AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

© 2025 American Psychological Association
ISSN: 0096-3445

https://doi.org/10.1037/xge0001830

Social Group Discovery, Structure, and Stereotype Updating

Joel E. Martinez[1], Rachel H. Krasner[1], Laura Rosero[1], Samuel J. Gershman[1, 2, 3], and Mina Cikara[1]

[1] Department of Psychology, Harvard University
[2] Center for Brain Science, Harvard University
[3] Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University

Group stereotypes are difficult to change and drive discriminatory behavior across numerous consequential contexts. Across seven experiments, we test predictions made by a domain-general structure learning model to understand how people decide what "counts" as a group and how those group representations inform our beliefs—here, stereotypes about what a group believes—about constituent members. We have two central hypotheses. First, given low levels of deviance within a collective, participants will perceive a single group among all agents; however, as deviance of one "counter-stereotypical" agent increases, that agent will be subtyped out of the group, yielding two perceived clusters. Second, as deviance increases, confidence in one's beliefs about the group and, correspondingly, a novel group member should decrease; however, once the deviating agent is subtyped out, confidence in one's beliefs about the remaining agents and novel member should increase again. We found consistent evidence for the first prediction: As one agent's deviation from the group increased from 0% to 25%, the deviant was *subgrouped*. As deviation increased to 50% and more, the deviant was *subtyped* out of the group. We only observed support for the second prediction in two of the experiments using the confidence measure. However, an exploratory analysis of these experiments revealed a new way to index group stereotype precision—quantifying perceived similarity of all the nondeviating agents to one another. Using this measure of group-based beliefs, we see support for our second hypothesis in a majority of the experiments.

Public Significance Statement

This work leverages recent insights from the cognitive science of structure learning to address major gaps in knowledge regarding how the mind solves the problem of social categorization, subgrouping versus subtyping, and cross-categorization. Extending the social structure learning model can provide a single framework to advance our understanding of how people decide what "counts" as a group. These models make specific predictions about (a) the mechanisms by which social structures influence individuals' beliefs and behaviors and (b) the temporal dynamics underlying the group discovery and updating process. Furthermore, integrating insights from these models into the study of social cognition allows for greater predictive precision and stimulates innovative strategies for stereotype change. Being able to formalize group-structure inference and stereotype updating has great impact potential; it provides greater purchase on how to dismantle stereotypes. For example, within a single model, we can specify just how "atypical" agents need to be to shift stereotypes effectively (accounting for a group's variability and the potency of explicit category labels). Too little atypicality will result in too small a shift; too much atypicality will result in subtyping and therefore no stereotype shift.

Keywords: groups, structure learning, stereotyping

Supplemental materials: https://doi.org/10.1037/xge0001830.supp

Tara Mandalaywala served as action editor.

Mina Cikara https://orcid.org/0000-0002-6612-4474

The authors report all measures, manipulations, and data exclusion criteria. De-identified data and analysis code are available on the Open Science Framework at https://osf.io/4hzr2. In conducting this research, the authors have complied with APA ethical standards in the treatment of their human sample, and these data have not been published previously, nor are they under consideration for publication elsewhere. The authors have not, at the time of this revision submission, presented these data at any meetings, nor are they publicly posted.

Portions of the introduction are adapted from the grant that funded this work (National Science Foundation Grant 2116543 awarded to Mina Cikara and Samuel J. Gershman).

Joel E. Martinez played a lead role in formal analysis and visualization, a supporting role in conceptualization, and an equal role in writing–original draft. Rachel H. Krasner played a lead role in data curation and a supporting role in project administration and writing–original draft. Laura Rosero played a supporting role in conceptualization, data curation, formal analysis, and writing–original draft. Samuel J. Gershman played a supporting role in formal analysis, funding acquisition, resources, supervision, and writing–original draft and an equal role in conceptualization and methodology. Mina Cikara played a lead role in funding acquisition, project administration, and writing–original draft, a supporting role in data curation and formal analysis, and an equal role in conceptualization and supervision.

Correspondence concerning this article should be addressed to Mina Cikara, Department of Psychology, Harvard University, 1420 William James Hall, 33 Kirkland Street, Cambridge, MA 02138, United States. Email: mcikara@fas.harvard.edu

How do humans construct social group representations? This is the question addressed by the emerging area of research on social structure learning, which leverages ideas from computational cognitive science that have been applied to nonsocial domains (see Austerweil et al., 2015). The core idea is that the brain uses statistical learning algorithms to sort individuals into *latent groups* on the basis of their behavioral patterns, such as choices (and possibly other features). Intuitively, individuals who behave similarly will tend to be grouped together (Schwyck et al., 2024); these group representations are updated as we accumulate more evidence (Du et al., 2021; Gershman & Cikara, 2023; Gershman et al., 2017; Lau et al., 2018; Weaverdyck & Parkinson, 2018).

Here, we investigate when it is that a collective of people becomes a *group*. Specifically, we ask the following: When do our representations of one group cleave into two distinct groups versus allow for two subgroups within a higher order superordinate group? How do these different structures affect our beliefs about said group(s) (Hamilton et al., 2009)? And what happens when explicit categories intersect with alternative cues to social group structure? Across seven experiments, we incorporate and test predictions made by a domain-general structure learning model (see Gershman & Cikara, 2020) to deepen our understanding of how people decide what "counts" as a group and how those group representations inform our beliefs about constituent members.

A Computational Framework for Social Structure Learning

How do people discover social groups in the absence of explicit group labels? Decades of work on entitativity (Campbell, 1958) highlight features like proximity, similarity, and common fate as markers of what makes a group, but where do judgments of, for example, similarity come from? Which dimensions of similarity matter and how much? Is common fate an input to or consequence of grouping? Social structure learning addresses the following problem: Given observed *behavioral* patterns for a set of individuals (e.g., their choices between movies), the observer must infer to which group each person belongs—or said a different way, the latent group assignment for each individual. The normative solution to this inference problem is given by Bayes' rule, which stipulates that the *posterior probability* over groupings given choices—P(grouping| choices)—is proportional to the product of the *likelihood*—P(choices| grouping)—and the *prior probability* P(grouping) (see Gershman et al., 2017, for more details). For instance, if two agents A and B always agree with each other, and both always disagree with a third agent C, an observer's posterior probability for two groups—specifically A and B in one group and C in a second group—should be higher than other possible groupings (e.g., where A and C go together without B). Thus, the posterior represents the observer's subjective confidence in each hypothetical grouping, the likelihood represents the match between a hypothetical grouping and the choices, and the prior represents a preference for particular groupings before observing the data.

To define the likelihood, we need to specify how a grouping gives rise to choices. A basic assumption of this framework is that individuals assigned to the same group will tend to behave similarly (e.g., make similar choices). Thus, groupings with greater within-group homogeneity will have a higher likelihood. This can, however, produce many small but homogenous groups, a tendency that can be tempered by enforcing a preference for a small number of groups as the prior via a concentration parameter (see Gershman & Blei, 2012, for an introduction). Specifically, this prior has the property that it favors a small number of latent groups, but allows for a possibly unbounded number of groups, so that new groups can be added as new individuals are observed. The degree to which a small number of groups is preferred by the prior's distribution is controlled by the concentration parameter.

This model is essentially an adaptation to social domains of Bayesian structure learning models developed for nonsocial domains, notably categorization and classical conditioning (e.g., Anderson, 1991; Gershman & Niv, 2010; Sanborn et al., 2010; see also the Supervised and Unsupervised STratified Adaptive Incremental Network model, Love et al., 2004). One advantage of the Bayesian framework is that it makes explicit an individual's assumptions about the environment, which can sometimes be used to ecologically constrain the prior. A second advantage of the Bayesian framework is that it formalizes subjective uncertainty about groups, which provides a principled way of modeling confidence judgments, adaptive learning rates, and decisions under uncertainty.

Although we have focused thus far on structure *learning*, the model can be applied to structure *inference* for well-learned groups (e.g., those based on race, age, and gender). These groups will tend to be frequently encountered and hence have high prior probability under the stipulated concentration parameter, which, again, prioritizes discovering fewer groups. This potentially explains why we rely on these groups even when more fine-grained groupings might be warranted by the data—as is the case when researchers assume features like race/ethnicity ought to be the principal drivers of social preferences and/or behavior. This is known as the multiple category or "cross-categorization problem."

Extending the Model to Account for More Complex Structures: Subtyping Versus Subgrouping

Though stereotypes can change, they are demonstrably resistant to short-term updating. Allport (1954) posited that stereotypes were difficult to change because

> there is a common mental device that permits people to hold prejudgments even in the face of much contradictory evidence. It is the device of admitting exceptions. … By excluding a few favored cases, the negative rubric is kept intact for all other cases. (p. 23)

This exclusion procedure, or *subtyping*, refers to the phenomenon in which stereotype-inconsistent individuals are cognitively represented as outside of the group, thereby inoculating the group from stereotype updating in response to the inconsistent information (Maurer et al., 1995; Taylor, 1981). Another alternative to updating and subtyping is *subgrouping*—the reclassification of stereotype-inconsistent individuals into a subordinate group that nevertheless remains a part of the superordinate group (Maurer et al., 1995; Park et al., 1992).

What factors determine whether a counter-stereotypical target updates the stereotype, gets subtyped, or gets subgrouped? If counter-stereotypical evidence is restricted to only a few targets, or counter-stereotypical targets are atypical along many additional dimensions, group-level stereotypes remain intact, suggesting that the "deviants" have been subtyped out (Hewstone, 1994; Johnston & Hewstone, 1992; Kunda & Oleson, 1995). By contrast, stereotypes about a group change more when stereotype-inconsistent attributes are dispersed across many members as opposed to concentrated, indicating the possibility of either updating or subgrouping (Weber & Crocker, 1983). This dispersion effect is mediated by typicality: Typical individuals affect group impressions more than atypical individuals (Rothbart & Lewis, 1988, Experiment 3; Weber & Crocker, 1983, Experiment 3). However, this dispersion effect only occurs when perceived group variability is low (Hewstone & Hamberger, 2000). In other words, we observe more stereotype change for dispersed stereotype-inconsistent attributes when the group is perceived as more homogenous. This indicates that the baseline variability of the group also matters for the effect of inconsistent members. Interestingly, moderate "deviants" cause more stereotype updating than extreme deviants (Kunda & Oleson, 1997), suggesting there is a threshold for updating/subgrouping; people who are too extreme in their inconsistency simply get subtyped out of the group.

In the decades since this work began, dozens of papers have documented the conditions under which updating, subgrouping, and subtyping occur; however, there is still *no unified theory* to account for all these findings. One strength of the social structure approach described above is the ease with which it can be adapted to make principled, quantitative predictions specifying the conditions under which people update, subtype, or subgroup. That is, we can extend the model to account for hierarchical structure by using a nonparametric distribution over tree structures (Blei et al., 2010). This is a generalization of the "flat" prior used in our earlier work (Gershman & Cikara, 2023).

Taking into account the heterogeneity of the agents in a group, the model makes predictions about just how stereotype-inconsistent or "deviant" a new agent has to be in order to (a) get represented as a member of the superordinate category, (b) be subtyped out, or (c) be represented as a member of a subgroup. In the case of subtyping, we would say that people would estimate the highest posterior probability on a two-group solution: one with all the current group members and a separate group made up of the new, deviant agent. In the case of subgrouping, we would say that people would estimate the highest posterior probability on a single-group solution, but with one subgroup that includes the new agent.

What predictions about group-related *beliefs* fall out of these different structure solutions? When an individual is subtyped out, they cannot affect the superordinate category stereotypes, because one can think of them as being assigned to an "exception" group on the same level as the superordinate group. By contrast, individuals who come to form subgroups will have a small updating effect. Finally, individuals lumped in with the superordinate category will have the strongest effect of updating beliefs about the group.

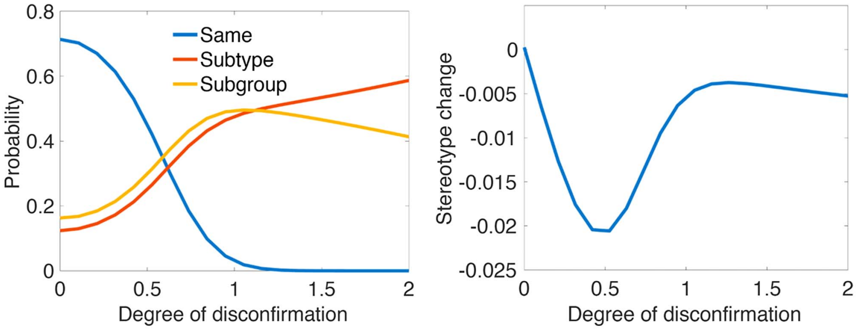
Initial Predictions Derived From the Model

Our analysis treats hierarchical social grouping as a problem of unsupervised structure discovery. The structure in this case takes the form of a tree, where each node in the tree corresponds to a group, whose parent node represents the superordinate group. The model simultaneously infers the tree structure and the assignments of individuals to nodes in the tree. Once all the assignments have been made, the model reports the group-level feature expectations, which allow us to measure how much prior beliefs over groupings have been updated.

Figure 1 shows the results of an illustrative simulation in which individuals are described by a single feature. The model is first given a set of homogenous individuals (n = 10) who all share feature = 1. We will call this feature setting the group stereotype. We then introduce a new agent who varies in the degree to which they disconfirm the stereotype (i.e., feature = x, where x gives the distance between the new agent's feature value and the feature value of the homogenous group). What we find is that for small disconfirmation values (e.g., degree = .2), the model favors assigning the new individual to the existing group. As the degree of disconfirmation or deviance grows larger (e.g., degree = .7), there is transition to the formation of a subgroup nested within the original group and then further to the formation of a novel subtype (Figure 1, left).

Figure 1

Illustrative Simulation of Our Model



Note. (Left) Posterior probability for each structure as a function of stereotype disconfirmation. As the new agent's distance from the group stereotype increases, different group assignments become more likely. (Right) Degree of stereotype change (i.e., group feature change) as a function of stereotype disconfirmation, where more negative values indicate more change. The stereotype changes with increasing disconfirmation up until the point that the subtype hypothesis becomes most likely. The disconfirmation values shown are arbitrary representations of magnitude, not directly mapped to the experimental design. See the online article for the color version of this figure.

How does this affect stereotype change toward the homogenous group? The model predicts a nonmonotonic relationship between the degree of disconfirmation and stereotype change (Figure 1, right): For small levels of disconfirmation from the new agent, the original stereotype is updated but not very much, whereas for very large levels of disconfirmation, the original stereotype is not updated at all. Only for intermediate levels of disconfirmation does the group stereotype get partially updated. Why? Because in this intermediate area, both the group and the subgroup take partial "responsibility" for the new agent's feature value.

The Current Experiments

While a vast literature has demonstrated the minimal preconditions for establishing prejudiced attitudes and discrimination (e.g., in the form of preferential resource allocation), we know far less about the minimal preconditions requisite for the formation and application of stereotypes (see, however, Bai, Fiske, & Griffiths, 2022; Bai, Griffiths, & Fiske, 2022).

Across a series of experiments, we test whether participants behave as predicted by the hierarchical structure learning model in their judgments of social targets. In contrast to previous studies that rely on preexisting categories, we start with "category-free" collectives and parametrically vary the deviance of single member to assess how it changes the inferred structure of the collective and associated beliefs (i.e., proto-stereotypes; Experiments 1a–1d). We then move on to test whether the model's predictions are supported when it is applied to known categories (i.e., collectives identified as political parties; Experiment 2). By parametrically varying the deviance of the counter-stereotypical exemplar and documenting the conditions under which human responses converge with versus diverge from model predictions, these experiments reveal several new insights: the extent of deviance required for participants to begin to update their latent structures and which features of the *social* context and stereotyping are unique relative to domain-general structure learning contexts.

Note also that people generally identify more subgroups for their in-group than for out-groups (Brewer & Lui, 1984; Huddy & Virtanen, 1995; Judd et al., 1995; Park & Judd, 1990; Park & Rothbart, 1982; Wallace et al., 1995). Why? Recall that subgrouping is more likely when perceivers observe greater group variability (Maurer et al., 1995; Park et al., 1992). People typically have more experience with in-group relative to out-group exemplars, which means they likely have a broader range of what constitutes "typical" for the in-group (i.e., the converse of the out-group homogeneity effect). The restricted variability in representations of out-groups, by contrast, makes subtyping more likely. Therefore, in the final two experiments, we test how results change when participants, themselves, are included within the collectives (Experiments 3 and 4).

We have two core hypotheses throughout, based on Figure 1. First, at low levels of individual deviance, participants will perceive a single group including all agents; however, as deviance of the one "counter-stereotypical" agent increases, that agent will be subtyped out of the group, yielding two clusters. Second, at low levels of individual deviance, confidence in one's beliefs about the group and, correspondingly, a new group member (about whom participants know nothing) should decrease; however, as the deviant is subtyped out into their own cluster, confidence in one's beliefs about the remaining agents in the group, including the new member, should increase again.

Overview of Analyses for All Experiments

Experiments 1a through 4 all employ a similar paradigm (see Figure 2). In this paradigm, participants guessed about and received feedback regarding the political issue positions of a series of agents. We manipulate between participants the extent to which deviant agents' positions differ from the positions of the other agents. In the next phase, participants rate the extent to which each pair of agents has similar views. In the final phase, we present a novel agent they have yet to see, ask participants to guess that agent's position on the last issue, and report their confidence about their guess.

Learning About Each Nondeviant Agent

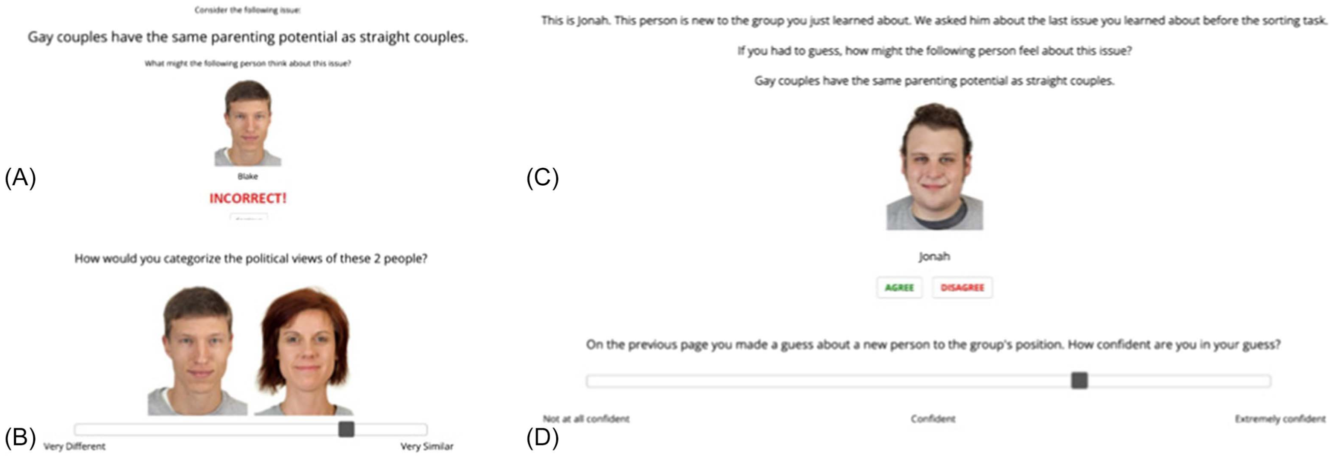
To first assess whether participants accurately learned the nondeviant agents' opinions, participant's predictions of the agents' opinions were coded as correct or incorrect and submitted to logistic mixed models. Participant accuracy was predicted from an interaction between deviance condition and the opinion learning round (from 1 to the max number of opinions in the design). To account for dependencies in the data due to repeated measures (Barr et al., 2013), the model estimated random intercepts per stimulus and participant alongside a random slope of opinion round by participant. Results are reported in log odds.

Learning About the Deviant Agent and Social Structure

We report two analyses that assess whether participants extracted social structure among the agents and if the deviants altered that structure. The first analysis uses the pairwise similarity ratings and applies to them an implementation of the hierarchical Bayesian nonparametric model detailed in the introduction, the infinite similarity model (ISM). The ISM is conceptually related to the infinite relational model (Kemp et al., 2006); see the mathematical details in Supplemental Note 1. This model can estimate the number of clusters each participant extracted among the agents. The model was implemented in Julia using the Turing package (Ge et al., 2018) and was initialized with the following hyperparameter values: $\nu = 2$ (controls the exponential distribution over the Chinese restaurant process concentration parameter—α—which influences the number of clusters created; a higher α typically leads to more clusters), $a = 11$ and $b = 0.5$ (both control the inverse γ distribution over observation variance, which is the variability within each cluster or the "width" of the clusters), and $\lambda = 0.1$ (controls sparsity of the β distribution over the mean similarity between clusters; a sparsity-inducing prior can lead to more distinct clusters). To obtain posterior estimates, the model was sampled 5,000 times using a sequential Monte Carlo sampler. One of the output metrics of the model is k, the number of clusters identified in the similarity ratings. If the degree of deviance of the deviant agent affects perceived structure, the number of clusters should increase to at least $k = 2$ as deviation increases (i.e., the deviant is represented as separate from the rest of the agents). The k estimates were predicted by the deviance condition in an ordinary least squares linear model. We tested condition marginal means when they were less than a null of $k = 2$ to identify at what deviance condition the agents were partitioned into two clusters.

Figure 2

Sample Stimuli and Experiment Overview



(A) Consider the following issue:

Gay couples have the same parenting potential as straight couples.

What might the following person think about this issue?

(B) How would you categorize the political views of these 2 people?

(C) This is Jonah. This person is new to the group you just learned about. We asked him about the last issue you learned about before the sorting task.

If you had to guess, how might the following person feel about this issue?

Gay couples have the same parenting potential as straight couples.

(D) On the previous page you made a guess about a new person to the group's position. How confident are you in your guess?

Study	1a	1b	1c	1d	2	3	4
n	284	273	278	560	277	309	287
design	8 issues 8 agents No reference to group	12 issues 6 agents No reference to group	12 issues 6 agents "group impression task"	12 issues 6 agents "group impression task" Only 25/50/75% deviance	12 issues 6 agents Group labeled political in-party	8 issues 8 agents Self-included Group labeled political in-party Deviance across different agents	8 issues 8 agents Self-included Group labeled political in-party

Note. (A) Learning and feedback trial. (B) Similarity judgment. (C) Guess for new agent. (D) Confidence rating on new agent judgment. Facial images are from the Chicago Face Database (Ma et al., 2015). Used with permission. See the online article for the color version of this figure.

The second analysis explicitly investigates the deviant's role in facilitating perceived structure by comparing the similarity ratings given to two types of agent pairs: deviants to nondeviants ("DN") versus nondeviants to nondeviants ("NN"). If the deviant was identified and treated as separate from other agents, more deviation should decrease the DN similarities compared to NN similarities. Likewise, in the last two experiments, participants' own similarity to the agents should decrease when the agents deviate more from the participants' answers. We tested whether there was an interaction between deviance condition and agent pair type (DN vs. NN) on similarity ratings using linear mixed models. The models included random intercepts and random slopes of pair type (DN vs. NN) per participant.

Confidence in New Agent's Opinion

The hierarchical structure learning model predicts that as the amount of group deviation increases, there is a shift in inference about group structure and a nonmonotonic change in stereotypes. We hypothesize this nonmonotonicity will be reflected in participants' confidence in how *new* group members will behave (i.e., if the stereotype has not changed, participants should be more confident a new person will share the features of the observed group). Specifically, the structure learning model predicts that participants' confidence about the new agent's opinion should decrease from the 0% to 50% deviation conditions and then increase from the 50% to 100% deviation conditions. We therefore analyzed the confidence ratings using a two-line test, which checks for significant slopes and a sign change before and after a break point (Simonsohn, 2018). We used two separate ordinary least squares regressions to predict confidence ratings from deviance conditions. The first model included 0%–50% deviation conditions, and the second model included 50%–100% conditions. To corroborate model predictions, the first regression's slope should be negative, the second slope should be positive, and both should be significant.

We also examined an exploratory moderator that could change the pattern of confidence ratings: whether the participants' predictions of the new agent's opinion matched the majority opinion (binary variable: true vs. false). Predicting a disagreement with the majority might indicate that participants had not encoded or had forgotten the majority opinion, in which case we would only observe the predicted confidence rating pattern in those who guessed the new agent would share the majority opinion. Within each level of the moderator, we computed the two-line test implemented as we have described above.

All above model estimates (slopes, marginal means) and comparisons were computed using the "emmeans" package in R (Lenth, 2020); degrees of freedom for linear mixed models were estimated using the Satterthwaite method (Luke, 2017). To ensure convergence in the mixed models, they were optimized using "bobyqa" and allowed to run for 500,000 iterations.

Transparency and Openness

We report all measures, manipulations, and data exclusion criteria. De-identified data and analysis code are available at https://osf.io/4hzr2 (Martinez et al., 2024).

Experiment 1a

Method

Participants

In Experiment 1, we recruited 301 participants from the online research platform Prolific. Participants were paid $3.00 as compensation for their time and provided informed consent in accordance with the university institutional review board (IRB).

> In addition to age, we asked participants to report their gender (i.e., to select one of the following: man, woman, non-binary, another gender not listed here, prefer not to answer) and their race based on the U.S. Census categories (American Indian or Alaska Native, Asian, Black or African-American, Hispanic/Latinx, White, Native Hawaiian or Other Pacific Islander, Other).

We excluded 17 participants who failed either or both attention checks (see below), leaving a total $n = 284$ (man = 104, woman = 170, nonbinary = 6, prefer not to answer = 4; average age = 35.9).

Stimuli

For agent pictures, we included 48 photos from the Chicago Face Database (Ma et al., 2015)—24 female and 24 male from the pool of faces classified as "white" (based on Chicago Face Database designations). Of these faces, four randomly selected females and four randomly selected males became the group of eight agents featured in the study.

Procedure

We told participants that they were about to learn about a collection of people polled on a series of political issues, with the following instructions:

> On the screens that follow you're going to learn about a collection of people we polled on a series of political issues. You are going to make guesses about and receive feedback on their positions on a series of 8 political issues. You don't have to remember what each person's position is, but try to see if you can figure out to what extent each person agrees with everyone else.

We presented eight agents for each issue, across eight political issues total. Unbeknownst to participants, for each given issue, seven of the eight agents were randomly assigned to either agree or disagree with a 5% chance that an agent would randomly deviate from that choice. This had the effect of creating a majority position from which each agent would rarely deviate. We randomly assigned participants to one of five deviant threshold conditions (0%, 25%, 50%, 75%, 100% deviance) where the eighth agent represented the deviant in the group. Deviant and nondeviant face identities were randomly assigned between participants (i.e., the face of the deviant varied across participants, but the identity of the deviant remained consistent within participant). This was the case for all experiments reported in the article with the exception of Experiment 3, where deviant identities varied between and within participants across trials (see below). The percentage threshold denoted the deviant's disagreement profile in relation to the group (i.e., at 25%, the deviant disagreed with the majority position on 25%, or 2, of the eight political issues).

In the first phase of the study, for each opinion learning round, each screen showed participants a political position at the top with an image of a male or a female agent with a corresponding name below. Participants clicked either the "agree" or "disagree" button to indicate their guess about that agent's position. "Correct!" or "Incorrect!" appeared after they selected their response. We randomized the order of political issues across participants and the order of agents within each issue (see Figure 2 for example stimuli).

In the second phase of the study, participants made a series of similarity judgments. Specifically, participants judged the similarity of each pair of agents from Phase 1 on a sliding scale from 1 = *very different* to 100 = *very similar*. We randomized the presentation order of the pairs across participants.

In the third phase, we presented participants with a new agent. Participants first guessed how this new agent felt ("agree" or "disagree") about the last political issue they saw prior to the similarity judgment task. They then answered the following question: "On the previous page you made a guess about a new person to the group's position. How confident are you in your guess?" on a scale ranging from 0 = *not at all confident* to 100 = *extremely confident*.

After completing the third phase of the study, participants then answered two attention check questions. The first question asked, "How many fatal heart attacks have you had?" with possible choices ranging from 0 to 10. Any answer other than 0 failed this attention check. The second question appeared with the following prompt: "This is an attention check. Consider the following numbers: [*number set*]. If these numbers were sorted based on their numeric value, which would be the value in the middle? Please write out the number in lower case." One of three possible number sets was randomly presented: [nine, eight, 4], [2, six, three], or [one, 7, five], with the correct answers being eight, three, and five, respectively. Any other answer for the corresponding number set failed this attention check.

The penultimate page of the study asked respondents to supply information as to which race/ethnicity they most identify, gender, as well as their age. Finally, we debriefed participants.

Analysis

See the overview above.

Results

Learning About Nondeviant Agents

First, we sought to confirm that participants learned about the agents. Both opinion round, $\chi^2(1) = 155.57$, $p < .0001$, and deviance, $\chi^2(4) = 29.91$, $p < .0001$, were significant predictors of accuracy about nondeviant agents' opinions, but not their interaction (p = .766). Participants' accuracy increased with each opinion round (b = .19, SE = .015, z = 12.28, p < .0001), and the 0% deviance condition showed greater accuracy than the other conditions ($M_{\text{differences}} = .44$ to

.58, *SE*s = .14 to .16, *z*s = 2.83 to 3.95, *p*s = .0008 to .037). The average accuracy on the last opinion round ranged between 75% and 80% across deviance conditions. These results suggest participants successfully learned the nondeviating agents' opinions, especially when there was no deviant.

Deviant and Structure Learning

We predicted that at low levels of individual deviance, participants would perceive a single group including all agents; however, as deviance of the one "counter-stereotypical" agent increased, that agent would be subtyped out of the group, yielding two clusters. Indeed, the number of clusters identified by the ISM, k, changed across deviance conditions, $F(4, 279) = 16.33$, $p < .0001$. The average k remained under 2 for the 0% ($k = 1.61$, CI [1.44, 1.78], $p < .0001$) and 25% conditions ($k = 1.66$, CI [1.51, 1.80], $p < .0001$), reached 2 at 50% deviation ($k = 1.87$, CI [1.72, 2.02], $p = .049$), and became 2 or greater at 75% ($k = 2.12$, CI [1.95, 2.29], $p = .911$) and 100% ($k = 2.40$, CI [2.24, 2.57], $p = 1.00$) deviation. With more deviation, participants perceived an increased number of clusters among the agents (Figure 3).

This perceived structure occurred in part because participants represented the deviant as belonging to a separate cluster as deviance increased. The interaction between deviance condition and agent pair type (NN vs. DN) was significant in predicting the similarity ratings, $F(1, 284) = 225.39$, $p < .0001$. The similarity between the deviant and the rest of the agents was high in the 0% deviation condition ($M = 84.4\%$, CI [80.1, 88.7]) and dropped with increasing deviation ($b = -62.3$, $SE = 3.45$, CI [−69.1, −18.1]), $t(284) = -18.06$, $p < .0001$, while the similarity between the nondeviating agents started high ($M = 83\%$, CI [78.7, 87.3]) and remained consistent as deviation increased: $b = .01$, $SE = 2.87$, CI [−5.63, 5.66], $t(284) = .01$, $p = .996$; $b_{\text{difference}} = -62.3$, $SE = 4.15$, CI [−70.4, −54.1], $t(284) = -15.01$, $p < .0001$. Collectively, these results suggest that participants learned both the agents' and deviants' positions well, which shaped the group structures extracted among them. Deviant subgrouping occurred at 25% (lower DN than NN similarity yet still less than two perceived clusters), and deviant subtyping began developing around 50% deviation (lower DN than NN similarity and two or more perceived clusters).

Confidence Ratings

Our confidence hypothesis was that at low levels of individual deviance, confidence in one's beliefs about the group and, correspondingly, a new group member (about whom participants know nothing) should decrease; however, as the deviant is subtyped out into their own cluster, confidence in one's beliefs about the remaining agents in the group, including the new member, should increase again. The nonmonotonic shape in confidence ratings prediction about the new agent's opinion was not supported. The two-line test showed a consistent linear decrease in confidence as deviance increased. The slope between 0% and 50% deviation was negative ($b = -14.2$, $SE = 9.8$, CI [−33.2, 5.2]), $t(182) = -1.45$, $p = .149$, as was the slope between 50% and 100% deviation ($b = -14.3$, $SE = 10.6$, CI [−35.2, 6.6]), $t(158) = -1.35$, $p = .178$. Neither was significant. The same pattern held in the two-line majority opinion moderator analyses, and all the slopes were negative and nonsignificant (bs < −8.3, ps > .242).

Experiment 1b

In Experiment 1a, participants clearly demonstrated a capacity to learn about the agents, including the deviant. Critically, our results supported our cluster hypothesis but not our confidence rating hypothesis. One possible explanation for why we did not observe our predicted effects on confidence ratings is because participants may not have had enough information about each agent to generate sufficient confidence in the constitutions of the majority group. In Experiment 1b we reduced the number of agents and increased the number of issues.

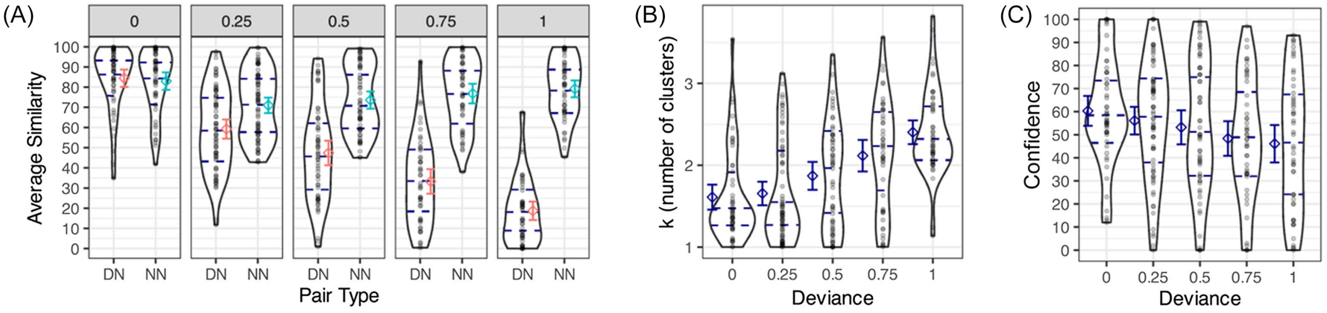
Method

Participants

In Experiment 1b, we recruited 299 participants from the online research platform Prolific. Participants were paid $3.00 as compensation for their time and provided informed consent in accordance with the university IRB. We excluded 26 participants who failed either or both attention checks (see below), leaving a total $n = 273$ (man = 130, woman = 134, nonbinary = 5, prefer not to answer = 3, another gender not listed here = 1; average age = 36.9).

Figure 3

Results From Experiment 1a



Note. (A) Average similarity ratings by deviance condition and pair type: deviant–nondeviant (DN; red) and nondeviant–nondeviant (NN; blue). (B) Estimates of k (number of clusters) across deviance conditions. (C) Confidence ratings for the new agent's opinion across deviance conditions. Error bars around the mean represent 95% confidence intervals. The violin shape displays the distribution of data, and the dashed lines represent quartiles. See the online article for the color version of this figure.

Procedure

Experiment 1b replicated the procedure in Experiment 1a with the following changes: We presented six agents for each issue, across 12 political issues total, with instructions in the first task updated to reflect the number of political issues. Otherwise, Experiment 1b presented the same tasks, attention checks, and demographic questions as Experiment 1a.

Analysis

The analysis followed the same structure as Experiment 1a.

Results

Learning About Nondeviant Agents

Again, we confirmed that participants learned about the agents. Both opinion round, $\chi^2(1) = 195.45$, $p < .0001$, and deviance, $\chi^2(4) = 33.24$, $p < .0001$, were significant predictors of accuracy about nondeviant agents' opinions, but not their interaction ($p = .133$). Participants' accuracy increased with each opinion round ($b = .113$, $SE = .008$, $z = 14.05$, $p < .0001$), and the 0% deviance condition showed greater accuracy than the other conditions ($M_{\text{differences}} = .53$ to $.68$, SEs $= .142$ to $.147$, zs $= 3.61$ to 4.8, ps $< .003$). The average accuracy on the last opinion round ranged between 73.9% and 83.9% across deviance conditions. These results suggest participants successfully learned the nondeviating agents' opinions, especially when there was no deviant.

Deviant and Structure Learning

We predicted that at low levels of individual deviance, participants would perceive a single group including all agents; however, as deviance of the one "counter-stereotypical" agent increased, that agent would be subtyped out of the group, yielding two clusters. Indeed, the number of clusters identified by the ISM, k, changed across deviance conditions, $F(4, 268) = 14.39$, $p < .0001$. The average k remained under 2 for the 0% ($k = 1.68$, CI [1.50, 1.85], $p = .0002$) and 25% conditions ($k = 1.65$, CI [1.45, 1.84], $p = .0002$), reached 2 at 50% deviation ($k = 2.11$, CI [1.92, 2.29], $p = .868$), and became 2 or greater at 75% ($k = 2.27$, CI [2.07, 2.46], $p = .996$) and 100% ($k = 2.40$, CI [2.24, 2.57], $p = 1.00$) deviation. With more deviation, participants perceived an increased number of clusters among the agents (Figure 4).

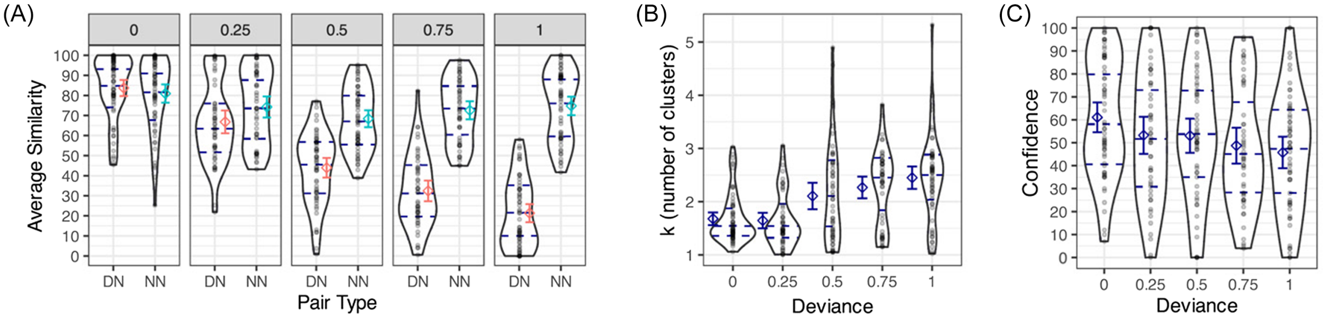
This perceived structure occurred in part because participants represented the deviant as belonging to a separate cluster as deviance increased. The interaction between deviance condition and agent pair type (NN vs. DN) was significant in predicting the similarity ratings, $F(1, 273) = 202.64$, $p < .0001$. The similarity between the deviant and the rest of the agents was high in the 0% deviation condition ($M = 83.7\%$, CI [79.7, 87.7]) and dropped with increasing deviation ($b = -63.7$, $SE = 2.97$, CI [−69.5, −57.8]), $t(273) = -21.42$, $p < .0001$, while the similarity between the nondeviating agents started high ($M = 80.9\%$, CI [76.4, 85.5]) and decreased slightly as deviation increased: $b = -5.91$, $SE = 2.89$, CI [−11.6, −.023], $t(273) = -2.05$, $p = .041$; $b_{\text{difference}} = -57.8$, $SE = 4.1$, CI [−65.8, −49.8], $t(273) = -14.2$, $p < .0001$. Collectively, these results suggest that participants learned both the agents' and deviants' positions well, which shaped the group structures extracted among them. Deviant subgrouping occurred at 25% (lower DN than NN similarity yet still less than two perceived clusters), and deviant subtyping began developing around 50% deviation (lower DN than NN similarity and two or more perceived clusters).

Confidence Ratings

Our confidence hypothesis was that at low levels of individual deviance, confidence in one's beliefs about the group and, correspondingly, a new group member (about whom participants know nothing) should decrease; however, as the deviant is subtyped out into their own cluster, confidence in one's beliefs about the remaining agents in the group, including the new member, should increase again. The nonmonotonic shape in confidence ratings prediction about the new agent's opinion was not supported. The two-line test showed a consistent linear decrease in confidence as deviance increased. The slope between 0% and 50% deviation was negative ($b = -16.3$, $SE = 10$, CI [−36.1, 3.49]), $t(165) = -1.63$, $p = .106$, as was the slope between 50% and 100% deviation ($b = -14.6$, $SE = 10.1$, CI [−34.6, 5.39]), $t(159) = -1.44$, $p = .151$. Neither was significant. The same pattern held in the two-line majority opinion moderator analyses,

Figure 4

Results From Experiment 1b



Note. (A) Average similarity ratings by deviance condition and pair type: deviant–nondeviant (DN; red) and nondeviant–nondeviant (NN; blue). (B) Estimates of k (number of clusters) across deviance conditions. (C) Confidence ratings for the new agent's opinion across deviance conditions. Error bars around the mean represent 95% confidence intervals. The violin shape displays the distribution of data, and the dashed lines represent quartiles. See the online article for the color version of this figure.

and all the slopes were negative and nonsignificant (*b*s < −3.97, *p*s > 1.02).

Experiment 1c

The results of Experiment 1b replicated the results of Experiment 1a: Participants learned about the agents, and we observed support for the cluster hypothesis but not the confidence hypothesis. Yet another reason we may not have observed the model-predicted effects on the confidence ratings was that the participants were not encoding the collective as a group, but rather just as a confederation of individuals. In Experiment 1c, we made it explicit that this was a "group impression" task. We also included an exploratory individual differences measure—personal need for structure (PNS)—which had previously been identified as a moderator of subtyping (Neuberg & Newsom, 1993).

Method

Participants

In Experiment 1c, we recruited 300 participants from the online research platform Prolific. Participants were paid $3.00 as compensation for their time and provided informed consent in accordance with the university IRB. We excluded 22 participants who failed either or both attention checks, leaving a total $n = 278$ (man = 134, woman = 136, nonbinary = 6, prefer not to answer = 2; overall average age = 36.4).

Procedure

Experiment 1c replicated the procedure in Experiment 1a with the following changes: We presented six agents for each issue, across 12 political issues total, as in Experiment 1b. In addition, we amended the first phase instructions to emphasize the group aspect:

> On the screens that follow you're going to learn about a collection of people we polled on a series of political issues and complete a group impression task. You are going to make guesses about and receive feedback on their positions on a series of 12 political issues. You don't have to remember what each person's position is, but try to see if you can figure out to what extent each person agrees with everyone else in the group.

After participants completed the third phase, they answered questions taken from the PNS scale (Neuberg & Newsom, 1993). A high score on the PNS scale suggested a propensity for organizing social and nonsocial information in a simple explanatory model. This tendency toward simplistic modeling had potential implications for the creation and adaption of stereotypes in ambiguous environments. Participants read instructions used in the introduction of the PNS scale:

> Read each of the following statements and decide how much you agree with each according to your attitudes, beliefs, and experiences. It is important for you to realize that there are no "right" or "wrong" answers to these questions. People are different, and we are interested in how you feel. Please respond according to the following 6-point scale (1 = *strongly disagree* to 6 = *strongly agree*).

We selected the four questions from the PNS scale's stereotype-related questions: "I enjoy having a clear and structured mode of life," "I like to have a place for everything and everything in its place," "I find that a well-ordered life with regular hours makes my life tedious" (reverse-coded), and "I find that a consistent routine enables me to enjoy life more."

Otherwise, Experiment 1c presented the same tasks, attention checks, and demographic questions as in the preceding experiments.

Analysis

The analysis procedures were the same as described in Experiment 1a with one exception: We also examined the moderating role of participants' PNS (binary variable: high vs. low, as produced by median split) on confidence ratings. Having a high need for structure could increase vigilance for deviations compared to a lower need, which in turn could produce the expected nonmonotonicity in confidence ratings, specifically among those with high need for structure. Within each level of the moderator, we computed the two-line test.

Results

Learning About Nondeviant Agents

Again, we confirmed that participants learned about the agents. Both opinion round, $\chi^2(1) = 220.97$, $p < .0001$, and deviance, $\chi^2(4) = 51.37$, $p < .0001$, were significant predictors of accuracy about nondeviant agents' opinions, but not their interaction ($p = .319$). Participants' accuracy increased with each opinion round ($b = .119$, $SE = .008$, CI [.103, .132], $z = 14.87$, $p < .0001$), and the 0% deviance condition showed greater accuracy than the other conditions ($M_{\text{differences}} = .59$ to .84, *SE*s = .131 to .142, *p*s < .0003). The average accuracy on the last opinion round ranged between 71.3% and 87.4% across deviance conditions. These results suggest participants successfully learned the nondeviating agents' opinions, especially when there was no deviant.

Deviant and Structure Learning

We predicted that at low levels of individual deviance, participants would perceive a single group including all agents; however, as deviance of the one "counter-stereotypical" agent increased, that agent would be subtyped out of the group, yielding two clusters. Indeed, the number of clusters identified by the ISM, *k*, changed across deviance conditions, $F(4, 273) = 15.95$, $p < .0001$. The average *k* remained under 2 for the 0% ($k = 1.72$, CI [1.52, 1.92], $p = .003$) and 25% conditions ($k = 1.75$, CI [1.54, 1.96], $p = .011$), reached 2 at 50% deviation ($k = 1.88$, CI [1.69, 2.07], $p = .107$), and became 2 or greater at 75% ($k = 2.12$, CI [2.12, 2.54], $p = .999$) and 100% ($k = 2.59$, CI [2.41, 2.77], $p = 1.00$) deviation. With more deviation, participants perceived an increased number of clusters among the agents.

This perceived structure occurred in part because participants represented the deviant as belonging to a separate cluster as deviance increased. The interaction between deviance condition and agent pair type (NN vs. DN) was significant in predicting the similarity ratings, $F(1, 278) = 129.96$, $p < .0001$. The similarity between the deviant and the rest of the agents was high in the 0% deviation condition ($M = 87.2\%$, CI [83.5, 90.8]) and dropped with increasing deviation ($b = -60.1$, $SE = 3.22$, CI [−66.4, −53.7]), $t(278) = -18.65$, $p < .0001$, while the similarity between the nondeviating agents started high (M = 85.2%, CI [81.5, 88.9]) and decreased slightly as deviation increased: $b = -11.7$, $SE = 2.78$, CI [−6.2, 4.2], $t(278) = -4.21$, $p < .0001$; $b_{\text{difference}} = -48.4$, $SE = 4.24$, CI [−56.7, −40], $t(278) = -11.4$, $p < .0001$. Collectively, these results suggest that participants learned both the agents' and deviants' positions well, which shaped the group structures extracted among them. Deviant subgrouping occurred at 25% (lower DN than NN similarity yet still less than two perceived clusters), and deviant subtyping began developing around 50% deviation (lower DN than NN similarity and two or more perceived clusters).

Confidence Ratings

Our confidence hypothesis was that at low levels of individual deviance, confidence in one's beliefs about the group and, correspondingly, a new group member (about whom participants know nothing) should decrease; however, as the deviant is subtyped out into their own cluster, confidence in one's beliefs about the remaining agents in the group, including the new member, should increase again. The nonmonotonic shape in confidence ratings prediction about the new agent's opinion was not supported. The two-line test showed a consistent linear decrease in confidence as deviance increased. The slope between 0% and 50% deviation was negative ($b = -26.2$, $SE = 10.5$, CI [−47, −5.5]), $t(159) = -2.50$, $p = .014$, as was the slope between 50% and 100% deviation ($b = -3.41$, SE = 10.3, CI [−23, 17]), $t(174) = -.33$, $p = .742$. The two-line majority opinion moderator analyses did not support the predicted patterns, and all the slopes were negative and nonsignificant when participants chose the majority opinion (bs < −16.5, ps > .057). The PNS two-line test showed consistent negative and nonsignificant slopes for the low PNS group (bs < −6.7, ps > .268). The high PNS group showed the predicted patterns before ($b = -42.5$, $SE = 14.9$, CI [−72.2, −12.8]), $t(67) = -2.86$, $p = .006$, and after ($b = 9.33$, $SE = 16.2$, CI [−22.9, 41.6]), $t(68) = .58$, $p = .566$, the breakpoint, but the positive slope was not significant (Figure 5).

Experiment 1d

The results of Experiment 1c replicated the results of Experiments 1a and 1b. To test whether our preceding results were due to being underpowered, in this experiment, we ran only the 25%, 50%, and 75% deviance conditions and doubled the sample size.

Method

Participants

In Experiment 1d, we recruited 599 participants from the online research platform Prolific. Participants were paid $3.00 as compensation for their time and provided informed consent in accordance with the university IRB. We excluded 39 participants who failed either or both attention checks, leaving a total $n = 560$ (man = 252, woman = 294, nonbinary = 6, prefer not to answer = 8; average age = 36.2).

Procedure

Experiment 1d replicated the procedure in Experiment 1c with the following changes: We limited the deviant threshold to include only the 25%, 50%, and 75% conditions. This experiment also included the PNS scale. Otherwise, Experiment 1d presented the same tasks, attention checks, and demographic questions as previously reported.

Analysis

The analysis procedures were the same as described in Experiment 1c.

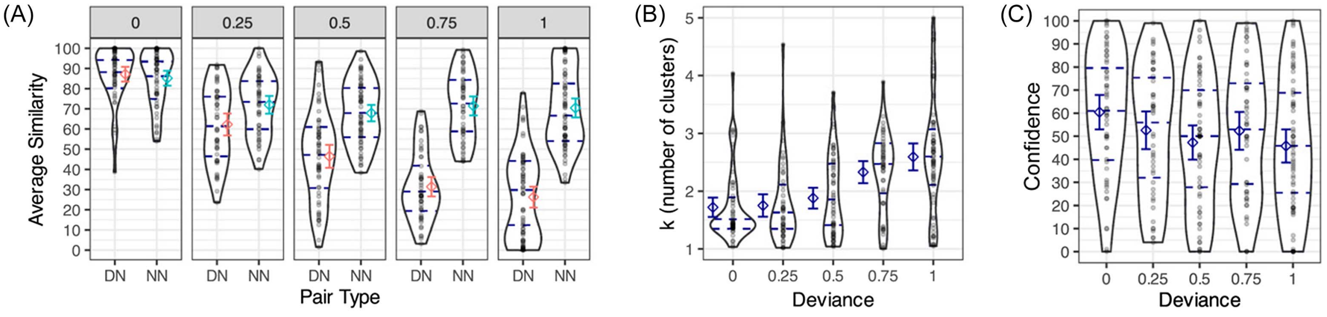
Results

Learning About Nondeviant Agents

Again, we confirmed that participants learned about the agents. Both opinion round, $\chi^2(1) = 280.1$, $p = .006$, and deviance, $\chi^2(2) = 10.31$, $p < .0001$, were significant predictors of accuracy about nondeviant agents' opinions, but not their interaction (p = .887). Participants' accuracy increased with each opinion round (b = .09, SE = .006, CI [.08, .10], $z = 16.72$, $p < .0001$). The 25% deviance condition showed the highest accuracy but was only significantly different than the 75% condition ($M_{\text{difference}}$ = .21, SE = .07,

Figure 5

Results From Experiment 1c



Note. (A) Average similarity ratings by deviance condition and pair type: deviant–nondeviant (DN; red) and nondeviant–nondeviant (NN; blue). (B) Estimates of k (number of clusters) across deviance conditions. (C) Confidence ratings for the new agent's opinion across deviance conditions. Error bars around the mean represent 95% confidence intervals. The violin shape displays the distribution of data, and the dashed lines represent quartiles. See the online article for the color version of this figure.

CI [.04, .38], $z = 2.89$, $p = .011$). The average accuracy on the last opinion round ranged between 73.2% and 76.9% across deviance conditions. These results suggest participants successfully learned the nondeviating agents' opinions, especially when there was no deviant.

Deviant and Structure Learning

We predicted that at low levels of individual deviance, participants would perceive a single group including all agents; however, as deviance of the one "counter-stereotypical" agent increased, that agent would be subtyped out of the group, yielding two clusters. Once again, the number of clusters identified by the ISM, k, changed across deviance conditions, $F(2,\ 557) = 33.15$, $p < .0001$. The average k remained under 2 for the 25% condition ($k = 1.75$, CI [1.65, 1.85], $p < .0001$) and became 2 or greater at 50% ($k = 2.12$, CI [2.02, 2.23], $p = .989$) and 75% ($k = 2.36$, CI [2.25, 2.46], $p = 1.00$) deviation. With more deviation, participants perceived an increased number of clusters among the agents.

This perceived structure occurred in part because participants represented the deviant as belonging to a separate cluster as deviance increased. The interaction between deviance condition and agent pair type (NN vs. DN) was significant in predicting the similarity ratings, $F(1,\ 560) = 114.32$, $p < .0001$. The similarity between the deviant and the rest of the agents was high in the 25% deviation condition ($M = 62.3\%$, CI [59.6, 65.1]) and dropped with increasing deviation ($b = -63.3$, $SE = 3.92$, CI [−71, −55.6]), $t(560) = -16.14$, $p < .0001$, while the similarity between the nondeviating agents started high ($M = 73.9\%$, CI [71.7, 76.2]) and decreased slightly as deviation increased: $b = -8.44$, $SE = 3.29$, CI [−14.9, −1.9], $t(560) = -2.56$, $p = .011$; $b_{\text{difference}} = -54.8$, $SE = 5.13$, CI [−64.9, −44.8], $t(560) = 10.69$, $p < .0001$. Collectively, these results suggest that participants learned both the agents' and deviants' positions well, which shaped the group structures extracted among them. Deviant subgrouping occurred at 25% (lower DN than NN similarity yet still less than two perceived clusters), and deviant subtyping began developing around 50% deviation (much lower DN than NN similarity and two or more perceived clusters).

Confidence Ratings

Our confidence hypothesis was that at low levels of individual deviance, confidence in one's beliefs about the group and, correspondingly, a new group member (about whom participants know nothing) should decrease; however, as the deviant is subtyped out into their own cluster, confidence in one's beliefs about the remaining agents in the group, including the new member, should increase again. The confidence ratings did exhibit the predicted nonmonotonic shape, but only qualitatively. The slopes before ($b = -17.8$, $SE = 10.9$, CI [−39.2, 3.61]), $t(365) = -1.63$, $p = .103$, and after ($b = 6.69$, $SE = 10.9$, CI [−14.7, 28.1]), $t(375) = .62$, $p = .539$, the breakpoint were not significant. None of the slopes in the majority opinion ($ps > .147$) and PNS ($ps > .199$) moderators were significant (Figure 6).

Summary

We observed a consistent pattern of results across Experiments 1a–1d. Across all studies, participants clearly demonstrated a capacity to learn about the agents, including the deviant. Critically, all four sets of results supported our cluster hypothesis but not our confidence rating hypothesis. This did not seem to be driven by insufficient information about individual agents, the absence of an identification of the agents as members of a group, or insufficient power to detect the predicted effect.

Experiment 2

One stark contrast between our Experiments 1a–1d and traditional subtyping/subgrouping literature is that we did not include any category labels. Here, we identified the collectives as either Democrats or Republicans (though note that issue positions were still randomly assigned to collectives).

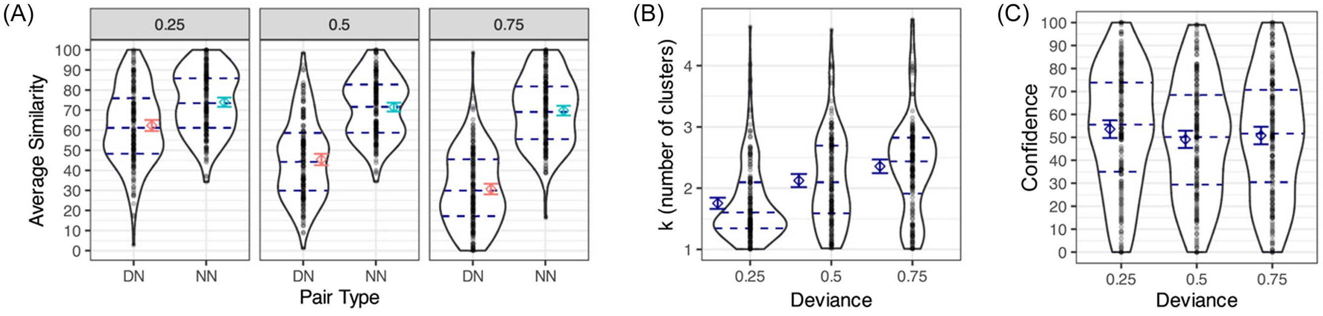
Method

Participants

In Experiment 2, we recruited 300 participants from the online research platform Prolific. Participants were paid $3.00 as compensation

Figure 6

Results From Experiment 1d



Note. (A) Average similarity ratings by deviance condition and pair type: deviant–nondeviant (DN; red) and nondeviant–nondeviant (NN; blue). (B) Estimates of k (number of clusters) across deviance conditions. (C) Confidence ratings for the new agent's opinion across deviance conditions. Error bars around the mean represent 95% confidence intervals. The violin shape displays the distribution of data, and the dashed lines represent quartiles. See the online article for the color version of this figure.

for their time and provided informed consent in accordance with the university IRB. We excluded 23 participants who failed either or both attention checks, leaving a total $n = 277$ (man = 135, woman = 135, nonbinary = 4, another gender not listed here = 1, prefer not to answer = 2; average age = 37.5).

Procedure

Experiment 2 replicated the procedure in Experiment 1b, with the following change: We amended the first phase instructions to randomly assign the agent group to the Democratic or Republican political party condition:

> On the screens that follow you're going to learn about a collection of registered [Democrats/Republicans] we polled on a series of political issues. You are going to make guesses about and receive feedback on their positions on a series of 12 political issues. You don't have to remember what each person's position is, but try to see if you can figure out to what extent each person agrees with everyone else in the group.

After completing the new agent task, participants then selected the party with which they most identified (Democratic, Republican, Independent, or none of these) and reported how strongly they identify with their party on a sliding scale from *not at all* to *very strongly*. As in the previous study, we also included the PNS scale in this experiment. Otherwise, Experiment 2 presented the same tasks, attention checks, and demographic questions as previously reported.

Analysis

The analysis procedures were the same as described in Experiment 1d.

Results

Learning About Nondeviant Agents

Again, we confirmed that participants learned about the agents. Both opinion round, $\chi^2(1) = 91.33, p < .0001$, and deviance, $\chi^2(4) = 23.43$, $p = .0001$, were significant predictors of accuracy about nondeviant agents' opinions, but not their interaction ($p = .637$). Participants' accuracy increased with each opinion round ($b = .084$, $SE = .009$, CI [.07, .10], $z = 9.36$, $p < .0001$), and the 0% deviance condition showed the highest accuracy, but only significantly higher than the 75% and 100% deviance conditions ($M_{\text{differences}} = .52$ to .58, SEs = .14, ps < .003). The average accuracy on the last opinion round ranged between 72.9% and 85.9% across deviance conditions. These results suggest participants successfully learned the nondeviating agents' opinions, especially when there was no deviant.

Deviant and Structure Learning

We predicted that at low levels of individual deviance, participants would perceive a single group including all agents; however, as deviance of the one "counter-stereotypical" agent increased, that agent would be subtyped out of the group, yielding two clusters. Once again, the number of clusters identified by the ISM, k, changed across deviance conditions, $F(4, 272) = 10.27$, $p < .0001$. The average k remained under 2 for the 0% condition ($k = 1.63$, CI [1.44, 1.82], $p = .0001$), reached 2 at 25% deviation ($k = 1.85$, CI [1.64, 2.06], $p = .081$), and became 2 or greater at 50% ($k = 2.14$, CI [1.95, 2.33], $p = .921$), 75% ($k = 2.16$, CI [1.97, 2.34], $p = .955$), and 100% ($k = 2.41$, CI [2.23, 2.60], $p = 1.00$) deviation. With more deviation, participants perceived an increased number of clusters among the agents.

This perceived structure occurred in part because participants represented the deviant as belonging to a separate cluster as deviance increased. The interaction between deviance condition and agent pair type (NN vs. DN) was significant in predicting the similarity ratings, $F(1, 277) = 138.1$, $p < .0001$. The similarity between the deviant and the rest of the agents was high in the 0% deviation condition ($M = 82.1\%$, CI [77.9, 86.2]) and dropped with increasing deviation ($b = -57.8$, $SE = 3.22$, CI [−64.1, −51.5]), $t(277) = -19.97$, $p < .0001$, while the similarity between the nondeviating agents started high ($M = 79.4\%$, CI [74.9, 83.9]) and decreased slightly as deviation increased: $b = -7.37$, $SE = 2.83$, CI [−12.9, −1.8], $t(277) = -2.61$, $p = .009$; $b_{\text{difference}} = -50.4$, $SE = 4.29$, CI [−58.9, −42], $t(277) = -11.75$, $p < .0001$. Collectively, these results suggest that participants learned both the agents' and deviants' positions well, which shaped the group structures extracted among them. Deviant subgrouping occurred at 25% (lower DN than NN similarity and almost less than two perceived clusters), and deviant subtyping began developing around 50% deviation (much lower DN than NN similarity and two or more perceived clusters).

Confidence Ratings

Our confidence hypothesis was that at low levels of individual deviance, confidence in one's beliefs about the group and, correspondingly, a new group member (about whom participants know nothing) should decrease; however, as the deviant is subtyped out into their own cluster, confidence in one's beliefs about the remaining agents in the group, including the new member, should increase again. The nonmonotonic shape in confidence ratings prediction about the new agent's opinion was not supported. The two-line test showed a consistent linear decrease in confidence as deviance increased. The slope between 0% and 50% deviation was negative ($b = -5.5$, $SE = 9.21$, CI [−23.7, 12.7]), $t(155) = -.59$, $p = .554$, as was the slope between 50% and 100% deviation ($b = -9.1$, $SE = 9.82$, CI [−28.4, 10.3]), $t(172) = -.92$, $p = .357$. Neither was significant. None of the slopes in the two-line tests within the moderator analyses were significant (ps > .428). Finally, we examined one additional moderator: whether the participant shared a political party affiliation with the agents (in-group vs. out-group relations); however, none of the slopes in the two-line tests were significant (ps > .200). That is, even when we restricted analysis to the participants who were doing the task with their own political party, we did not observe support for the confidence rating hypothesis.

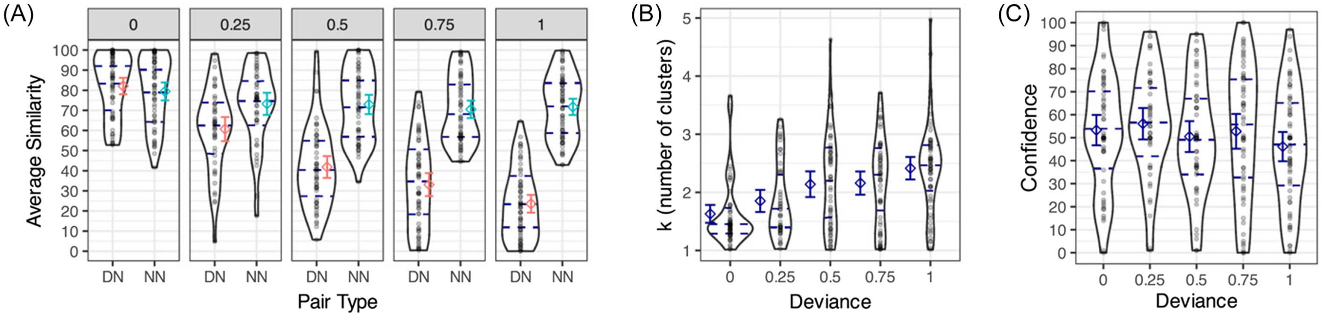
Summary

The results of this experiment comported with the preceding experiments' results. Furthermore, an exploratory analysis indicated that even when participants completed the task for their in-party, we did not observe the predicted U-shaped pattern of confidence ratings (Figure 7C).

Experiment 3

Social categorization is unique compared to nonsocial categorization, because social categorization and identification are intimately

Figure 7
Results From Experiment 2



Note. (A) Average similarity ratings by deviance condition and pair type: deviant–nondeviant (DN; red) and nondeviant–nondeviant (NN; blue). (B) Estimates of *k* (number of clusters) across deviance conditions. (C) Confidence ratings for the new agent's opinion across deviance conditions. Error bars around the mean represent 95% confidence intervals. The violin shape displays the distribution of data, and the dashed lines represent quartiles. See the online article for the color version of this figure.

intertwined—we do not sort people into categories the way we do fruits and vegetables. We sort people into in-groups and out-groups, which are egocentrically defined: those to which I do or do not belong. As we noted in the introduction, people tend to subgroup their in-groups more than their out-groups because they have more information about in-groups (i.e., increased perceived variability within in-groups relative to out-groups). Therefore, in the current experiment, we changed the task so that participants became *part* of the collective and deviant agents' responses were different from the participants' own. We also dispersed deviance across different agents (i.e., different agents disagreed with the participant on different issues) rather than locating deviance of varying degrees in a single agent.

Method

Participants

In Experiment 3, we recruited 345 participants from the online research platform Prolific. Participants were paid $3.00 as compensation for their time and provided informed consent in accordance with the university IRB. We excluded 36 participants who failed either or both attention checks (see below), leaving a total n = 309 (man = 135, woman = 164, nonbinary = 6, another gender not listed here = 2, prefer not to answer = 2; overall average age = 38.3).

Procedure

Unlike the experiments already discussed, this version placed the participant within the context of the group itself. This experiment included the three main phases used in Experiments 1a–1d and 2, as well as the PNS scale and subsequent attention checks and demographic questions. We presented eight agents for each issue, across eight political issues total.

As before, we randomly assigned participants to one of five deviant threshold conditions (0%, 25%, 50%, 75%, 100% deviance). However, we no longer restricted deviance to one agent. Instead, we spread deviance across agents, randomizing *which* agents disagreed with the participants on each issue. In other words, the percentage threshold determined how many agents disagreed with the participants and the rest of the group (i.e., at 25%, for each of the eight issues, two deviants were repeatedly randomly selected from the eight agents to disagree with the participant and other group members).

After completing the initial consent form, participants first encountered the question "Which of the following political parties do you most identify with?" and selected one of the following options: Democratic, Republican, or Independent. Participants next read, "How strongly do you identify with the party you selected?" and recorded their response on a sliding scale with a range of 1 = *not at all* to 100 = *very strongly*.

After completing these two preliminary questions, participants now entered the first phase of the main experiment with instructions about the group amended to reflect their chosen political party: "Like yourself, they also most identify with the [participant's selected political party] party." Prior to guessing the opinions of the group of agents, participants read the issue statement and replied to "What do you think about this statement?" with the option of either clicking "agree" or "disagree."

In the second phase of the study, as before, participants made a series of similarity judgments. Because this group included the participant themselves, a silhouette labeled "You" also appeared as a pairing option with the rest of the agents.

After exiting the final phase, participants rated "How strongly do you identify with the set of people you just encountered?" on a sliding scale of 1 = *not at all* to 100 = *very strongly*. Once again, we asked "How strongly do you identify with the [political party] party?" inserting the political party selected by the participant in the beginning. Participants responded using a sliding scale ranging from 1 = *not at all* to 100 = *very strongly*. The repetition of this question allowed for a before and after comparison of party identification sentiment. The study concluded with the PNS scale, attention checks, and demographic questions as previously discussed.

Analysis

The analysis procedures were the same as described in Experiment 2 with one exception. In this experiment, participants indicated how much they identified with their affiliated political party before and after the main experiment. We conducted exploratory analyses on the pre–post differences in identification ratings using linear mixed models. Ratings were predicted by an interaction between deviance condition and a time indicator (pre vs. post). The models included random intercepts for participants.

Results

Agent Learning

Again, we confirmed that participants learned about the agents. The interaction between opinion round and deviance was significant, $\chi^2(4) = 160.68, p < .0001$. The 0% ($b = .31, SE = .04$, CI [.24, .38], $z = 8.15, p < .0001$) and 100% ($b = .34, SE = .03$, CI [.29, .38], $z = 13.40, p < .0001$) deviance conditions exhibited the largest learning slopes compared to the rest of the deviation conditions, which showed flatter slopes (bs = .02 to .03; $p\text{s}_{\text{differences}} < .0001$). These results indicate participants successfully learned the agents' opinions when there was no deviant or everyone deviated. Note that this pattern occurred because deviation was dispersed randomly across agents (i.e., was not tied to specific agents). In other words, it was impossible to learn who the deviant was in the 25%, 50%, and 75% conditions because the agent disagreeing with the participant and the rest of the group changed issue by issue.

Deviant and Structure Learning

We predicted that at low levels of individual deviance, participants would perceive a single group including all agents; however, as deviance of the one "counter-stereotypical" agent increased, that agent would be subtyped out of the group, yielding two clusters. Once again, the number of clusters identified by the ISM, k, changed across deviance conditions, $F(4, 303) = 10.27, p < .0001$. The average k remained under 2 for the 0% ($k = 1.38$, CI [1.18, 1.59], $p < .0001$), 25% ($k = 1.37$, CI [1.16, 1.57], $p < .0001$), and 50% conditions ($k = 1.56$, CI [1.35, 1.77], $p < .0001$) and became 2 or greater at 75% ($k = 1.99$, CI [1.79, 2.18], $p = .448$) and 100% ($k = 2.20$, CI [2.01, 2.40], $p = .979$) deviation. With more deviation, participants perceived an increased number of clusters among the agents and themselves.

This perceived structure occurred in part because participants represented the deviant as belonging to a separate cluster as deviance increased. The interaction between deviance condition and agent pair type (participant–agents vs. agents–agents) was significant in predicting the similarity ratings, $F(1, 309) = 256.26, p < .0001$. The similarity between the participant and the agents was high in the 0% deviation condition ($M = 84.5\%$, CI [79.3, 89.8]) and dropped with increasing deviation ($b = -62.4, SE = 2.65$, CI [−67.6, −57.2]), $t(309) = -23.5, p < .0001$, while the similarity between the agents started high ($M = 87.4\%$, CI [83.4, 91.3]), decreased for 25% to 75% deviation, and increased again with 100% deviation: $b = 1.24, SE = 3.45$, CI [−5.55, 8.03], $t(309) = .36, p = .719$; $b_{\text{difference}} = -63.6$, $SE = 3.97$, CI [−71.3, −55.8], $t(309) = -16.01, p < .0001$. Collectively, these results suggest that participants learned the agents' positions well in the 0% and 100% deviation conditions, which shaped the group structures extracted among them and how the participants related to them. Participants began subtyping themselves away from the agents by 75% deviation (Figure 7A and 7B).

Confidence Ratings

Our confidence hypothesis was that at low levels of individual deviance, confidence in one's beliefs about the group and, correspondingly, a new group member (about whom participants know nothing) should decrease; however, as the deviant is subtyped out into their own cluster, confidence in one's beliefs about the remaining agents in the group, including the new member, should increase again. By contrast to the preceding studies, the nonmonotonic shape in confidence ratings prediction about the new agent's opinion was supported. The two-line test showed a linear decrease and then increase in confidence as deviance increased. The slope between 0% and 50% deviation was negative ($b = -41.1, SE = 9.39$, CI [−59.6, −22.5]), $t(172) = -4.37, p < .0001$, and the slope between 50% and 100% deviation was positive ($b = 45.6, SE = 9.45$, CI [26.9, 64.2]), $t(189) = 4.82, p < .0001$. Both were significant (Figure 8A and 8C).

To confirm the difference between this experiment and the preceding experiments was significant, we ran the interaction test for each of the two-line tests comparing the confidence rating pattern pooled across Experiments 1a–2 against the confidence rating pattern for Experiment 3. We pooled the first set of studies to reduce the number of analyses. This yielded two interaction tests, both of which were significant: the interaction between deviation levels 0% and 50% by Experiments 1a–2 versus Experiment 3, $F(1, 1{,}285) = 28.25, p < .0001$, and the interaction between deviation levels 50% and 100% by Experiments 1a–2 versus Experiment 3, $F(1, 1{,}334) = 27.195, p < .0001$.

Party Identification

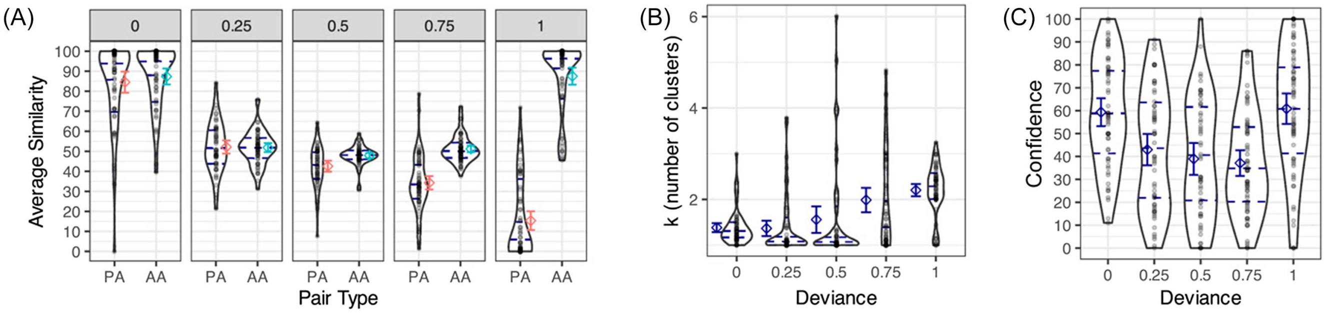
The exploratory analysis indicated that the interaction between deviance and the pre/post time variable was not significant, $F(4, 309) = 2.35, p = .055$, nor was deviation, $F(4, 309) = 1.25, p = .291$. Only time was significantly associated with party identification, $F(1, 309) = 15.89, p < .0001$. On average, the postidentification ratings ($M = 64.4, SE = 1.5$, 95% CI [61.4, 67.4]) were lower than the preratings ($M = 67.6, SE = 1.5$, 95% CI [64.7, 70.6]; $M_{\text{difference}} = 3.22, SE = .81$, 95% CI [1.63, 4.8]), $t(309) = 3.98, p = .0001$.

Experiment 4

One potential reason we observed the predicted U shape in confidence ratings in Experiment 3 is because agent learning was impossible in the 25% to 75% conditions. When deviance is randomly dispersed across agents for each opinion, rather than consistently anchored to specific agents, learning who disagrees from oneself is only possible when everyone (100%) or no one (0%) deviates. This can be observed in the chance accuracies in the opinion learning task (Supplemental Figure S1B), the middle-of-the-scale similarity ratings between agents (Figure 7A), and the perceived number of clusters increasing to 2 at greater deviation than the previous experiments (Figure 7B). Put simply, the U shape in confidence ratings may have occurred from a learning confound that differentiates the 0 and 100% deviance conditions from the 25 to 75% conditions. To address this issue, we replicated Experiment 3, but this time, deviance was tied to specific agents in the group. In other words, participants could learn which other person was consistently disagreeing with them. Evidence of agent learning in all conditions alongside U-shaped confidence ratings would provide clearer support for the predicted downstream consequences of perceived group structures on group-based beliefs.

Figure 8

Results From Experiment 3



Note. (A) Average similarity ratings by deviance condition and pair type: participant–agents (PA; red) and agents–agents (AA; blue). (B) Estimates of *k* (number of clusters) across deviance conditions. (C) Confidence ratings for the new agent's opinion across deviance conditions. Error bars around the mean represent 95% confidence intervals. The violin shape displays the distribution of data, and the dashed lines represent quartiles. See the online article for the color version of this figure.

Method

Participants

In Experiment 4, we recruited 298 participants from the online research platform Prolific. Participants were paid $3.00 as compensation for their time and provided informed consent in accordance with the university IRB. We excluded 11 participants who failed either or both attention checks (see below), leaving a total $n = 287$ (man = 121, woman = 155, nonbinary = 11; overall average age = 36.6).

Procedure

Experiment 4 replicated the procedure in Experiment 3 with the following changes. We randomly assigned participants to one of five deviant threshold conditions (0%, 25%, 50%, 75%, 100% deviance). However, in this version of the experiment, the percentage threshold determined the number of deviant agents and assigned deviancy to specific agents from issue to issue (i.e., at 25%, two specific deviants disagree with the majority position throughout each of the eight political issues).

This study concluded with the PNS scale, attention checks, and demographic questions as previously discussed.

Analysis

The analysis procedures were the same as described in Experiment 3 with one exception. Due to the design of this experiment, which includes the participant (P), deviants (D; deviate from participant opinion), and majority (M; agents who agree with the participant), the comparison of pair type in the similarity analysis becomes mismatch (DM, PD) versus match (DD, MM, PM), which separates pairs that disagree versus agree.

Results

Agent Learning

Again, we confirmed that participants learned about the agents. The interaction between opinion round and deviance was significant, $\chi^2(4) = 32.91$, $p < .0001$. The 100% deviance condition ($b = .42$, $SE = .04$, CI [.35, .50], $z = 10.75$, $p < .0001$) exhibited the largest learning slopes compared to the rest of the deviation conditions (bs = .17 to .28, $p\text{s}_{\text{differences}} \leq .0001$ to .019). On the last rounds, accuracies across conditions ranged between 73.3% and 92.9%. These results suggest participants successfully learned the agents' opinions, especially when everyone deviated.

Deviant and Structure Learning

We predicted that at low levels of individual deviance, participants would perceive a single group including all agents; however, as deviance of the one "counter-stereotypical" agent increased, that agent would be subtyped out of the group, yielding two clusters. Once again, the number of clusters identified by the ISM, k, changed across deviance conditions, $F(4, 282) = 9.18$, $p < .0001$. The average k remained under 2 for 0% deviation ($k = 1.59$, CI [1.30, 1.88], $p = .003$) and became 2 or greater at 25% ($k = 2.58$, CI [2.28, 2.89], $p = .999$), 50% ($k = 2.52$, CI [2.16, 2.89], $p = .998$), 75% ($k = 2.69$, CI [2.39, 2.99], $p = 1.00$), and 100% ($k = 2.58$, CI [2.24, 2.91], $p = .999$) deviation. With more deviation, participants perceived an increased number of clusters among the agents and themselves.

This perceived structure occurred in part because participants represented the deviant as belonging to a separate cluster as deviance increased. The interaction between deviance condition and agent pair type (match vs. mismatch with the participant) was significant in predicting the similarity ratings, $F(1, 329.31) = 7.81$, $p = .006$. The similarity between the participants and mismatching agents started low in the 25% deviation condition ($M = 24.1\%$, CI [19.9, 28.3]) and decreased slightly with increasing deviation ($b = -3.94$, $SE = 3.57$, CI [−0.97, 3.09]), $t(334) = -1.10$, $p = .271$, while the similarity between the participants and matching agents started high ($M = 75.5\%$, CI [72.1, 78.9]) and increased as deviation increased: $b = 12.52$, $SE = 2.90$, CI [6.81, 18.2], $t(282) = 4.32$, $p < .0001$; $b_{\text{difference}} = 16.5$, $SE = 5.89$, CI [4.87, 28], $t(329) = 2.79$, $p = .006$. Collectively, these results suggest that participants quickly learned both the agents' and deviants' positions, which shaped the group structures extracted among them and how the participants related to them. Participants already subtyped themselves away from the mismatching agents by 25% deviation (Figure 9A and 9B).

Confidence Ratings

Our confidence hypothesis was that at low levels of individual deviance, confidence in one's beliefs about the group and, correspondingly, a new group member (about whom participants know nothing) should decrease; however, as the deviant is subtyped out into their own cluster, confidence in one's beliefs about the remaining agents in the group, including the new member, should increase again. The nonmonotonic shape in confidence ratings prediction about the new agent's opinion was again supported. The two-line test indicated a linear decrease and then increase in confidence as deviance increased. The slope between 0% and 50% deviation was negative ($b = -38.9$, $SE = 9.80$, CI [−58.2, −19.6]), $t(171) = -3.97$, $p = .0001$, and the slope between 50% and 100% deviation was positive ($b = 58.3$, $SE = 10.3$, CI [37.9, 78.7]), $t(155) = 5.64$, $p < .0001$. Both were significant (Figure 9C).

To confirm the difference between this experiment and Experiments 1a–2 was significant, we ran the interaction test for each of the two-line tests comparing the confidence rating pattern pooled across Experiments 1a–2 against the confidence rating pattern for Experiment 4. Again, we pooled the first set of studies to reduce the number of analyses. This yielded two interaction tests, one of which was significant: the interaction between deviation levels 0% and 50% by Experiments 1a–2 versus Experiment 4, $F(1, 1{,}294) = 3.195$, $p = .074$, and the interaction between deviation levels 50% and 100% by Experiments 1a–2 versus Experiment 4, $F(1, 1{,}285) = 28.25$, $p < .0001$.

Party Identification

Exploratory analyses indicated the interaction between deviance and the pre/post time variable was not significant, $F(4, 287) = .59$, $p = .670$, nor was deviation, $F(4, 287) = .35$, $p = .841$. Only time was significantly associated with party identification, $F(1, 287) = 7.83$, $p = .006$. On average, the postidentification ratings ($M = 66.2$, $SE = 1.5$, 95% CI [63.3, 69.2]) were lower than the preratings ($M = 68.5$, $SE = 1.5$, 95% CI [65.5, 71.4]; $M_{\text{difference}} = 2.23$, $SE = .80$, 95% CI [.66, 3.8]), $t(287) = 2.79$, $p = .006$.

Summary

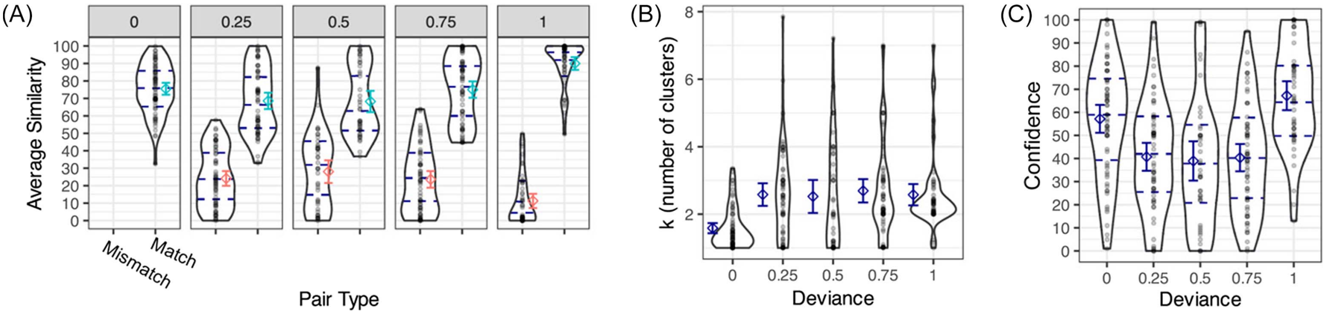
Once participants were part of the group and deviance was yoked to specific agents, we observed support for both the cluster and confidence rating hypotheses.

Discussion

The social structure learning model aims to capture how perceiving the presence and structure of groups in social environments arises through statistical learning (Gershman & Cikara, 2020; Gershman et al., 2017; Lau et al., 2018). Here, we tested predictions made from a domain-general version of the model that further aims to develop a unified explanation, which links perceptual inferences about group structures to stereotype change (Gershman & Cikara, 2023). Specifically, we test the prediction that feature/behavioral alignment (in this case, alignment of policy-related preferences) is an important ingredient for how groups are perceptually constructed, and therefore, behavioral deviance or dissimilarity will affect the group structures perceived among a collection of people. These perceived group structures should, in turn, produce downstream consequences on the malleability of perceived group stereotypes—here measured by confidence in beliefs about the opinion of a new incoming group member.

Across seven experiments, we found consistent evidence for the first prediction that increasing an agent's deviance from other agents changed the perceived group structure among them (Figure 1, left panel). Without deviation, all agents were perceived as similar, part of the same cohesive group. As deviation increased to 25%, the deviant was *subgrouped* as evidenced by the perceived decrease in similarity between the deviant and nondeviants, yet the perceived structure stayed below 2, meaning the agents were still perceived as one group with some internal structure. As deviation increased to 50% and more, the deviant was *subtyped* out of the group: Perceived clusters increased to two or more, and the perceived deviant–nondeviant similarity continued to decrease. However, the perceived group structure did not consistently influence downstream

Figure 9
Results From Experiment 4



Note. (A) Average similarity ratings by deviance condition and pair type: deviant–nondeviant (mismatch; red) and nondeviant–nondeviant (match; blue). (B) Estimates of *k* (number of clusters) across deviance conditions. (C) Confidence ratings for the new agent's opinion across deviance conditions. Error bars around the mean represent 95% confidence intervals. The violin shape displays the distribution of data, and the dashed lines represent quartiles. In the case of 0% deviation from the participant, there are no agents who mismatch. See the online article for the color version of this figure.

confidence judgments about a new group member as predicted by the model. Specifically, participants' confidence in their predictions about a novel group member's alignment did not reflect the predicted nonmonotonic relationship to deviance (i.e., that participants should have reported greater confidence in deviance conditions where group structures are clearest [0% deviation = one cohesive group, 100% deviation = two distinct groups: nondeviants vs. deviant]) relative to the intermediate conditions. It was only in Experiments 3 and 4 that the U shape in confidence ratings across deviation conditions emerged.

Why did we only observe support for our confidence rating prediction in Experiments 3 and 4? First, it is important to note that the inconsistent downstream consequences were not a result of participant inattention or lack of learning the agent's behavioral patterns. Accuracies were consistently above 70% on the last opinion learning rounds in all of the experiments where deviation was anchored to specific deviants (Supplemental Figure S1). Moreover, perceived similarity between deviants and nondeviants decreased and the number of perceived clusters increased with greater deviation in every experiment. This means participants learned about individual agents and formed the corresponding clusters.

Experiments 3 and 4 incorporated two major design differences from the other experiments. First, following up on previous research showing that dispersion in stereotype-inconsistent behaviors facilitates stereotype change (Weber & Crocker, 1983), Experiment 3 dispersed deviant opinions across agents rather than tying deviation to specific agents. However, dispersion also affected agent learning in our experimental design—agents' opinions were impossible to learn in conditions where agents randomly exhibited a mix of deviance and nondeviance. Therefore, the relationship between dispersed deviance and stereotype change in the social structure learning model was confounded by the inability to learn about specific deviants. However, Experiment 4 addressed this problem. There, we saw that even when participants could learn about deviant agents, learning was as good as in the preceding experiments (in fact, the deviant was subtyped out of the group at 25%—as indicated by the two-cluster result from the ISM; Figure 8B) and the confidence ratings conformed to the predictions made by the model.

The other difference between Experiments 3 and 4 and the preceding studies was that participants were in the group. One possibility is that inclusion in the group allowed the participants to create ideologically realistic constellations of policy preferences from which deviants could deviate (by contrast to the other studies where policy positions for the majority were randomly determined). Though this "coherence criterion" is not presently represented in the model as described in the introduction (and resulting prediction depicted in Figure 1), it may represent one way that social grouping differs from nonsocial grouping and could be integrated as a prior—as people's beliefs about which kinds of behaviors are expected to be similar (Tamir & Thornton, 2018; Thornton & Tamir, 2021).

To address whether coherence rather than mere inclusion of self in the group was driving the difference across the two sets of studies, we examined how ideologically coherent respondents' positions even *could* be given that we had their party affiliation data in Experiments 3 and 4. For example, if these issues allowed for very clear party-aligned ideologies to emerge within the group of agents about whom they learn, participants would have to seed the group with coherence in the first place. Specifically, we would expect higher consensus within parties relative to between parties across issues (e.g., 90% of Democrat participants but only 15% of Republican participants endorse "Planned Parenthood is underfunded"). Notably, the political science literature has documented individual voters are rarely perfectly aligned with their party's platform (except on a few high-profile issues) in part because they are relatively unaware of changes in party platforms (e.g., Adams et al., 2011; Fernandez-Vazquez, 2014).

Consistent with the literature in political science, we found low levels of within (relative to between) party coherence across the issues in these experiments. Moreover, only two out of the 15 possible issues were consistently polarized across the experiments: "Strict voter ID laws are needed to prevent voter fraud" and "Planned Parenthood is underfunded" (see Supplemental Figure S2). These results, coupled with the finding that the group labels "Democrat" and "Republican" did not generate a U-shaped curve in the confidence ratings in Experiment 2, strongly suggest that the differences between the two sets of experiments (1a–2 vs. 3 and 4) are better explained by participants merely being in the groups rather than by any pattern of ideological coherence across the issues that might have emerged from participants seeding the groups with their own constellations of preferences.

A Different Approach to Measuring Stereotype Updating

Yet another reason our results did not support the confidence rating prediction in the first several studies may simply be that the confidence ratings about a novel agent were an inadequate way to index stereotype updating. One very interesting pattern that appeared (to varying degrees) across studies is the changes in the rated nondeviant-to-nondeviant similarity across conditions. For example, Figure 2A depicts that the similarity between the deviant and the nondeviants (in red) clearly decreases as deviance decreases. What we have not yet discussed is that the similarity among all the majority members (in blue) also changes across conditions: It first decreases and then increases again nonmonotonically as deviance increases. From a purely statistical standpoint, this should not occur—the probability of agreement among all the majority members is held constant across all conditions.

One possibility is that this change in perception of how "aligned" the majority group members are with one another is an indirect measure of the group stereotype, which changes across conditions as predicted by the model. Consider the following scenario: If all majority members were identical to one another, then the policy-issue similarity among them should be 1. In other words, each agent should be perfectly substitutable for every other agent (except the deviant), so the underlying beliefs or stereotype about the collective is very strong, with zero variability (a very sharp peak, if beliefs were plotted as a distribution). Because we included some noise into agents' positions—a 5% chance that any agent would flip on a given issue—it created some dispersion in the underlying stereotype (a small widening of the peak). What is so interesting in the present data is that this dispersion seems to increase as the deviance of the one deviant agent increases, but then it decreases once the deviant is subtyped out (after 50%). Said another way, the deviant affected how "sharp" the stereotype was of the remaining agents.

In a follow-up set of exploratory analyses, we submitted the nondeviant–nondeviant similarities to the two-line test (see Supplemental Table S1). Across all experiments, we see a negative slope between 0% and 50%; across Experiments 1a–c, 3, and 4, we see a positive slope between 50%, though it does not always reach statistical significance.

Limitations

The collective results highlight a limitation in the current specification of the social structure learning model. A more substantial change is required to explain the genesis of stereotypes as opposed to beliefs about any other statistical structure. For example, the model could incorporate parameters that distinguish how and whether the self or expected behavioral coherence is represented within the various group representations. A related limitation is that our experiments indirectly assessed stereotype change by using confidence ratings as a proxy for measuring participants' beliefs about the groups' beliefs. Future work could more directly test stereotype change by measuring perceived stereotypes and changes in those perceptions as perceived group structures change. Other exciting extensions of the model and accompanying empirical work could test how these structures update when there are multiple groups in the environment or what happens when people have sufficient evidence about any one agent to begin to individuate them: that is, represent them along a continuum of a given feature dimension rather than as a member of a cluster. A third limitation is that we defined stereotypes here as alignment of policy-related beliefs and preferences. Future work should interrogate whether these results extend to structure learning based on other features that may be more or less mutable (e.g., traits, physical features, group norms).

Conclusions

Understanding how group stereotypes change requires understanding what people think constitutes a group in the first place. As such, we tested predictions from a domain-general computational model that aimed to provide a unified account that connects social structure learning to stereotype change (Gershman & Cikara, 2023). Through careful experimentation, we identified a set of important factors that influence both group structure and group-related beliefs. The first is the predictability of behavior. First, we were able to provide initial quantitative evidence for how much deviance is required before a deviant is subtyped from a group. In our data, 25% was enough to begin perceiving internal structure within one group (i.e., subgrouping) and 50% deviation was enough for participants to reliably perceive more than one cluster among all the agents (i.e., subtyping). This 25% tipping point converges with the amount of deviation needed within a majority group for social norm structures to shift (Centola & Baronchelli, 2015). The second, and crucial, factor, particularly for affecting group-related beliefs, was participants' membership in the perceived group structure. Once participants were in the groups themselves, their structure learning *and* group-based beliefs began to match the model's predictions. These factors highlight how learning or inferring structures in a social world might differ from domain-general statistical learning and the need for computational models to account for this added social complexity.

Constraints on Generality

Here, we have tested the model using nonrepresentative, convenience samples of online respondents. Specifically, all of our experiments were run with convenience samples from the United States with political issues and parties specific to the U.S. context. Moreover, we have tested the model using constrained lab experiments with fabricated agents and opinion profiles. These testing conditions suggest care should be taken when trying to extrapolate beyond these conditions, and future work could assess the contextual generalizability of this model more directly. Though we were focused mostly on understanding how people construct groups de novo, political preferences are only one possible dimension among many that act as inputs to creating socially meaningful groups. Consequently, we cannot know if and how exactly our pattern of results translates to groups and individuals beyond those sampled here. However, it remains important to consider where and when predictions derived from the social structure learning model should not apply. One potential boundary condition is the size of a society. Specifically, in smaller scale societies where members intimately know everyone in their network (Smaldino, 2019), there may not be a need to build abstracted feature–attribute associations of the kinds the social structure learning model is designed to approximate.

References

Adams, J., Ezrow, L., & Somer-Topcu, Z. (2011). Is anybody listening? Evidence that voters do not respond to European parties' policy statements during elections. *American Journal of Political Science*, *55*(2), 370–382. https://doi.org/10.1111/j.1540-5907.2010.00489.x

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

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429. https://doi.org/10.1037/0033-295X.98.3.409

Austerweil, J. L., Gershman, S. J., Tenenbaum, J. B., & Griffiths, T. L. (2015). Structure and flexibility in Bayesian models of cognition. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 187–208). Oxford University Press.

Bai, X., Fiske, S. T., & Griffiths, T. L. (2022). Globally inaccurate stereotypes can result from locally adaptive exploration. *Psychological Science*, *33*(5), 671–684. https://doi.org/10.1177/09567976211045929

Bai, X., Griffiths, T., & Fiske, S. (2022). *Multidimensional stereotypes emerge spontaneously when exploration is costly*. PsyArxiv. https://doi.org/10.31234/osf.io/mbuhv

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the Association for Computing Machinery*, *57*(2), 1–30. https://doi.org/10.1145/1667053.1667056

Brewer, M. B., & Lui, L. (1984). Categorization of the elderly by the elderly: Effects of perceiver's category membership. *Personality and Social Psychology Bulletin*, *10*(4), 585–595. https://doi.org/10.1177/0146167284104012

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science*, *3*(1), 14–25. https://doi.org/10.1002/bs.3830030103

Centola, D., & Baronchelli, A. (2015). The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(7), 1989–1994. https://doi.org/10.1073/pnas.1418838112

Du, M., Basyouni, R., & Parkinson, C. (2021). How does the brain navigate knowledge of social relations? Testing for shared neural mechanisms for shifting attention in space and social knowledge. *NeuroImage*, *235*, Article 118019. https://doi.org/10.1016/j.neuroimage.2021.118019

Fernandez-Vazquez, P. (2014). And yet it moves the effect of election platforms on party policy images. *Comparative Political Studies*, *47*(14), 1919–1944. https://doi.org/10.1177/0010414013516067

Ge, H., Xu, K., & Ghahramani, Z. (2018). Turing: A language for flexible probabilistic inference. In A. J. Storkey & F. Pérez-Cruz (Eds.), *International conference on artificial intelligence and statistics* (pp. 1682–1690). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v84/ge18b.html

Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1–12. https://doi.org/10.1016/j.jmp.2011.08.004

Gershman, S. J., & Cikara, M. (2020). Social-structure learning. *Current Directions in Psychological Science*, *29*(5), 460–466. https://doi.org/10.1177/0963721420924481

Gershman, S. J., & Cikara, M. (2023). Structure learning principles of stereotype change. *Psychonomic Bulletin & Review*, *30*(4), 1273–1293. https://doi.org/10.3758/s13423-023-02252-y

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology*, *20*(2), 251–256. https://doi.org/10.1016/j.conb.2010.02.008

Gershman, S. J., Pouncy, H. T., & Gweon, H. (2017). Learning the structure of social influence. *Cognitive Science*, *41*(Suppl. 3), 545–575. https://doi.org/10.1111/cogs.12480

Hamilton, D. L., Sherman, S. J., Crump, S. A., & Spencer-Rodgers, J. (2009). The role of entitativity in stereotyping: Processes and parameters. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 179–198). Psychology Press. https://doi.org/10.4324/9781841697772

Hewstone, M. (1994). Revision and change of stereotypic beliefs: In search of the elusive subtyping model. *European Review of Social Psychology*, *5*(1), 69–109. https://doi.org/10.1080/14792779543000020

Hewstone, M., & Hamberger, J. (2000). Perceived variability and stereotype change. *Journal of Experimental Social Psychology*, *36*(2), 103–124. https://doi.org/10.1006/jesp.1999.1398

Huddy, L., & Virtanen, S. (1995). Subgroup differentiation and subgroup bias among Latinos as a function of familiarity and positive distinctiveness. *Journal of Personality and Social Psychology*, *68*(1), 97–108. https://doi.org/10.1037/0022-3514.68.1.97

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

Judd, C. M., Park, B., Ryan, C. S., Brauer, M., & Kraus, S. (1995). Stereotypes and ethnocentrism: Diverging interethnic perceptions of African American and white American youth. *Journal of Personality and Social Psychology*, *69*(3), 460–481. https://doi.org/10.1037/0022-3514.69.3.460

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. *Proceedings of the 21st national conference on artificial intelligence and the 18th innovative applications of artificial intelligence conference, AAAI-06/IAAI-06* (pp. 381–388). Princeton University.

Kunda, Z., & Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: Constructing grounds for subtyping deviants. *Journal of Personality and Social Psychology*, *68*(4), 565–579. https://doi.org/10.1037/0022-3514.68.4.565

Kunda, Z., & Oleson, K. C. (1997). When exceptions prove the rule: How extremity of deviance determines the impact of deviant examples on stereotypes. *Journal of Personality and Social Psychology*, *72*(5), 965–979. https://doi.org/10.1037/0022-3514.72.5.965

Lau, T., Pouncy, H. T., Gershman, S. J., & Cikara, M. (2018). Discovering social groups via latent structure learning. *Journal of Experimental Psychology: General*, *147*(12), 1881–1891. https://doi.org/10.1037/xge0000470

Lenth, R. V. (2020). *emmeans: Estimated marginal means, aka least-squares means* (Version 1.5.3) [Computer software]. https://github.com/rvlenth/emmeans

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

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

Martinez, J. E., Krasner, R., Rosero, L., Cikara, M., & Gershman, S. J. (2024, June 13). *Social group discovery, structure, and stereotype updating*. https://osf.io/4hzr2

Maurer, K. L., Park, B., & Rothbart, M. (1995). Subtyping versus subgrouping processes in stereotype representation. *Journal of Personality and Social Psychology*, *69*(5), 812–824. https://doi.org/10.1037/0022-3514.69.5.812

Neuberg, S. L., & Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simpler structure. *Journal of Personality and Social Psychology*, *65*(1), 113–131. https://doi.org/10.1037/0022-3514.65.1.113

Park, B., & Judd, C. M. (1990). Measures and models of perceived group variability. *Journal of Personality and Social Psychology*, *59*(2), 173–191. https://doi.org/10.1037/0022-3514.59.2.173

Park, B., & Rothbart, M. (1982). Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. *Journal of Personality and Social Psychology*, *42*(6), 1051–1068. https://doi.org/10.1037/0022-3514.42.6.1051

Park, B., Ryan, C. S., & Judd, C. M. (1992). Role of meaningful subgroups in explaining differences in perceived variability for in-groups and out-groups. *Journal of Personality and Social Psychology*, *63*(4), 553–567. https://doi.org/10.1037/0022-3514.63.4.553

Rothbart, M., & Lewis, S. (1988). Inferring category attributes from exemplar attributes: Geometric shapes and social categories. *Journal of Personality and Social Psychology*, *55*(6), 861–872. https://doi.org/10.1037/0022-3514.55.6.861

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167. https://doi.org/10.1037/a0020511

Schwyck, M. E., Du, M., Li, Y., Chang, L. J., & Parkinson, C. (2024). Similarity among friends serves as a social prior: The assumption that "birds of a feather flock together" shapes social decisions and relationship beliefs. *Personality and Social Psychology Bulletin*, *50*(6), 823–840. https://doi.org/10.1177/01461672221140269

Simonsohn, U. (2018). Two lines: A valid alternative to the invalid testing of U-shaped relationships with quadratic regressions. *Advances in Methods and Practices in Psychological Science*, *1*(4), 538–555. https://doi.org/10.1177/2515245918805755

Smaldino, P. E. (2019). Social identity and cooperation in cultural evolution. *Behavioural Processes*, *161*, 108–116. https://doi.org/10.1016/j.beproc.2017.11.015

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences*, *22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005

Taylor, S. E. (1981). A categorization approach to stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 88–114). Erlbaum.

Thornton, M. A., & Tamir, D. I. (2021). People accurately predict the transition probabilities between actions. *Science Advances*, *7*(9), Article eabd4995. https://doi.org/10.1126/sciadv.abd4995

Wallace, D. S., Lord, C. G., & Ramsey, S. L. (1995). Relationship between self-typicality and the in-group subtypes effect. *Personality and Social Psychology Bulletin*, *21*(6), 581–587. https://doi.org/10.1177/0146167295216004

Weaverdyck, M. E., & Parkinson, C. (2018). The neural representation of social networks. *Current Opinion in Psychology*, *24*, 58–66. https://doi.org/10.1016/j.copsyc.2018.05.009

Weber, R., & Crocker, J. (1983). Cognitive processes in the revision of stereotypic beliefs. *Journal of Personality and Social Psychology*, *45*(5), 961–977. https://doi.org/10.1037/0022-3514.45.5.961

Received June 3, 2024
Revision received June 20, 2025
Accepted July 1, 2025 ■