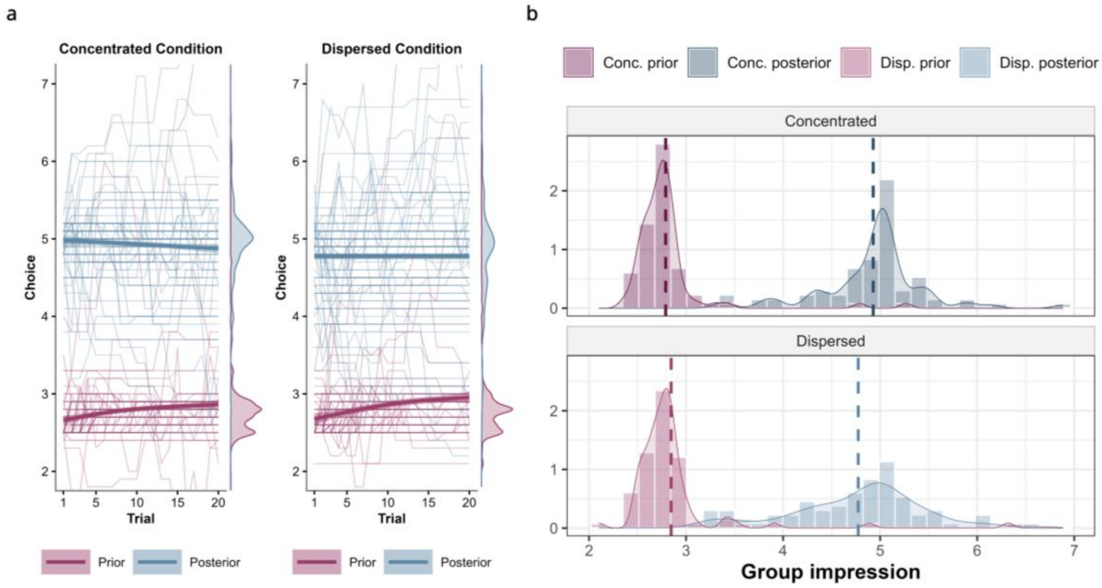

Figure 8. Dispersed counter-stereotype evidence across individuals hindered stereotype updating in human participants in Experiment 3c. a, The chain of selected choices in Markov Chain Monte Carlo with People (MCMCP) for prior and posterior measures. Each line represents one participant; the bold line shows the overall trend estimated by a generalized additive model. Marginal densities were computed from all individual choices across participants. **b,** Distributions of participants' mean choices across 20 trials in the MCMCP task, with prior and posterior shown in different colors. Dashed lines indicate distribution means. Although priors were similar across conditions, the dispersed condition elicited a lower posterior mean and greater variance.

Support for hierarchical inference 4: Stereotypes persisted when stereotype-inconsistent group members hold a secondary group membership

Experiment 3 shows that stereotypes are hard to change when people perform hierarchical Bayesian inference because social information is adaptively allocated across levels (group vs. individuals) depending on context—a central computational principle underlying hierarchical Bayesian inference. Here, we further test a fourth prediction of hierarchical Bayesian inference based on this context-dependent adaptive information allocation: whether stereotype-inconsistent evidence is preferentially attributed to uncertain alternative parameters (e.g., a secondary group membership) rather than the established stereotypes of the target group.

Specifically, hierarchical Bayesian inference predicts that multiple group memberships help preserve stereotypes: when individuals belong to several social groups (e.g., a person can be both male and Californian), stereotype-inconsistent evidence is attributed to the group with less established stereotypes, thereby protecting established stereotypes from changing. For example, an individual with strong stereotypes about men but weak stereotypes about Californians may attribute a men-stereotype-inconsistent observation from a Californian man to "Californians" rather than updating their beliefs about men.

Simulation study 4a. Similar to Study 3, we simulated belief updating in hierarchical Bayesian perceivers in response to stereotype-inconsistent evidence from individuals who belong to two groups.

The hierarchical Bayesian perceiver's task was to estimate the mining capability of an alien group. When an alien belongs to multiple groups, the prior belief that this perceiver holds about the alien can be conceptualized as a mixture of multiple group stereotypes. Based on this assumption, we simulated perceivers making inferences based on a hierarchical beta mixture model. For the first group (Group A), we specified a relatively strong prior to represent an established stereotype. For the second group (Group B), we specified a weak prior to capture a less established group impression. To test generalizability, we manipulated the number of group members encountered, observations per targets, and prior stereotype strength, yielding 45 conditions. In each condition, we also simulated a control model where targets belonged to only one group (e.g., Group A). The dependent variable was the update difference score: the difference in the posterior mean of the group-level estimate for Group A between the multi-group and single-group conditions. Positive values indicate greater stereotype rigidity (i.e., less stereotype update for Group A) due to multi-group memberships.

We first conducted a one-sample t-test on the update difference score to examine stereotype update. The update difference was significantly greater than zero, $t(44) = 14.60$, $p < .001$, Cohen's $d = 2.17$, indicating weaker stereotype update for Group A when a second group membership (Group B) was present. Consistent with this finding, the estimated mixing ratio assigned to the stereotyped group (Group A) in the beta mixture model was significantly below 50%, $t(44) = -24.50$, $p < .001$, Cohen's $d = -3.65$, indicating that impressions of new individuals were less influenced by the stereotyped group (e.g., less than 50%). A linear regression predicting the update difference score from prior strength, number of targets, and observations per targets showed that all predictors were significant (all $ps < .025$), but their individual impacts remained small (all $\beta s < .008$). The large intercept ($\beta = 0.227, p < .001$) indicated that stereotype persistence was primarily driven by the presence of the additional group membership. In summary, hierarchical Bayesian perceivers are reluctant to update established stereotypes when observing stereotype-inconsistent individuals with an additional group membership.

Experiment 4b-4c. We next tested whether human perceptions were consistent with those of hierarchical-Bayesian perceivers (n = 100 in Experiment 4b, n = 120 in Experiment 4c). Participants were randomly assigned to either a multi-group or single-group condition and read a cover story similar to Study 3 (Fig. 9). In the single-group condition, participants were told that they would observe and estimate competence of an alien group (e.g., the red-eye group). In the multi-group condition, participants were additionally told that the aliens lived in different areas (desert vs. forest), regardless of eye color. Thus, each alien belonged to two groups simultaneously (eye color and living area), allowing impressions of an alien to be contingent on both group memberships.

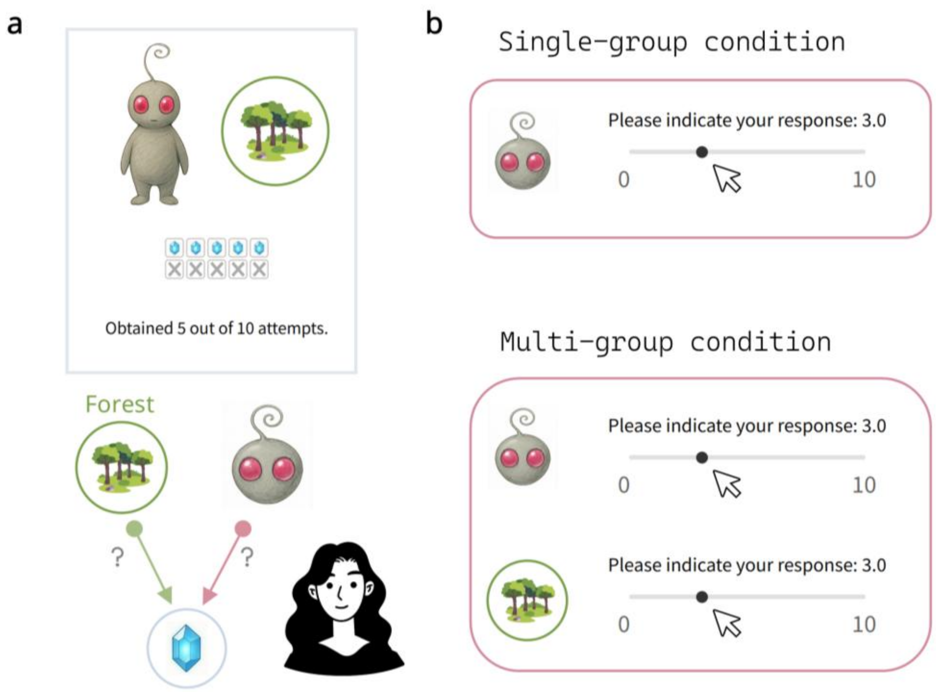

Figure 9. Experiment design for learning mining capability of aliens in Exp 4b-4c . a, After learning an initial stereotype of the red-eyed alien group, participants in the multi-group condition observed mining records of individual aliens with additional information about their living area. This allowed potential attributions of stereotype-inconsistent evidence to two groups (e.g., eye color and living area). In the single-group condition, no information about living area was provided. **b,** After observing all aliens' mining records, participants indicated their beliefs about the alien

group's overall mining competence. In the multi-group condition, participants indicated their beliefs about the competence of both groups, independently.

Similar to Experiment 3c, we induced initial stereotypes by showing participants estimates given by three prior "participants". Based on these reports, participants provided their prior estimate of the group's average mining capability using a slider from 0 to 10. Next, participants observed five aliens' mining records, which collectively countered the initial stereotypes. In the single-group condition, mining records were shown with the alien. In the multi-group condition, mining records were shown with the alien and its living area, with the first alien always belonging to a different living area than the remaining four. This ensured the living-area group membership was salient. After observing all five aliens' mining records, participants rated the extent to which mining capability was influenced by eye color, living area, and individual differences on a 7-point Likert scale. Subsequently, they provided an updated estimate of the group's capability (Fig. 9b). Participants in the multi-group condition also rated the mining capability of the aliens from the living area of the last four aliens (e.g., forest aliens).

The manipulation check indicated successful manipulation: In Experiment 4b, participants in the multi-group condition attributed mining capability less strongly to eye color than participants in the single-group condition, $t(98) = -2.41$, $p = .011$, Cohen's $d = -0.48$. A robust ANOVA using condition (single vs. multi-group) as between-subject factor and measurement time (prior vs. posterior) as the within-subject factor on the group mining capability estimation revealed a main effect of measurement time, $F(1, 60.74) = 5.92$, $p < .001$, and a significant interaction effect between condition and measurement time, $F(1, 60.74) = 5.92$, $p = .018$. This interaction (Fig. 10) indicates that posterior estimates of the red-eyed alien group's mining capability were lower in the multi-group condition ($M = 4.79$, $SD = 1.62$) than in the single-group condition ($M = 5.21$, $SD = 1.09$). Experiment 4c replicated this stereotype rigidity under multi-group membership, $F(1, 58.90) = 10.09$, $p = .002$. Together, when people hold a relatively strong stereotype about a target group, observing stereotype-inconsistent individuals who have other group memberships that the perceiver does not hold strong stereotypes about hinders the updating of beliefs about the target group, consistent with hierarchical Bayesian inference.

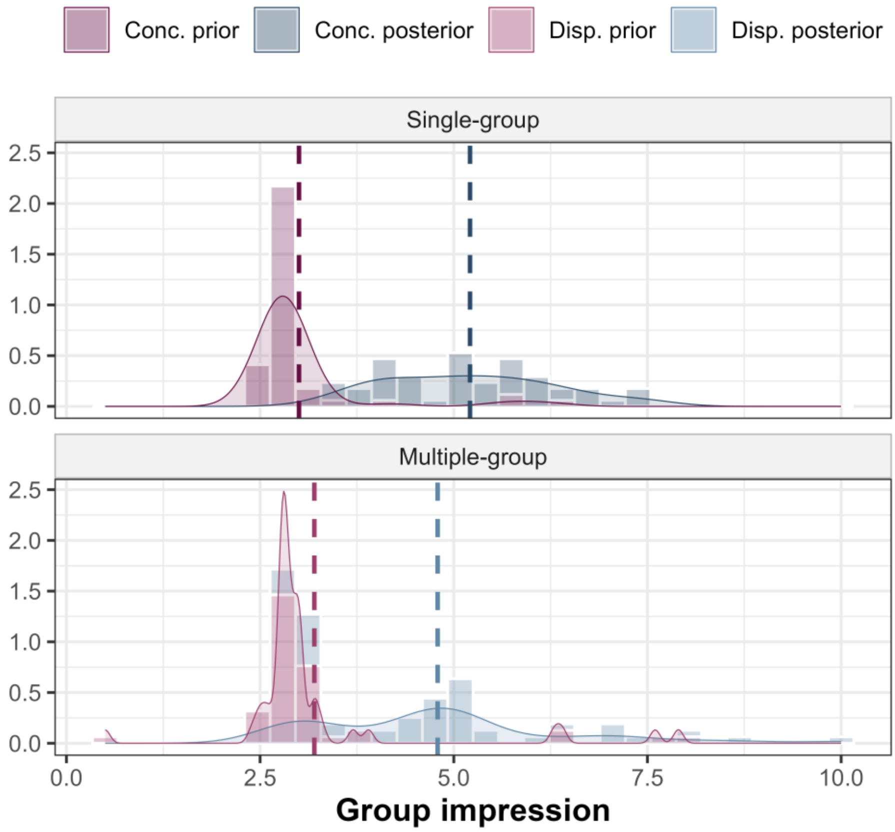

Figure 10. Multiple group memberships hindered stereotype updating in human participants. Distributions of participants' estimate of group mining capability in Experiment 4b, with prior and posterior distributions shown in different colors. Dashed lines indicate distribution means. While priors were comparable across conditions, the multi-group condition elicited a lower posterior mean, indicating less stereotype updating in response to stereotype-inconsistent behavior from multi-group aliens.

Discussion

In the current research, we propose and test a novel account of stereotyping—stereotypes emerge as byproducts of maximizing perceptual accuracy about individual targets, a process formalized as hierarchical Bayesian inference. Hierarchical Bayesian inference helps perceivers reduce the cognitive cost of representing individuals separately by assuming that individuals with the same group membership have traits drawn from a common distribution. As a byproduct, hierarchical inference facilitates biased beliefs about the group in contexts where the assumed hierarchical structure does not match the true data-generating process (i.e., when individuals' traits are not dependent on their group membership).

Specifically, hierarchical Bayesian inference makes four specific predictions about how stereotypes are formed and resist changes. Across four studies (5 computational simulations and 7 human experiments), we found empirical support for all four predictions. First, consistent with hierarchical Bayesian inference, human participants' stereotype strengths were positively correlated with asymmetrical exposures to members from different groups (Study 1; Fig. 3). Second, even without asymmetric exposure, stereotypes were formed (e.g., more confident group-level impressions than individual-level impressions) when human participants had shallow observations of many group members (Study 2; Fig. 6). Third, after acquiring a group stereotype, human participants were reluctant to update it when observing heterogeneous, stereotype-inconsistent members (Study 3; Fig. 8), and fourth, when those members held a secondary group membership (Study 4; Fig. 10). Together, these findings support that people may form and maintain stereotypes in a manner consistent with hierarchical Bayesian inference, providing a novel computational explanation for why stereotypes are easily formed and hard to change despite stereotype-inconsistent evidence.

Our findings on hierarchical Bayesian inference as a computational mechanism underlying stereotyping connect to a broader principle of human cognition, spanning domains such as language acquisition[48], causal learning[49], and category learning[46,50], in which knowledge is organized across multiple levels of abstraction. For example, in early visual learning, infants gradually build object representations from fine-grained features and parts accumulated through experience[49,50]. Similarly, we argue that social inference operates across multiple levels of abstraction: individual-level observations update group-level representations, and group-level representations (i.e., stereotypes) in turn guide inferences about individuals.

Such hierarchical process reflects a general cognitive mechanism that supports efficient learning[33,36,46]. When the hypothesis space (e.g., about an individual's traits) is large and unconstrained, learning is costly because substantial data are required to eliminate alternative hypotheses. Hierarchical inference reduces this cost by adaptively learning higher-level knowledge (e.g., group impressions) that constrain the hypothesis space (e.g., individual impressions)[33,46]. For instance, observing a few members of a novel group may be sufficient to infer the group's competence, which guides subsequent inferences about other group members. When a new group member deviates from this expectation, the perceiver can readily attribute the discrepancy to individual-specific factors (e.g., the individual's own traits, or their other group memberships). In addition, imposing a prior at the higher group-level minimizes prediction error by preventing the inference at the individual-level being overly sensitive to noisy or ambiguous cues in social context. For instance, when a Michelin-star chef overcooked a steak on a particular occasion, a prior belief that Michelin-star chefs are good at cooking helps attribute the current stereotype-inconsistent observation to the specific circumstance rather than overturning the belief about Michelin-star chefs. In the absence of such prior, higher-level inferences would be easily swayed by momentary observations. However, this buffer to group-level beliefs readily explains why group stereotypes can be hard to change.

Our research also contributes to the growing literature on understanding the nature of stereotypes. Prior research has predominately argue that stereotypes are products of automatic categorization [11,51]. However, this account falls short of explaining how these representations arise computationally and how they shape social inferences of individuals in daily life. Here, we argue that the mental representations of people are hierarchical, connecting both the individual-level and group-level representations. Crucially, this hierarchy reflects the perceiver's mental representation rather than the true generative structure of social traits. Because social groups are often culturally constructed rather than biologically grounded[38], imposing hierarchical structure for reasons of efficiency thus introduces systematic biases. In our experiments (2b, 3b-c, 4b-c), stereotypes emerged and persisted even when individuals' traits did not depend on their group membership (Fig. 6, 8, 10).

The findings that human stereotypes were associated with asymmetric exposure in a manner predicted by hierarchical Bayesian inference (Study 1; Fig. 3) builds on Bayesian accounts of illusory correlation[43] and recent biased-sampling theories of stereotyping[16]. As noted by Bai et al.[16], one limitation of biased-sampling accounts is that they predict idiosyncratic stereotypes. If group impressions are driven by early interactions such that perceivers form favorable impressions of initially encountered groups while neglecting other groups, then which groups are encountered first depends on each perceiver's unique experience. This would lead to substantial variability in stereotypes across perceivers, contrary to empirical evidence of shared cultural stereotypes[51]. Our account addresses this limitation by linking shared environmental structures across perceivers within the same society to stereotypes. For example, when the relative proportion of social groups is fixed, different perceivers will have similar exposures to each group (Study 1b) and are therefore expected to form more similar group impressions as shown in our supplementary analyses (Table S1-S2).

While asymmetric exposure is not uncommon in everyday life, it is not required for stereotype formation. Consistent with hierarchical inference, Study 2 found that stereotypes emerged when participants had shallow observations of many group members (Fig. 6) because group-level impression formation outpaced individual-level impressions. Statistically, limited observations within a cluster (e.g., limited mining records of an individual group member) cannot support reliable estimates at the individual level. Hierarchical Bayesian inference addresses this by pooling information across individuals to generate group-level estimates that facilitate individual-level learning. This pooling naturally accelerates group-level learning. Consistent with this prediction, participants formed more confident group-level impressions when they only had shallow observations of individual group members (Fig. 6). These findings help explain why stereotypes are acquired so readily. While we explicitly assigned individuals to groups in the experiments (e.g., red-eyed aliens), in the real world people may spontaneously categorize others into groups based on verbal and visual cues[43,52–56], creating the conditions needed for forming stereotypes under hierarchical inference.

Beyond explaining how stereotypes are formed, hierarchical Bayesian inference also explains how stereotypes are updated. Recent work argues that stereotypes resist to update because self-reinforcing behaviors, such as avoidance, that limit one's exposure to stereotype-inconsistent information [19,21]. Here, we showed that even when participants were exposed to stereotype-inconsistent information, they still resisted to update stereotypes (Studies 3-4; Figs. 8 and 10). In both studies, once stereotypes were established through social transmission[57–59], new evidence was preferentially used to update less certain parameters to maximize information gain—a common Bayesian principle. Consequently, stereotype-inconsistent evidence was attributed to individual-level traits in Study 3 and to the secondary, unfamiliar group membership in Study 4, both of which hindered stereotype updating.

The interplay between group- and individual-level impressions shown in Study 3 was consistent with the continuum model of impression formation[11] and the broader literature on individuation [40,41]. These theories posit that category-based processes activating group-level stereotypes and attribute-based processes focusing on individuating information can operate in tandem, potentially through shifts in attention [40]. Our approach extends these theories by providing a precise computational account of how this dynamic interplay unfolds and when stereotypes are maintained despite conflicting information between categories and individuals. Attenuated group-individual contingency, however, is not the only mechanism that protects stereotypes from counterevidence. In other contexts, group representations can be modified, as reflected in processes such as subtyping and subgrouping[42,60]. These reorganizations also connect to a Bayesian interpretation proposed in recent studies [43,55]: perceivers tradeoff between simpler representations and those that better fit the data, "fencing off" deviant individuals into new clusters (subtypes) to preserve the statistical integrity of the original group as deviations increase. Our hierarchical Bayesian account extends these studies by showing that even when group representations remain unchanged, stereotypes were sustained by attributing stereotype-inconsistent evidence to individual-level impressions (Fig. 8).

Building on the same computational principle that stereotype-inconsistent evidence can be preferentially used to update uncertain parameters, Study 4 showed why stereotypes were resistant to change when individuals belong to multiple group memberships [61,62], advancing the literature on crossed categorization and intersectional stereotypes[63–65]. Prior theories on crossed categorizations (e.g., black males) struggle to formalize computations of multiple group impressions [63]. Our findings address this gap by providing a formal computational account: when targets belong to two groups (e.g., red-eyed and from a forest), participants attributed stereotype-inconsistent evidence disproportionately to the more uncertain group membership, hindering the update of established stereotypes (Fig. 10). This finding is consistent with the recent lens-based account of intersectional stereotypes, which proposes that the category or lens applied in the current context depends on how well it fits the observed data [64].

The computational process of how people attributed stereotype-inconsistent evidence to a secondary category in Study 4 also speaks to feature competition in instrumental learning [66,67], where shared outcomes across predictors weaken learned contingencies between cues and rewards. However, our account is theoretically distinct from an instrumental learning perspective on stereotypes [57,67,68], which characterizes social learning as reward-contingent and conceptualizes stereotypes as estimates of state values (e.g., group traits) from feedback. As shown in Study 2, holding total success rate constant, learning depended on how evidence was distributed across group members and on the perceiver's group-level priors (Fig. 6)—patterns that cannot be captured by reward-based updating alone. A key contribution of our hierarchical Bayesian account is that perceivers form structured representations, enabling flexible inference and efficient social learning [36]. Indeed, Bayesian learning outperform simple heuristics and some model-free reinforcement learning algorithms in dynamic environments [69], as the absence of explicit mental representations limits generalization across contexts and delays updating when environments change.

With these theoretical contributions in mind, we highlight a few limitations that constrain the generalizability of the conclusions from this research. First, we used a simplified paradigm in which information about individuals was numerically represented, whereas in the real world, individual impressions and group stereotypes are also encoded in richer semantic spaces. Second, to minimize cognitive load, participants in our experiments formed impressions of individuals and groups with limited individual-level information, which may have attenuated individual-level impression formation. Future research could implement more realistic designs that provide richer individual-level information. Third, our analysis operates at the computational level and does not specify the underlying psychological processes or neural implementations. Prior work has proposed heuristic[70] and neural approximations [71] to Bayesian inference. Whether such mechanisms or related mental structures in social inference[4,72] extend to hierarchical inference in stereotyping remains an important direction for future research.

Despite these limitations, the present research proposed and empirically demonstrated a novel computational account of stereotypes: stereotypes emerged as byproducts of hierarchical Bayesian inference, which helps maximize perceptual accuracy of individual targets but facilitate the formation and persistence of biased group stereotypes. This account reconciles the literature on group stereotypes which typically focus on groups and the literature on person perception which typically focuses on individuals, providing an integrative framework for understanding mental representations of other people in naturalistic contexts. Our findings also provide a novel explanation for why stereotypes can form when exposure to different groups is relatively symmetrical and sufficient, and why stereotypes can persist when counter-stereotype evidence is inevitable.

Methods

All empirical studies were approved by the Institutional Review Board at the University of California, San Diego (808857) and Columbia University (AAAV8633), and all participants gave informed consent in accordance with approved procedures before starting the experiments.

Study 1

Simulations. In Study 1a, we simulated a segregated society of three groups in NetLogo[73] using a Schelling model[74,75], in which agents preferentially assort and interact with ingroup members. The simulation ran for 200 iterations. At each iteration, agents relocated when the proportion of same-group neighbors (within their eight adjacent cells) fell below a threshold, gradually producing spatial clustering. We introduced a perceiver performing hierarchical Bayesian inference into the society at time T = 171, 181, or 191, which determined the travel time (30, 20, or 10 iterations, respectively). We also manipulated the similarity threshold (0.3, 0.5, 0.7, 0.9), population density (agents occupying 35%, 50%, 65%, 80%, or 95% of the environment), and the perceiver's travel radius (2, 4, 6, 8, or 10 units), yielding 900 simulation conditions. In each condition, the perceiver interacted with encountered individuals and observed binary social outcomes (positive vs. negative), which were used to update beliefs about a latent trait (e.g., warmth) at both the individual and group levels. In the hierarchical condition, individual traits were sampled from a Beta(16, 4) distribution, from which five social interaction outcomes were generated for each encountered group member. In the mismatched condition, outcomes were sampled from a common Bernoulli distribution independent of group membership; thus, no true hierarchical structure was present in the data-generating process. The perceiver's trajectory was generated using a Metropolis–Hastings algorithm with a bivariate normal proposal distribution. Trait values were defined on the interval [0, 1], and impressions were modeled using a Beta distribution. Stereotypes were estimated using a hierarchical beta–binomial model via Gibbs sampling:

$$
\begin{aligned}
y_i &\sim \text{Binomial}(n_i, \theta_i),\\
\theta_i &\sim \text{Beta}(\mu\kappa, (1-\mu)\kappa),\\
\mu &\sim \text{Beta}(2, 2),\\
\kappa &\sim \text{Gamma}(6, 0.3),
\end{aligned}
$$

where y_i denotes the number of positive outcomes observed for target i out of n_i interactions, θ_i represents the latent trait of target i, μ is the group-level mean trait, and κ controls the dependency of θ_i on μ. We defined the strength of stereotyping as the standard deviation of the posterior means of μacross the three groups. Social segregation was quantified as one minus the mean exposure rate, defined as the average proportion of outgroup members in individuals' local neighborhoods.

In Study 1b, we simulated societies with three groups. Group size imbalance is defined as $\sum_{k=1}^{K} S_k^2$, where S is the population share of the group. We generated 30 sets of group shares using a sparse Dirichlet distribution favoring unequal shares (Dirichlet($\alpha\beta$), $\alpha = 0.5, \beta = [1,1,1]$). Under the model-mismatched assumption where the trait was fixed across all target individuals and groups, we crossed these shares with a set of factors: number of targets encountered (5, 10, 20), number of interactions per target (2, 5, 10), the probability of positive social outcomes (θ: 0.2, 0.5, 0.8). This resulted in 810 conditions. For the model-matched assumption where the individual target's traits were sampled based on group-level distributions and thus the hierarchical Bayesian observer's mental model matches the reality, we crossed these shares with number of targets encountered (5, 10, 20), number of interactions per target (2, 5, 10), the probability of positive social outcomes (group-level mean μ: 0.2, 0.5, 0.8), and the variance across targets (concentration parameter κ: 4, 20, 100). This resulted in 2430 conditions. The Bayesian model and the definition of stereotype are identical to Study 1a.

Data collection and analyses. In Study 1c, we used the stereotype dispersion data across the 50 U.S. states from Bai et al., which were derived from ratings of perceived traits of 20 immigrant groups (e.g., Mexicans) provided by 1502 Amazon Mechanical Turk participants [44]. Stereotypes are defined as the average Euclidean distance between the perceived warmth and competence of *n* ethnic groups and their collective centroid $(\bar{x}_j, \bar{y}_j)$ in a two-dimensional space:

$$Stereotype\ Dispersion = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(x_i - \bar{x}_j)^2 + (y_i - \bar{y}_j)^2}$$

where, x_i and y_i represent the perceived warmth and competence ratings for each group i, while $(\bar{x}_j, \bar{y}_j)$ is the centroid representing the average of these dimensions across all groups in a specific state j.

For social segregation, we used the American Community Survey (ACS) from the U.S. census bureau at the census tract level, which documented detailed demographic information about the U.S. population. We used the ACS published in 2019 since it is temporally close to the publication time of Bai et al[44]. We collected the race-specific population counts of white, black, Indian, Asian, Hawaiian, and other groups using the tidycensus R package[76]. These data comprised 510,139 observations at the tract-by-race level, corresponding to 72,877 census tracts. We then organized the data by states. Social segregation for each state is described by a normalized segregation index ($D_{g,i}$), which quantifies the extent to which a specific social group is unevenly distributed across subunits relative to the total population in the state[77]:

$$D_{g,i} = \frac{\sum_m \left|\frac{N_{g,m,i}}{N_{m,i}} - \frac{N_{g,i}}{N_i}\right| \cdot N_{m,i}}{2 \cdot N_i \cdot \frac{N_{g,i}}{N_i} \cdot (1 - \frac{N_{g,i}}{N_i})}$$

where $N_{g,m,i}$ represents the population of group g in subunit m within the state i, while $N_{m,i}$ is the total population of subunit m. The variables $N_{g,i}$ and N_i represent the total population of group g and the total population of the state i, respectively. Therefore, the segregation index was computed separately for each ethnic group. We examined its association with stereotyping both at the group level and using an average segregation index across all groups.

Group size imbalance was also derived from the American Community Survey (ACS) 2019. For each state, we calculated the population share of each ethnic group by dividing its count by the total state population, and defined group size imbalance as the sum of squared group shares.

In Study 1d, we used data from the General Social Survey (GSS), a national survey of U.S. adults conducted repeatedly over time. We analyzed respondents' ratings of White and Black Americans on three traits measured on 7-point Likert scale: intelligence, work ethic, and wealth. Ratings were recoded so that higher scores reflect more negative stereotyping. These measures were available for 15 to 17 years between 1990 and 2024; other traits and groups were excluded due to limited temporal coverage (2–7 years). We also used the measure "liveblks", which assesses respondents' preferences regarding the racial composition of their neighborhood (1 = strongly favor, 5 = strongly oppose), to derive measures of social segregation and group size imbalance. The total sample included 75,699 respondents, with sample sizes varying across waves due to missing data. Stereotyping was operationalized as the Euclidean distance between mean ratings of Black and White Americans across the three traits within each Census division and year:

$$\sqrt{\sum_k (x_{k,\text{Black}} - _{k,\text{White}})^2},$$

where $x_{k,g}$ is the mean rating of group g on trait k. We then estimated a fixed effects panel regression predicting stereotypes from the two derived variables and a linear time trend, with Census division fixed effects to account for stable regional differences in both stereotyping and residential patterns.

Study 2

Simulations. The Bayesian perceiver's inference was simulated using the same hierarchical beta-binomial model implemented in Study 1c-1d. We used a factorial design to study the effect of shallow interaction. Total observations ($N_{total} \in \{8,16,32,64\}$) were crossed with different competence priors ($a \in \{2,4,6,8\}$), and the number of encountered group members ($N_{target} \in \{N_{total}, N_{total}/2, 2\}$) to manipulate observations per target. This resulted in 48 conditions. Across these 48 primary conditions, we used integer partitioning to generate all valid distributions of mining successes across targets, yielding 177 unique conditions. For each unique condition, we calculated the posterior SD of μ as a measure of certainty (smaller SD = stronger stereotype) and a dispersion score (SD of success counts across individuals). We examined the effects of these manipulations using a linear mixed model with a random intercept of primary conditions.

Experiment. The experiment was preregistered (https://osf.io/9w5pq/overview?view_only=972af1882aef4f96a6d314e827642835). We preregistered both ANOVA and linear mixed-effects models for the main analyses. However, after data collection, the linear mixed models failed to converge due to low intraclass correlations, making the inclusion of random effects statistically unjustified. Therefore, we report only the ANOVA results. The same issue occurred in Experiment 3a, so linear mixed models were not used either.

We recruited 365 participants from CloudResearch Connect (age: M = 40.40, SD = 12.90; 48.5% females). To align with the simulation, participants inferred the underlying competence from observed mining records of multiple aliens from the same group, with the total number of observations (e.g., mining attempts) across targets fixed at 32. We manipulated dispersion (low vs high) and the number of targets (2 vs 16), yielding 4 conditions (low-2: [7, 9]; high-2: [4, 12]; low-16: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 0, 0]; high-16: [2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 1, 1]).

Participants were randomly assigned to one of the 4 conditions. They were asked to estimate the competence of a single alien group in extracting blue minerals. Specifically, participants viewed mining records from multiple individual aliens from the same group and then provided their estimation of competence. Each alien's mining record consisted of multiple attempts, with each attempt resulting in either one mineral (success) or nothing (failure). Regardless of the number of targets, two mining attempts were displayed per screen to equate task length across conditions. For example, in the 2-target condition, each alien's 16 attempts were shown across 8 screens. For each alien, the alien image was initially presented for 1.2-1.6 s, then simultaneously with its mining outcomes for 3 s, followed by a 2 s transition (inter-trial interval) before the next alien.

After viewing all records, participants rated the competence of the last individual alien they observed and the group as a whole on a 7-point Likert scale (“Based on your observations, how competent the alien group with red eyes is at mining the target mineral?”). They then rated their confidence of their judgment (“How confident you are about your prior judgment?”). The order of individual and group judgments was randomized.

Study 3

Simulations. We implemented the beta-binomial model as in Study 2 to simulate belief updating in response to stereotype-inconsistent evidence. A prior stereotype was specified by assigning the group average μ a Beta(10, 20) distribution, reflecting a strong belief of low competence (mean ≈ 0.33), and placing a weak Gamma(0.01, 0.01) prior on κ. The Bayesian perceiver then derives posterior belief by observing new records from multiple group members and updating both individual- and group-level impressions. In a factorial design, we manipulated the number of group members ($N_{target} \in \{5,10,15\}$), observations per targets ($N_{obs\_each} \in \{2,5,8\}$), and the distribution of ground-truth competence ($a \in \{30,60,90\}$ for Beta(a, 20)). This resulted in 27 primary conditions. For each condition, we enumerated all possible evidence configurations via integer partitioning, spanning perfectly uniform to highly heterogeneous distributions. For example, if there were 5 extracted minerals under the condition $N_{target} = 5, N_{obs\_each} = 2$, a list of more concentrated observations would be [1,1,1,1,1], while a dispersed one would be [2,2,1,0,0]. This resulted in 11,539 finer conditions. Dispersion was quantified as the variance of observations. The dependent measure was an update score, defined as the posterior minus prior mean of μ; smaller values indicate stronger stereotype persistence.

Experiments. In Experiment 3a, we recruited 333 participants from CloudResearch Connect (age: *M* = 38.70, *SD* = 11.50; 49% females). This study was preregistered (https://osf.io/c8bq3/overview?view_only=ffd45015ec4c42b5a7e6905249ab0e1f). Participants read the cover story similar to Study 2 and asked to estimate the mining capability of an alien group. Different from Study 2, the competence metric was standardized and defined as the average number of minerals collected in 10 attempts.

Participants' prior beliefs about the group (i.e., established stereotypes) were induced using a social learning task. Participants were told that, to estimate the group's mining capability, they would observe evaluations made by prior participants who had viewed multiple aliens from the group. Their task was to infer these prior participants' judgments and use feedback to learn about the group's competence. For each prior participants, current participant completed 10 choices, resulting in 40 social learning trials in total. In each trial, they saw two possible estimates of the group's performance and guessed which one had been chosen by the prior participant. The true choice was revealed immediately as feedback. Participants were informed that values could include decimals because they represented group averages. Unbeknownst to participants, the prior participants' choices were generated using a Metropolis–Hastings algorithm, a Markov chain Monte Carlo method commonly used to approximate posterior distributions[37]. We specified a proposal distribution of Normal(2.5, 0.5) to simulate a relatively low success rate in mineral extraction, thereby leaving room for participants to update their group-level stereotypes when they later observe individuals with higher success rates. All participants are presented with the same sequence of choices.

After finishing the 40 social learning trials, participants viewed mining records of 10 attempts (e.g., extracted 5 minerals out of 10 attempts) for multiple aliens from the group depending on their assigned conditions. For each alien, the alien image was initially presented for 2 s, then simultaneously with its mining outcomes for 4 s, followed by a 3 s transition before the next alien. Mining records were specified as: concentrated/5 aliens [4, 5, 5, 6, 5], dispersed/5 aliens [1, 4, 5, 7, 8], concentrated/10 aliens [4, 5, 5, 6, 5, 4, 5, 5, 6, 5], and dispersed/10 aliens [1, 4, 5, 7, 8, 1, 4, 5, 7, 8].

After viewing these estimates, participants' group-level beliefs were elicited using Markov Chain Monte Carlo with People (MCMCP), a method for eliciting mental representations in the form of probability distributions [47]. In MCMCP, participants completed a two-alternative forced-choice task consisting of 20 trials. In each trial, they indicated which of two values better reflected the average number of minerals extracted by red-eye aliens in 10 attempts. The options were generated using the Metropolis-Hastings algorithm. On the first trial, one option was randomly generated, and the other was sampled from a normal distribution centered on the value of the other option; on subsequent trials, the option selected in the previous trial was retained while the alternative option was resampled around it. All values were discretized to one decimal place, and the resulting choices approximate samples from participants' posterior belief over group-level competence.

In Experiment 3b, we recruited 160 participants (age: *M* = 41.60, *SD* = 11.90; 46% females). Based on the feedback from eight participants from Experiment 3a, we simplified the social learning task. Participants were assigned to either a concentrated or dispersed condition. To induce a prior stereotype, participants viewed three verbal estimates provided by prior participants who had observed the 10-attempt mining records of multiple aliens and reported the group's average mining ability. These estimates were fixed at 2.6, 2.8, and 3.0 minerals (out of 10), each presented for 10 seconds. After this social learning phase, participants completed an MCMCP task to elicit their prior beliefs. Participants then observed mining records from seven aliens, each consisting of outcomes from 10 attempts. In the concentrated condition, the records were either [4, 6, 4, 6, 5, 5, 5] or [4, 6, 5, 5, 5, 5, 5]. These sequences had low variance and were presented in random order for each participant. In the dispersed condition, the records exhibited high variance and were fixed as [0, 2, 4, 5, 7, 8, 9]. Trial order was randomized with the constraint that adjacent values were not highly similar. Importantly, the mean number of extracted minerals was identical across conditions.

Once finished viewing all mining records, participants indicated the extent to which mining performance was driven by the group ("Based on your impressions, to what extent did the aliens' category influence how many minerals the aliens extracted?") versus individual differences ("To what extent do you think stable individual abilities explain the differences in how many minerals the aliens extracted?") on two 7-point Likert items. Finally, participants completed MCMCP again to provide their posterior estimation.

Study 4

Simulations. When individual targets belong to multiple groups, impressions of them can be conceptualized as a mixture of multiple stereotypes. To formalize this idea, we specified a hierarchical beta-binomial model in which individual-level impressions are generated as a mixture of two group-level Beta distributions. For the first group (Group A), we imposed a relatively strong prior to represent an established stereotype. This prior was parameterized as Beta(a, 3a), where *a* controls the prior strength. For the second group (Group B), we specified a weak prior, $\text{Beta}(1,1)$, to capture a less established or uninformative group impression. This hierarchical mixture model was specified as:

$$\begin{aligned} y_i &\sim \text{Binomial}(n_i, \theta_i) \\ \theta_i &\sim \textstyle\sum w_i \,\text{Beta}(\mu_i\kappa, (1-\mu_i)\kappa) \\ \mu_1 &\sim \text{Beta}(a, 3a), \mu_2 \sim \text{Beta}(1,1) \\ \kappa &\sim \text{Gamma}(6, 0.2) \end{aligned}$$

The notations follow the beta-binomial model in Study 1. The new parameter, w, represents the mixing ratio, specifying the weight of each group in shaping the individual-level impression. We used a factorial design manipulating the number of group members ($N_{target} \in \{5,10,15\}$), observations per targets ($N_{obs\_each} \in \{2,5,10\}$), and prior stereotype strength ($a \in \{1,2,5,13,30\}$), yielding 45 conditions. In each condition, we also simulated a control model where individuals belonged to only one group. The dependent variable was the update difference score: the difference in the posterior mean of the group-level estimate between the multi-group and single-group models. Positive values indicate greater stereotype rigidity due to multi-group memberships.

Experiments. We recruited 100 participants in Experiment 4b (age: M = 41.30, SD = 11.60; 56% females) and 120 participants in Experiment 4c (age: M = 40.80, SD = 13.60; 54% females) from CloudResearch Connect. Experiment 4c was preregistered (https://osf.io/j2znb/overview?view_only=e32dbc09669d47dabc236864cfc1a35a). Participants were randomly assigned to a multi-group or single-group condition and asked to estimate the mining competence of aliens. In the single-group condition, participants observe only one alien group based on the eye color (e.g., the red-eye group). In the multi-group condition, participants were further told that the aliens lived in one of two different areas (desert vs. forest), regardless of eye color. In the single-group condition, all aliens were described as inhabiting the same living area (desert), rendering living area non-diagnostic.

Participants viewed competence estimates of the eye color defined group provided by three prior participants, which are fixed to be 2.6, 2.8, or 3.0 minerals out of 10 attempts. They then indicated their initial estimate of the group's average mining capability using a slider ranging from 0 to 10. Next, participants observed mining records from five aliens ([3, 7, 6, 7, 5]). The first record (3) was always presented first and was consistent with the prior; the remaining records were shuffled. In the single-group condition, only the mining records were shown; in the multi-group condition, the mining records were shown together with the alien's living area, with the first alien always belonging to a different living area (e.g., desert) than the remaining four (e.g., forest). As a manipulation check, participants then rated the extent to which mining capability was influenced by eye color, living area, and individual differences in a 7-point Likert scale. Finally, they provided an updated estimate of the alien group's (e.g., the red-eye group) capability using the slider ranging from 0 to 10. In the multi-group condition, participants also rated the mining capability of aliens living in an area corresponding to the last four aliens they saw.

References

1. Uleman, J. S., Adil Saribay, S. & Gonzalez, C. M. Spontaneous Inferences, Implicit Impressions, and Implicit Theories. *Annu. Rev. Psychol.* **59**, 329–360 (2008).
2. Sutherland, C. A. M. & Young, A. W. Understanding trait impressions from faces. *British J of Psychology* **113**, 1056–1078 (2022).
3. Olivola, C. Y., Funk, F. & Todorov, A. Social attributions from faces bias human choices. *Trends in Cognitive Sciences* **18**, 566–570 (2014).
4. Lu, J. & Lin, C. A Small-World Mind: Towards a General Computational Principle of Social Cognition across Contexts. Preprint at https://osf.io/preprints/psyarxiv/zsh6u_v2/ (2026).
5. Olivola, C. Y. & Todorov, A. Elected in 100 milliseconds: Appearance-Based Trait Inferences and Voting. *J Nonverbal Behav* 28 (2010).
6. Willis, J. & Todorov, A. First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face. *Psychol Sci* **17**, 592–598 (2006).
7. Olivola, C. Y., Tingley, D., Bonica, A. & Todorov, A. The donkey in elephant's clothing: revisiting the impact and validity of political facial stereotypes. Pittsburgh, PA: Tepper Sch. Bus. (2014).
8. Todorov, A., Olivola, C. Y., Dotsch, R. & Mende-Siedlecki, P. Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annu. Rev. Psychol.* **66**, 519–545 (2015).
9. Ashmore, R. D. & Del Boca, F. K. Sex stereotypes and implicit personality theory: Toward a cognitive—Social psychological conceptualization. *Sex roles* **5**, 219–248 (1979).
10. Eagjy, A. H., Ashmore, R. D. & Makhijani, M. G. What Is Beautiful Is Good, But. . .: A Meta-Anatytic Review of Research on the Physical Attractiveness Stereotype. *Psychological Bulletin* **110**, 109–128 (1991).
11. Fiske, S. T. & Neuberg, S. L. 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 1–74 (Elsevier, 1990).
12. Dobs, K., Isik, L., Pantazis, D. & Kanwisher, N. How face perception unfolds over time. *Nat Commun* **10**, 1258 (2019).
13. Lippmann, W. *Public Opinion*. (Routledge, 1922).
14. Dovidio, J. F., Hewstone, M., Glick, P. & Esses, V. M. Prejudice, Stereotyping and Discrimination: Theoretical and Empirical Overview. in *The SAGE Handbook of Prejudice, Stereotyping and Discrimination* 3–28 (SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom, 2010). doi:10.4135/9781446200919.n1.
15. Allport, G. W. *The Nature of Prejudice.* (Addison-wesley, 1954).
16. Bai, X., Fiske, S. T. & Griffiths, T. L. Globally Inaccurate Stereotypes Can Result From Locally Adaptive Exploration. *Psychological Science* (2022).

17. Bai, X., Griffiths, T. L. & Fiske, S. T. Costly exploration produces stereotypes with dimensions of warmth and competence. *Journal of Experimental Psychology: General* **154**, 347–357 (2025).
18. Cuddy, A. J. C. *et al.* Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology* **48**, 1–33 (2009).
19. Allidina, S. & Cunningham, W. A. Avoidance begets avoidance: A computational account of negative stereotype persistence. *Journal of Experimental Psychology: General* **150**, 2078–2099 (2021).
20. Harris, C., Fiedler, K., Marien, H. & Custers, R. Biased preferences through exploitation: How initial biases are consolidated in reward-rich environments. *Journal of Experimental Psychology: General* **149**, 1855 (2020).
21. Denrell, J. Why Most People Disapprove of Me: Experience Sampling in Impression Formation. *Psychological Review* **112**, 951–978 (2005).
22. Le Mens, G. & Denrell, J. Rational learning and information sampling: On the "naivety" assumption in sampling explanations of judgment biases. *Psychological Review* **118**, 379–392 (2011).
23. Costello, F. & Watts, P. The rationality of illusory correlation. *Psychological Review* **126**, 437–450 (2019).
24. Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. (MIT press, 1982).
25. Jussim, L. *et al.* Stereotype accuracy: One of the largest and most replicable effects in all of social psychology. in *Handbook of prejudice, stereotyping, and discrimination* 31–63 (Psychology Press, 2015).
26. Hamilton, D. L. & Gifford, R. K. Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology* **12**, 392–407 (1976).
27. Meiser, T. & Hewstone, M. Illusory and spurious correlations: distinct phenomena or joint outcomes of exemplar-based category learning? *Euro J Social Psych* **36**, 315–336 (2006).
28. Eyal, T. & Epley, N. Exaggerating Accessible Differences: When Gender Stereotypes Overestimate Actual Group Differences. *Pers Soc Psychol Bull* **43**, 1323–1336 (2017).
29. Terracciano, A. *et al.* National Character Does Not Reflect Mean Personality Trait Levels in 49 Cultures. *Science* **310**, 96–100 (2005).
30. Efron, B. & Morris, C. Data Analysis Using Stein's Estimator and its Generalizations: Journal of the American Statistical Association: Vol 70 , No 350 - Get Access. *Journal of the American Statistical Association* https://www.tandfonline.com/doi/abs/10.1080/01621459.1975.10479864 (1975).
31. Ma, W. J., Kording, K. P. & Goldreich, D. *Bayesian Models of Perception and Action: An Introduction*. (The MIT Press, 2023).
32. Efron, B. & Morris, C. Stein's paradox in statistics. *Scientific American* **236**, 119–127 (1977).
33. McElreath, R. *Statistical Rethinking*. (Chapman and Hall/CRC, 2015).

34. Kruschke, J. K. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. (Academic Press, Boston, 2015).
35. Matsuura, K. *Bayesian Statistical Modeling with Stan, R, and Python*. (Springer Nature Singapore, Singapore, 2022). doi:10.1007/978-981-19-4755-1.
36. Griffiths, T. L., Chater, N. & Tenenbaum, J. B. *Bayesian Models of Cognition: Reverse Engineering the Mind*. (MIT Press, 2024).
37. Gelman, A. *et al. Bayesian Data Analysis*. (CRC Press, Boca Raton, 2014).
38. Cikara, M., Martinez, J. E. & Lewis, N. A. Moving beyond social categories by incorporating context in social psychological theory. *Nat Rev Psychol* **1**, 537–549 (2022).
39. Pettigrew, T. Intergroup Contact Theory. *Annual review of psychology* **49**, 65–85 (1998).
40. Sherman, J. W., Stroessner, S. J., Conrey, F. R. & Azam, O. A. Prejudice and Stereotype Maintenance Processes: Attention, Attribution, and Individuation. *Journal of Personality and Social Psychology* **89**, 607–622 (2005).
41. Brewer, M. B. A Dual Process Model of Impression Formation. in *Advances in Social Cognition, Volume I* (Psychology Press, 1988).
42. Maurer, K. L., Park, B. & Rothbart, M. Subtyping versus subgrouping processes in stereotype representation. *Journal of personality and social psychology* **69**, 812 (1995).
43. Gershman, S. J. & Cikara, M. Structure learning principles of stereotype change. *Psychon Bull Rev* **30**, 1273–1293 (2023).
44. Bai, X., Ramos, M. R. & Fiske, S. T. As diversity increases, people paradoxically perceive social groups as more similar. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 12741–12749 (2020).
45. Smith, T. W., Marsden, P., Hout, M. & Kim, J. General social surveys. *National Opinion Research Center* https://gss.norc.org/content/dam/gss/get-documentation/pdf/reports/project-reports/GSSProject%20report32.pdf (2012).
46. Kemp, C., Perfors, A. & Tenenbaum, J. B. Learning overhypotheses with hierarchical Bayesian models. *Developmental Science* **10**, 307–321 (2007).
47. Sanborn, A. N., Griffiths, T. L. & Shiffrin, R. M. Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology* **60**, 63–106 (2010).
48. Perfors, A., Tenenbaum, J. B. & Regier, T. The learnability of abstract syntactic principles. *Cognition* **118**, 306–338 (2011).
49. Kemp, C., Goodman, N. D. & Tenenbaum, J. B. Learning to Learn Causal Models. *Cognitive Science* **34**, 1185–1243 (2010).
50. Landau, B., Smith, L. B. & Jones, S. S. The importance of shape in early lexical learning. *Cognitive development* **3**, 299–321 (1988).
51. Fiske, S. T., Cuddy, A. J. C., Glick, P. & Xu, J. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology* **82**, 878–902 (2002).
52. Aboud, F. E. *Children and Prejudice*. (Blackwell Pub, Oxford, 1989).

53. Baron, A. S., Dunham, Y., Banaji, M. & Carey, S. Constraints on the Acquisition of Social Category Concepts. *Journal of Cognition and Development* **15**, 238–268 (2014).
54. Huebner, P. A. & Willits, J. A. Structured Semantic Knowledge Can Emerge Automatically from Predicting Word Sequences in Child-Directed Speech. *Front. Psychol.* **9**, (2018).
55. Martinez, J. E., Krasner, R. H., Rosero, L., Gershman, S. J. & Cikara, M. Social Group Discovery, Structure, and Stereotype Updating. *Journal of Experimental Psychology: General* (2025).
56. Navarro, D. J., Griffiths, T. L., Steyvers, M. & Lee, M. D. Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology* **50**, 101–122 (2006).
57. Schultner, D., Lindström, B. R., Cikara, M. & Amodio, D. M. Transmission of social bias through observational learning. *Science Advances* (2024).
58. Wille, E. *et al.* Gender stereotypes in a children's television program: Effects on girls' and boys' stereotype endorsement, math performance, motivational dispositions, and attitudes. *Frontiers in Psychology* **9**, 2435 (2018).
59. Fujioka, Y. Television Portrayals and African-American Stereotypes: Examination of Television Effects when Direct Contact is Lacking. *Journalism & Mass Communication Quarterly* **76**, 52–75 (1999).
60. Richards, Z. & Hewstone, M. Subtyping and Subgrouping: Processes for the Prevention and Promotion of Stereotype Change. *Pers Soc Psychol Rev* **5**, 52–73 (2001).
61. Purdie-Vaughns, V. & Eibach, R. P. Intersectional invisibility: The distinctive advantages and disadvantages of multiple subordinate-group identities. *Sex roles* **59**, 377–391 (2008).
62. Crenshaw, K. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. in *Feminist legal theories* 23–51 (Routledge, 2013).
63. Nicolas, G., De La Fuente, M. & Fiske, S. T. Mind the overlap in multiple categorization: A review of crossed categorization, intersectionality, and multiracial perception. *Group Processes & Intergroup Relations* **20**, 621–631 (2017).
64. Petsko, C. D., Rosette, A. S. & Bodenhausen, G. V. Through the looking glass: A lens-based account of intersectional stereotyping. *Journal of Personality and Social Psychology* **123**, 763–787 (2022).
65. Wang, Y. & Lin, C. Stereotypes at the intersection of perceivers, situations, and categories: analyzing stereotypes from storytelling using natural language processing. Preprint at https://doi.org/10.31234/osf.io/dvurm_v3 (2025).
66. Rescorla, R. A. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical conditioning, Current research and theory* **2**, 64–69 (1972).

67. Schultner, D., Molleman, L. & Lindström, B. Feature-based reward learning shapes human social learning strategies. *Nature Human Behaviour* 2183–2198 (2025) doi:10.31234/osf.io/zm7ta.
68. Schultner, D. *et al.* Transmission of societal stereotypes to individual-level prejudice through instrumental learning. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2414518121 (2024).
69. Qi, W. & Vul, E. The evolution of theory of mind on welfare tradeoff ratios. *Evolution and Human Behavior* **43**, 381–393 (2022).
70. Shi, L., Griffiths, T. L., Feldman, N. H. & Sanborn, A. N. Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review* **17**, 443–464 (2010).
71. Langlois, T. A., Charlton, J. A. & Goris, R. L. T. Bayesian inference by visuomotor neurons in the prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2420815122 (2025).
72. Lu, J. & Lin, C. Network models reveal high-dimensional social inferences in naturalistic settings beyond latent construct models. *Commun Psychol* **3**, (2025).
73. Tisue, S. & Wilensky, U. Netlogo: A simple environment for modeling complexity. in *International conference on complex systems* vol. 21 16–21 (Boston, MA, 2004).
74. Schelling, T. C. Dynamic models of segregation†. *The Journal of Mathematical Sociology* **1**, 143–186 (1971).
75. Smaldino, P. E. *Modeling Social Behavior: Mathematical and Agent-Based Models of Social Dynamics and Cultural Evolution*. (Princeton University Press, Princeton Oxford, 2023).
76. Walker, K., Herman, M., Eberwein, K. & Walker, M. K. Package 'tidycensus'. *MIT* http://148.205.148.16/web/packages/tidycensus/tidycensus.pdf (2021).
77. Stier, A. J., Sajjadi, S., Karimi, F., Bettencourt, L. M. A. & Berman, M. G. Implicit racial biases are lower in more populous more diverse and less segregated US cities. *Nat Commun* **15**, 961 (2024).