Structure learning principles of stereotype change

Samuel J. Gershman[1,2] and Mina Cikara[1]

[1]Department of Psychology, Harvard University, Cambridge, MA

[2]Center for Brains, Minds, and Machines, Cambridge, MA

November 1, 2022

Acknowledgments. We are grateful to Mahzarin Banaji for inspiring discussions and to Joel Martinez and Amit Goldenberg for comments on a draft of the manuscript. This work was supported by a grant from the National Science Foundation (BCS-2116543) and the Center for Brains, Minds, and Machines (CBMM), funded by NSF STC award CCF-1231216. Simulation code is publicly available at `https://github.com/sjgershm/Hierarchical-groups`. Correspondence should be addressed to Samuel Gershman, 52 Oxford St., Cambridge, MA 02138. Email: gershman@fas.harvard.edu.

Abstract

Why, when, and how do stereotypes change? This paper develops a computational account based on the principles of structure learning: stereotypes are governed by probabilistic beliefs about the assignment of individuals to groups. Two aspects of this account are particularly important. First, groups are flexibly constructed based on the distribution of traits across individuals; groups are not fixed, nor are they assumed to map on to categories we have to provide to the model. This allows the model to explain the phenomena of group discovery and subtyping, whereby deviant individuals are segregated from a group, thus protecting the group's stereotype. Second, groups are hierarchically structured, such that groups can be nested. This allows the model to explain the phenomenon of subgrouping, whereby a collection of deviant individuals is organized into a refinement of the superordinate group. The structure learning account also sheds light on several factors that determine stereotype change, including perceived group variability, individual typicality, cognitive load, and sample size.

Keywords: Bayesian modeling; intergroup cognition; social structure learning; stereotypes

Introduction

Stereotypes are notoriously resistant to change. Lippmann (1922) famously quipped that "there is nothing so obdurate to education or criticism as the stereotype" (p. 99). Eight decades later, this view was echoed by Banaji (2002):

> Stereotypes are the vehicles of essentialist thinking about social groups. Dispositional group attributions, or the belief that groups are inherently the way they are, can lead to the assessment that attributes associated with groups are stable and unchanging. (p. 15102)

According to this cognitive view, the obduracy of stereotypes is grounded in beliefs about the structure of social groups. Stereotypes do not change because people believe groups do not change.

And yet stereotypes *do* change over longer periods of time (Bergsieker et al., 2012). For example, public opinion polls show that gender stereotypes track changes in social and occupational roles (Eagly et al., 2020; Miller et al., 2015), a pattern also reflected in measures of implicit attitudes (Charlesworth and Banaji, 2021). As we review below, stereotypes can be changed under certain circumstances, even in the short term. Our goal in this paper is to understand the principles governing these changes: why, when, and how do stereotypes change in response to experience?

Stereotypes' reputation for obduracy derives in part from studies showing that people often fail to update group stereotypes in response to counter-stereotypical individuals. Evidence suggests that this occurs because observers mentally segregate counter-stereotypical individuals into "subtypes" such that the stereotype is effectively protected from disconfirmation (see Hewstone, 1994, for a review). In other words, counter-stereotypical targets get subtyped *out* of the group.[1] Subtyping occurs when counter-stereotypical individuals are sufficiently deviant that they can be classified as outliers; by "deviant" we mean distance from the mode of the trait distribution. Under lower levels of deviance, however, counter-stereotypical individuals are assimilated into the

[1] Allport (1954) used the term "refencing," which he characterized as a mental device for holding onto prejudgments in the face of contradictory evidence. The first direct references to subtyping appear in Ashmore (1981), Brewer et al. (1981), and Taylor (1981).

group, resulting in the stereotype change we referenced above.

An alternative to updating and subtyping is subgrouping, which refers to the reclassification of stereotype-inconsistent individuals into a subordinate group that nevertheless remains a part of the superordinate group. Specifically, if multiple individuals share a common pattern of deviance, then they may be assimilated into a "subgroup" of the superordinate group (Maurer et al., 1995; Park and Judd, 1990; Park et al., 1992; Richards and Hewstone, 2001). In this case, the group stereotype still undergoes a change; however, it is a lesser degree of change relative to assimilation without subgrouping. This is because the stereotype is still anchored to the superordinate group but also has to accommodate the deviation represented by the subgroup. While the concepts of subtyping and subgrouping have strongly shaped the study of stereotyping, a precise understanding of these processes remains elusive. When should we observe one versus the other?

Stereotypes reflect beliefs about covariation between group membership and traits; we begin from the premise that the relationship between traits and group assignment is bidirectional. Specifically, we propose that stereotyping and group assignment are two aspects of a single process. Assignment of a person to a category or a group governs which associated traits we attribute to them. Individuals' traits, in turn, govern group assignment (e.g., a person may or may not get assigned to a given group as a function of how different they are from the group), which then determines whether and how much associated group stereotypes change. These two aspects of stereotyping have typically been investigated separate from one another. Our goal in this paper is to formalize both aspects and incorporate them into a unified computational model of stereotype change. This model makes two core contributions. First, it helps organize existing findings by offering a flexible model which can account for a host of previously documented effects. Second, it generates new predictions (and potentially new targets for intervention) regarding when, why, and how stereotypes are updated.

Another notable strength of our approach is how it treats the concept of traits. We use "traits" to refer to *any* features of the targets under consideration, including behavior. Deviance may therefore arise both from features that are intrinsic to the target (e.g., counter-stereotypical personality traits, physical features) as well as those that are manifested in their behavior (e.g., neighborhoods

they choose to live in, clothes they wear, protests they attend). This flexibility expands our approach's generalizability and modeling scope considerably.

The paper proceeds in three parts. First, we develop a "rational analysis" (cf. Anderson, 1990) that addresses *why* we have stereotypes. The answer, in brief, is that stereotypes enable probabilistic inferences about traits in the presence of unreliable individuating information, or in the absence of individuating information altogether. This analysis also clarifies why stereotypes change: they must track the probability distribution of traits within a group. If group members change or new group members exhibit different traits, the probability distribution of traits changes, too. Importantly, observing a counter-stereotypical individual does not necessarily indicate that the distribution has changed. It could alternatively indicate a new distribution (i.e., a new group), and hence generate a new stereotype while leaving the original stereotype intact. Thus, the answer to the *when* question is that stereotypes change when counter-stereotypical individuals are assimilated into an existing group. The phenomenon of subgrouping exposes a more nuanced answer to this question, whereby individuals can be *partially* assimilated into a group when they cohere with a subset of individuals in the group that remains a part of the superordinate group.

Next, we will answer the *how* question in several steps, building from simple to more complex models of stereotyping. We begin with the setting in which group membership is known or observable (the most widely studied context of stereotype change). This allows us to formalize a principled probabilistic definition of stereotypes as beliefs about the distribution of traits conditional on group membership. Stereotype formation is then modeled as learning about the parameters governing the trait distribution for each group. We show that this model can account for several well-known aspects of stereotyping, including illusory correlation, accentuation, and outgroup homogeneity effects. In each of these applications we note novel predictions made by our account and compare it to alternative existing accounts to highlight its added utility.

We then consider the setting in which group membership is unobservable and hence must be inferred—the major innovation of our paper. This stands in stark contrast to previous work on the topic of stereotypes and updating in which categories are made explicit (e.g., via labels, phenotypic features) to observers. We show how this setting corresponds to a *structure learning*

problem similar to the kind that has been studied in other cognitive domains (Austerweil et al., 2015). Structure learning refers to the acquisition of representations that organize domain-specific knowledge—the discovery of hypothesis spaces. For example, before we can learn about the perceptual similarities between colors or the biological relationships between animals, we need to learn that colors are organized in a circle and animals are organized in a taxonomic tree (Kemp and Tenenbaum, 2008). We apply this framework to the discovery of an important social hypothesis space: the organization of individuals into groups, what we refer to as *social structure learning* (Gershman and Cikara, 2020; Lau et al., 2020; Gershman et al., 2017b; Lau et al., 2018; Spicer and Sanborn, 2017). We show how social structure learning can provide an account of how trait observation governs group assignment: specifically, subtyping and the various factors that moderate it (e.g., dispersion, variability, sample size, typicality, degree of deviance, and cognitive load). To formalize subgrouping, we extend the model to hierarchically structured groups, where subordinate groups can be nested within superordinate groups. This model can explain, for example, how instructing people to subtype versus subgroup interacts with trait dispersion and perceived outgroup homogeneity to affect stereotype change. The end product is anccount of how both the *structure* and *content* of stereotypes are acquired from experience.

A few words about our modeling approach are in order before proceeding. We have chosen to present a sequence of models rather than a single model that we use to simulate all the relevant empirical phenomena. This choice was guided by our goal of identifying minimal principles sufficient to explain a set of phenomena. These principles integrate coherently as we progress to more complex models, but we have chosen to present them in their simplest form in order to avoid post hoc (and possibly ambiguous) dissection of the more complex models into simpler constituents.

Why: The function of stereotypes

The question of why we have stereotypes has exercised social psychologists since the earliest studies (see Snyder and Miene, 1994, for a review). Why should the mind equip itself with representations that contribute to biases in social perception and prejudice in intergroup attitudes? This question is not likely to be resolved by any single answer, because stereotypes play a multifaceted

role in social cognition. A classic answer locates the function of stereotypes in their benefit to information processing economy: by simplifying the representation of individuals, fewer resource demands are made on the observer (Allport, 1954; Tajfel, 1969; Macrae et al., 1994; Bodenhausen and Lichtenstein, 1987). We will return to this idea later in connection with our own account.

Other answers locate the function of stereotypes in the maintenance and justification of social structures, such as social roles, power hierarchies, and coalitions: describing not only how certain groups are perceived but also prescribing how those groups *should* think, feel, and behave (Jost and Banaji, 1994; Koenig and Eagly, 2014; Cikara, 2021; Fiske and Neuberg, 1990). These structural functions can interact with cognitive functions; for example, more powerful individuals don't need to pay as much attention to less powerful individuals, and hence will rely more on stereotypes (Fiske, 1993). Reduced attention in turn reinforces the power hierarchy.

A third answer to the why question is also cognitive, but focuses on *statistical*, rather than information processing, constraints (Lee et al., 2013; McCauley et al., 1980). When we meet a new person, there is much that we do not know about them. However, we are not completely in the dark because unobserved traits may covary with observed traits. Thus, if we could learn the patterns of covariation, we could exploit them in the service of social inference. Stereotypes, on this view, correspond to beliefs about the covariation between a set of traits and a category or group label (a particular kind of trait). This view accords with several quantitative measures of stereotyping, such as the conditional probability that an individual has a particular trait given that they belong to a particular group (Brigham, 1971; Krueger, 1996), or the ratio between the conditional probability and the trait base rate (McCauley and Stitt, 1978). Empirical support for the statistical view comes from studies showing that observers rely more on covariation information when individuating information is absent, ambiguous, or uninformative (Krueger and Rothbart, 1988; Crawford et al., 2011; Nelson et al., 1990; Locksley et al., 1980). We will discuss other sources of empirical support in subsequent sections.

If stereotypes are to fulfill their statistical function, the group labels must be chosen such that traits can be effectively predicted. This implies that stereotypes should be accurate—a highly contentious proposition (Jussim et al., 2009). While studies have indicated that some stereotypes are

moderately accurate (e.g., Chan et al., 2012; Diekman et al., 2002; Rogers and Wood, 2010), they are more often not, at which point they merely serve to induce systematic biases, such as distortions of perceived covariation and homogeneity. The existence of such biases does not, however, imply that stereotypes make irrational use of evidence. In fact, an ideal observer *must* be biased (Gershman, 2021), but note that we do not mean "bias" in the colloquial sense of exhibiting a *preference*. Because the data available to an observer are typically insufficient to completely disambiguate all unobserved traits, statistically accurate trait inference requires an inductive bias that favors some inferences over others. This will inevitably produce systematic errors, despite reducing error on average. Although he did not employ the technical vocabulary for formalizing this idea, Allport (1954) recognized the necessity of inductive bias for statistical reasoning and prediction:

> Open-mindedness is considered to be a virtue. But, strictly speaking, it cannot occur. A new experience must be redacted into old categories. We cannot handle each event freshly in its own right. (p. 27)

We aim to place some mathematical flesh on the bones of Allport's insight. A complementary goal is to move beyond a reliance on explicit "categories" as the sole organizing unit of social structure.

In the next section, we introduce a simple Bayesian model that formalizes the statistical function of stereotypes. The purpose of this model is not to account for all aspects of stereotyping, but rather to establish the *prima facie* plausibility of the general approach, which we elaborate in subsequent sections to explain more complex aspects of stereotyping. The simple model will also serve a didactic purpose for those readers unfamiliar with Bayesian statistics. Note that the more complex models that we introduce later still retain the ability to explain the same phenomena captured by the simpler models.

How stereotypes form: Bayesian inference

We stated earlier that Bayesian inference can provide a rational analysis of stereotyping—an answer to the *why* question. We will begin by briefly spelling out what this means. A behavior or cognitive process is rational if we can describe it as the solution to an optimization problem.

Stereotypes arise in the setting where an observer is confronted with partial or ambiguous information about groups, from which they infer statistical properties of those groups. Thus, the optimization problem in this case concerns accurate statistical inference. Bayesian inference can be rationalized as the optimal solution to this problem (see for example Robert, 2007). Since below we focus on point estimation using the posterior mean (i.e., we summarize the posterior distribution with its expected value), we will offer a decision-theoretic justification for this estimator: it can be shown that the posterior mean minimizes the expected error when error is defined as the squared difference between the true and estimated parameter. In other words, we can show that this approach is *rational* in a statistical sense, but not necessarily *accurate* in the sense of always corresponding to the ground truth.

We now turn to the formal details of our model to explain *how* stereotypes are formed. An observer has access to information about N individuals, where the information about individual n is represented by a vector $\mathbf{x}_n = [x_{n1}, \ldots, x_{nD}]$ consisting of D trait values (e.g., kindness, industriousness, etc.). These vectors are collected into the $N \times D$ matrix $\mathbf{X}$, where each row corresponds to an individual.

In addition to this matrix, the observer has access to each individual's group membership (e.g., gender, nationality, etc.), represented by the binary vector $\mathbf{z}_n = [z_{n1}, \ldots, z_{nK}]$, where K is the number of groups. For now, we will assume that each individual belongs to a single group, such that $z_{nk} = 1$ if individual n belongs to group k, and $z_{nk} = 0$ otherwise (we will relax this constraint later). The group membership vectors are collected into the $N \times K$ matrix $\mathbf{Z}$, where (like the trait matrix) each row corresponds to an individual.

In our model, a stereotype for group k corresponds to the conditional distribution over traits given membership in group k, $P(\mathbf{x}_n | z_{nk} = 1)$: that is, the perceived distribution of traits associated with each group. This formalizes the widely cited definition given by Ashmore and Del Boca (1981): "A stereotype is a set of beliefs about the personal attributes of a social group" (p. 21).[2] It is important to distinguish between the objective conditional distribution (which the observer does not access directly) and the observer's subjective *beliefs* about the conditional distribution.

[2]See Greenwald and Banaji (1995) for a compendium of other definitions.

It is only the latter which defines a stereotype. The accuracy or inaccuracy of these beliefs about these groups in the world lies outside the bounds of this inquiry.

From the observer's point of view, the observed data are assumed to be generated from some parametric distribution $P(\mathbf{x}_n|z_{nk} = 1, \theta)$ with unobserved parameters θ. We will refer to this subjective conditional distribution as the *observation model*, which may differ from the objective conditional distribution. For example, an observer might assume that heights (x) are drawn from a gender-specific distribution with an unknown mean (θ_k, where k indexes gender).

In order to form stereotypic beliefs, an ideal observer needs to *marginalize* over uncertainty about the parameters:

$$P(\mathbf{x}_{n+1}|z_{nk} = 1) = \int_\theta P(\mathbf{x}_{n+1}|z_{nk} = 1, \theta)P(\theta|\mathbf{X}, \mathbf{Z})d\theta, \tag{1}$$

where now the goal is to predict the trait vector for a new individual after observering the traits for N previous individuals. Marginalization simply means summing (or integrating in this case) over different values of one variable (e.g., average height) to obtain the marginal distribution of another variable (e.g., a particular individual's height).

Intuitively, this equation says that the observer weights each possible stereotype by the *posterior probability* of its parameter value. This posterior probability is the observer's belief about the stereotype parameters given observations ($\mathbf{X}$, the person x trait matrix, and $\mathbf{Z}$ the person x group matrix). Continuing the example from the previous paragraph, an observer who wants to predict the height of a man (without knowing any other individuating information) would weight each possible average height (θ_{male}) by its posterior probability conditional on the set of observed heights and genders. Believing that men tend to be taller, the observer would then genearate a higher posterior probability for 5′ 9″ than 5′ 2″ for this new target. Of course, instead of height the feature could be competence, probability of shoplifting, and so on.

Bayes' rule stipulates how to compute the posterior probability:

$$P(\theta|\mathbf{X}, \mathbf{Z}) \propto P(\mathbf{X}|\mathbf{Z}, \theta)P(\theta), \tag{2}$$

where $P(\theta)$ is the *prior probability* assigned by the observer to θ. This is the formal definition of inductive bias (i.e., that new experiences must be redacted into old categories) referred to in the previous section. Bayes' rule says that the prior should be combined multiplicatively with the *likelihood* of θ (the probability of the data conditional on θ) and renormalized to obtain the posterior probability. Thus, the observer's belief about the average height for men depends both on the likelihood of each hypothetical average height (how consistent is each average height with the observed data) and the observer's prior beliefs about average height.

The Bayesian model sketched above can be understood as a formalization of the classic "bookkeeping" model proposed by Rothbart (1981), according to which all evidence, both stereotype confirming and disconfirming, is assimilated into the estimated trait distribution. Hewstone (1994) described the bookkeeping model as a "feature-frequency" model in the same class as similar models used to describe non-social category learning (e.g., Fried and Holyoak, 1984). The representation of a stereotype corresponds to the set of sufficient statistics (feature frequencies in the case of discrete traits) for the trait distribution. As we will show below, this kind of model has broad explanatory power. Nonetheless, research on subtyping and subgrouping has called into question some of its basic assumptions. Later in this paper we will introduce a structure learning framework that addresses the deficiencies of the bookkeeping model.

In summary, stereotype formation can be modeled as learning about the parameters governing the trait distribution for each group. We will now show that this model can account for several well-known aspects of stereotyping, including illusory correlation, accentuation, and outgroup homogeneity effects.

How the model accounts for illusory correlation

Illusory correlation refers to the phenomenon in which people perceive a relationship between infrequent behaviors or traits and infrequent classes of people where there is none. In the context of stereotyping, it is invoked to explain why negative traits or behaviors (which are relatively rare) get erroneously associated with minoritized groups (Hamilton and Gifford, 1976). How does the model account for this tendency?

Here we assume that the observation model is parametrized as a factorized Gaussian, where the traits are conditionally independent given the group membership:

$$P(\mathbf{x}_n|z_{nk}=1,\theta)=\prod_{d=1}^{D}\mathcal{N}(x_{nd};\mu_{kd},\sigma_{kd}^2), \tag{3}$$

with mean μ_{kd} and variance σ_{kd}^2 for group k and trait d. Initially, let us assume that there is a single trait ($D=1$) and that the variances are known; therefore the unobserved parameters are the means, $\theta=[\mu_1,\ldots,\mu_K]$, where we have dropped the trait index d for notational simplicity. If the prior over the means is Gaussian, $\mu_k\sim\mathcal{N}(m_k,\sigma_0^2)$, then the posterior after observing N_k individuals in group k is also Gaussian, with mean

$$\hat{\mu}_k=w_k\bar{x}_k+(1-w_k)m_k, \tag{4}$$

where $\bar{x}_k$ is the trait value averaged across individuals in the group, and

$$w_k=\frac{1}{1+\frac{\sigma_k^2}{N_k\sigma_0^2}} \tag{5}$$

is the weight attached to the observed data. Intuitively, the posterior mean is always somewhere between the average trait value and the prior mean for the group.[3]

Eq. 4 shows that when the number of observations is small, the posterior mean is pulled toward the mean of the prior m_k because the weight on the observed average trait value is smaller. Thus, if the mean of the prior is less than the average trait value ($m_k<\bar{x}_k$),[4] the posterior mean $\hat{\mu}_k$ will always be less than the average trait value. This bias diminishes with a larger sample size because more observations engender greater weight on the observed average trait value. This property is sufficient to reproduce several well-known aspects of stereotyping.

First, consider two groups (A and B) consisting of individuals with positive ($x_n>0$) and

[3] A derivation of this expression can be found in Murphy (2012).

[4] Because the prior mean is not typically measured in experiments, it is unclear how often this condition is satisfied. That said, researchers tend to use negative behaviors to demonstrate the effect in intergroup contexts. Negative behaviors are perceived, on average, as relatively rare compared to neutral and positive behaviors (Phillips and Cushman, 2017).

negative ($x_n < 0$) traits. Both groups have an equal proportion of positive and negative traits represented across individuals (hence average trait value $\bar{x}_A = \bar{x}_B$), but group A is larger ($N_A > N_B$). According to Eq. 4, as long as observed averaged traits $\bar{x}_A$ and $\bar{x}_B$ are positive (i.e., are greater than prior m_k), the "majority" group A will be *perceived* as more positive than the minority group B ($\hat{\mu}_A > \hat{\mu}_B$) because group B gets pulled toward 0 despite having identical underlying parameters. In other words, an illusory correlation will result. One important feature of this model is that it stipulates *when* illusory correlation should be most pronounced: at intermediate sample sizes, because when N_k is close to 0 the posterior mean will be dominated by the prior mean (which is the same for both groups), and when N_k gets very large the posterior mean will be dominated by the true mean (which is also the same for both groups). This prediction is consistent with the experimental results reported by Murphy et al. (2011), where absolute sample sizes were manipulated while holding the relative sample size fixed.

The model will also produce illusory correlation even when the sample sizes for the two groups are equal, provided the prior means are unequal (e.g., the prior associated with group A is less positive than the prior associated with group B). Because the posterior mean will in general be biased towards the prior mean (in the absence of a great deal of evidence), identical trait averages will produce non-identical posterior means. This prediction is consistent with experimental results reported by Hamilton and Rose (1980), where prior expectations were manipulated while holding sample sizes fixed (see Spears et al., 1987, for similar results).

We are not the first to discuss illusory correlations from a Bayesian perspective. Costello and Watts (2019) proposed a Beta-Binomial model (also known as Laplace's Rule of Succession) for contingency tables encoding, for example, the number of individuals with a particular trait in each group. The model outputs a posterior mean estimate of the probability that the trait will be observed in each group. Because the prior over these probabilities is uniform, the posterior mean is pulled towards $1/2$. This pull will be stronger for smaller groups (i.e., the minority group), thereby generating an illusory difference between groups. This account is conceptually very similar to ours, though it differs in the mathematical details.

Bott et al. (2021) criticized the model of Costello and Watts on several empirical and theoretical

grounds (some of which apply to our model as well). A complete discussion of these arguments would take us too far afield, but briefly they argue that the model may not necessarily be the optimal solution to the estimation problem, and that the model has trouble capturing some forms of illusory correlation. Bott and colleagues develop an alternative Bayesian model based on the *pseudocontingency heuristic* (Fiedler et al., 2009), according to which "If two things occur often then assume they are associated." Mathematically, this corresponds to estimating correlation based on marginal frequencies, rather than the full joint distribution.

It is not our goal here to consider in detail the relative merits of these different models, but rather to present some illustrative implications of a simple stereotyping model. We will progressively extend this model to accommodate other stereotyping phenomena. Our principal goal for this section was just to lay the foundation for a more sophisticated structure learning model that we present later.

How the model accounts for accentuation

Accentuation refers to the exaggeration of between-group differences in the judgment of individuals (Tajfel and Wilkes, 1963; Krueger, 1992; McGarty and Penny, 1988; McGarty and Turner, 1992). For example, in the classic demonstration by Tajfel and Wilkes (1963), lines were categorized into two groups (corresponding to short and long lines). Compared to conditions in which labels were randomly assigned or no labels were assigned at all, the lengths of lines from different categories were perceived as more different.

Within the Bayesian framework laid out above, accentuation arises from the regularizing effect of the prior: trait inferences for *individual* group members are biased towards their associated group stereotype, simultaneously pulling them away from the other group's stereotypes.[5] It follows essentially the same logic that we used to explain illusory correlation, but now applied to inferences about individuals rather than groups.

To model inferences about individuals, we assume that memory of an individual's trait vector

[5]Tajfel's pioneering study of accentuation can be understood as a form of *categorical perception* (Harnad, 1987). Our Bayesian explanation of accentuation follows the same line of reasoning that has been applied to other categorical perception phenomena (Feldman et al., 2009).

is corrupted by Gaussian noise, yielding a noisy memory trace $\tilde{x}_n \sim \mathcal{N}(x, \tau^2)$, where τ is the perceptual noise standard deviation (for simplicity we continue to work with 1-dimensional trait vectors, but the theory generalizes straightforwardly to multiple dimensions). This leads to an expression for the individual posterior mean similar to Eq. 4:

$$\hat{x}_n = b_k \tilde{x}_n + (1 - b_k)\mu_k, \tag{6}$$

where k refers to the group membership of individual n, and

$$b_k = \frac{1}{1 + \frac{\tau^2}{\sigma_k^2}} \tag{7}$$

is the weight attached to the memory trace for that individual, a monotonically decreasing function of the ratio between the memory noise variance τ and within-group trait variance σ_k.[6] Here again the posterior mean is somewhere in between the sample mean and the prior mean, but in this case the sample is drawn from memory.

One implication of this model is that memory recall of an individual's traits should be biased towards the group mean μ_k, thereby accentuating between-group differences. The bias should strengthen when memory is more unreliable (that is, when the weight attached to the memory trace for that individual is low). For example, accentuation is stronger in the Tajfel-Wilkes paradigm when the units of line lengths are unfamiliar (i.e., Belgian participants estimating in inches; Corneille et al., 2002), possibly because these units are more easily confusable in memory. A second implication of this model is that greater dispersion of traits within the group (higher σ_k) should reduce the bias towards the group mean. Indirect evidence comes from a study of racial stereotypes (Ryan et al., 1996), which found that high dispersion groups were perceived as less stereotypic (the degree to which a group is perceived to conform to the group stereotype), and stereotypicality predicted the bias towards the group stereotype in judgments of individuals.[7] A

[6] Note that we have assumed here direct access to the group mean and variance, though these could be estimated as in the previous section.

[7] The results of this study are somewhat hard to interpret because perceptions of stereotypicality and dispersion were negatively correlated, raising the question of whether these are truly measuring different constructs. When they were simultaneously used as predictors of individual trait judgments, stereotypicality (but not dispersion) predicted

third implication, untested as far as we know, is that increasing memory noise (e.g., by increasing cognitive load during encoding or test, or lengthening the retention interval) should strength the bias towards the group mean.

Eq. 6 assumes that the memory for each individual's group membership is perfect, but evidence suggests that memory errors also occur for group membership, and that these errors dilute accentuation effects (Krueger and Rothbart, 1990). To capture unreliability of memory for group membership, we assume that people reconstruct membership from the noisy memory trace. Assuming a uniform prior over groups, the posterior is given by:

$$P(z_{nk} = 1|\tilde{x}_n) \propto P(\tilde{x}_n|z_{nk} = 1) = \mathcal{N}(\tilde{x}_n; \mu_k, \sigma_k^2 + \tau^2). \tag{8}$$

The posterior mean is then obtained by marginalizing over group membership:

$$\hat{x}_n = \sum_k P(z_{nk} = 1|\tilde{x}_n)[b_k\tilde{x}_n + (1 - b_k)\mu_k]. \tag{9}$$

Recall that marginalization simply means summing over different values of one variable (an individual's group membership in this case) to obtain the marginal distribution of another variable (an individual's trait value).

One implication of Eq. 8 is that group membership errors during memory retrieval will be more likely for individuals in regions of overlap between stereotypes (i.e., near category boundaries), and therefore (by Eq. 9) accentuation effects will be weaker in these regions. Consistent with this hypothesis, Krueger and Rothbart (1990) showed that increasing the variance of the trait distribution resulted in more memory errors, and that excluding miscategorized individuals resulted in stronger accentuation effects. A second implication is that increasing memory noise (τ) should weaken accentuation by blurring the memories of individual group membership. This prediction has not been tested as far as we know.

bias towards the stereotype. However, the collinearity of the predictors means that some null effects could be type II errors.

How the model accounts for the outgroup homogeneity effect

Yet another phenomenon widely documented in the stereotyping literature is the *outgroup homogeneity effect*. Outgroups tend to be viewed as more homogeneous than ingroups, even when their true variances are identical (Linville et al., 1989; Ostrom and Sedikides, 1992; Judd and Park, 1988; Quattrone and Jones, 1980; Park and Rothbart, 1982). Importantly, evidence suggests that this effect is not mediated by the ingroup-outgroup distinction *per se*, but rather by the differential amount of information about each group available to subjects (who necessarily must belong to either the ingroup or the outgroup). When the ingroup is smaller than the outgroup, the outgroup heterogeneity effect is reversed, with the ingroup now being perceived as more homogeneous (Simon and Brown, 1987; Simon and Pettigrew, 1990; Mullen and Hu, 1989).

So far, we have assumed that the variances of the observation model were known; we now extend the model to inferences about variances.[8] A simple form for the posterior mean can be obtained if we assume a Jeffreys prior over variance, according to which $P(\sigma_k^2) \propto 1/\sigma_k^2$:

$$\hat{\sigma}_k^2 = \frac{s_k^2}{1 + 2/N_k}, \tag{10}$$

where $s_k^2 = \frac{1}{N_k} \sum_n z_{nk}(x_n - \mu_k)^2$ is the sample variance. A key feature of this model is that the estimated variance is shrunk when the sample size is small. Specifically, for a single sample ($N_k = 1$), the estimated variance is $1/3$ of the sample variance. As N_k grows, the estimated variance becomes progressively closer to the sample variance.

This is consistent with the evidence that perception of variability tracks sample size. Eq. 10 just formalizes this idea: the same sample variance translates to larger or smaller estimated variance depending on the group size. Eq. 10 also makes a stronger and novel quantitative prediction that the effect of sample variance on estimated variance increases with the sample size; thus, distortions are greatest for small sample sizes.

Our account of the outgroup homogeneity effect is similar in spirit to the model developed by Linville et al. (1989), which assumed that people estimate group dispersion using the uncorrected

[8] For simplicity, we will assume the means to be known, but it is straightforward to analyze the case where both means and variances are unknown.

sample variance. The sample variance estimator is biased downwards, but the bias diminishes with larger samples, thereby producing both the observed underestimation of dispersion and its characteristic dependence on sample size. One difference between the sample variance account and the Bayesian account developed here is that in principle the Bayesian account (but not the sample variance account) can incorporate prior beliefs about homogeneity. Some evidence suggests that people have beliefs about homogeneity even before observing samples from a particular group (Wilder, 1984b).

Both the sample variance and Bayesian accounts locate the origin of the effect in information processing: even when the sample variance is equated between groups, differences in sample size will produce differences in perceived dispersion due to the nature of the mental estimator. Konovalova and Le Mens (2020) have developed a conceptually different statistical explanation which locates the origin in sampling biases. The key premise is that the effect arises for natural groups due to the fact that the sample variance is *not* equated between groups. If people tend to preferentially encounter ingroup members, then surprisingly even an unbiased estimator of sample variance will yield an outgroup homogeneity effect, assuming the dependent measure is the probability that the ingroup has higher variance than the out-group. This arises due to the skewed sampling distribution of variance. One disadvantage of this account is that the effect only arises when making probability judgments about ordinal relations between groups, even though most studies measure estimates of perceived dispersion separately for each group. In any case, biased sampling and biased estimation accounts are not mutually exclusive, as pointed out by Konovalova and Le Mens (2020).

As we will discuss later, inferences about variability are more complex than the picture developed above. For example, beliefs about subgroups play an important role in determining inferences about variability (Park et al., 1992; Kraus et al., 1993). A complete account will thus require us to explain the structure and origin of subgroups. We now develop such an account using principles of structure learning.

When and how: A structure learning model of stereotype change

The models developed thus far concern parameter learning with known groups: how do we build our representation of the trait distribution within a group? The parameters governing each distribution represent the content of the corresponding group stereotype. What such models do not tell us is where the groups come from in the first place. They do not solve the structure learning problem.

One answer to the structure learning problem is that groups consist of individuals who share a common group identity or category label. This is consistent with the modeling of the previous section, where we assumed that individuals have access to category labels of all individuals (except in the context of memory-based judgments about individuals, where they have to reconstruct those labels from information stored in memory). There are three problems with this answer. First, it doesn't explain explain subtyping: why do we mentally segregate deviants even when they share a category label? We could, in theory, just assimilate deviants to the category and update the stereotype accordingly. Second, it doesn't explain subgrouping: why do we mentally construct hierarchically organized sets of groups when all the members share the same superordinate category label? Third, it doesn't explain how people reason about groups in the absence of explicit category labels. For example, when security officers patrol college campuses they generate guesses about which people are students versus professors versus staff versus unauthorized visitors, in the absence of explicit labels or markers of campus-affiliation. We argue that, irrespective of the (in)accuracy of their guesses, they accomplish this by inferring *latent groups* on the basis of observable features.

To realize this idea computationally, we extend the Bayesian model of stereotyping to incorporate uncertainty about the group membership matrix $\mathbf{Z}$. We will proceed in two steps, again building from simple to complex models. First, we will analyze the case where each individual can only belong to a single group. This will provide an account of subtyping. Then we will analyze the case where each individual can belong to multiple, hierarchically organized groups. This will provide an account of subgrouping.

Group discovery and assignment

Before we lay out the explanation for subtyping, we develop a model for how people may be assigned to groups in the first place (though note our explanation of subtyping can easily be applied to cases where groups are known because structure learning models can take explicit category labels into account). The *latent groups model* follows closely the setup of previous sections, but now the posterior is defined over both parameters (stereotype content or traits) and group membership (stereotype structure):

$$P(\mathbf{Z}, \theta|\mathbf{X}) \propto P(\mathbf{X}|\mathbf{Z}, \theta)P(\theta)P(\mathbf{Z}), \tag{11}$$

where we have now introduced a prior over the membership matrix $\mathbf{Z}$. Recall that $z_{nk} = 1$ if individual n belongs to group k, and is 0 otherwise. In this section, we assume that each individual belongs to a single group, and hence each row contains a single 1. Because the number of groups is unknown *a priori*, we define a prior over membership matrices with an unbounded number of columns. A standard nonparametric prior for such matrices is the *Chinese restaurant process* (Aldous, 1985; Gershman and Blei, 2012):

$$P(z_{nk} = 1) = \begin{cases} \frac{g_{nk}}{n-1+\alpha}, & k \leq K_n \\ \frac{\alpha}{n-1+\alpha}, & k = K_n + 1, \end{cases} \tag{12}$$

where $g_{nk} = \sum_{j=1}^{n-1} z_{jk}$ is the number of individuals assigned to group k prior to n (i.e., the state of the process prior to n), K_n is the number of unique groups created prior to n, and $\alpha \geq 0$ is a concentration parameter that probabilistically controls the number of groups. When $\alpha = 0$, all individuals are assigned to the same group; in the limit $\alpha \to \infty$, all individuals are assigned to their own group. The expected number of groups $\mathbb{E}[K_n]$ scales according to $\alpha \log n$.

Variants of this model have been widely applied in cognitive science (see Austerweil et al., 2015, for a review). Most relevant for present purposes is the model developed by Spicer and Sanborn (2017), which used essentially the same model to explain certain aspects of subtyping. We expand on this idea below. Also closely related is the model presented in Gershman et al.

(2017b) to analyze patterns of social influence. According to that model, social influence between individuals is stronger to the extent that they believe they belong to the same group. Membership is inferred on the basis of observed choices. We subsequently applied this idea to real-world political attitudes (Lau et al., 2018) and used it to understand the neural correlates of social influence (Lau et al., 2020). In the same vein, we assume that observers use trait data (which may include an individual's preferences) to infer latent group membership. They then use these inferences about group membership to structure their inferences about the group trait distribution (i.e., the stereotype), following the logic of the previous sections.

To obtain the posterior over group membership, we marginalize over the parameters θ:

$$\begin{aligned} P(\mathbf{Z}|\mathbf{X}) &= \int_{\theta} P(\mathbf{Z}, \theta|\mathbf{X}) d\theta \\ &\propto \mathcal{N}(\mathbf{X}; \boldsymbol{\mu}, \sigma_0^2 \mathbf{Z}\mathbf{Z}^\top + \sigma^2 \mathbf{I}) P(\mathbf{Z}), \end{aligned} \tag{13}$$

where we have assumed for simplicity that all groups share a common prior ($\boldsymbol{\mu} = \mathbf{0}$ in our applications) and that the prior and likelihood covariances are isotropic (i.e., different traits are uncorrelated with each other conditional on group membership).[9] We have also assumed here that the variances are known, though it is possible to extend the model to inferences about variances, as we did above.

Finding the group assignment with highest posterior probabilities requires an intractable search in the space of group assignments. In practice, most of these assignments have negligible probability. We therefore enumerate a small set of plausible group assignments that have relatively high probability and score these using Eq. 13. The plausible set was chosen manually for each simulation (adding more assignments to the plausible set did not typically change the results dramatically). To obtain parameter estimates, we marginalize over this plausible set (denoted by $\mathcal{Z}$):

$$P(\theta|\mathbf{X}) = \sum_{\mathbf{Z} \in \mathcal{Z}} P(\mathbf{Z}, \theta|\mathbf{X}). \tag{14}$$

[9]The isotropic covariance assumption is unlikely to be true in general, and is not mathematically necessary, but makes the model presentation simpler and is sufficient to account for the phenomena we address here.

Because we are interested in qualitative correspondences between the model and empirical data, we do not fit the parameters of the model. Instead, we fix the parameters across all simulations except those where they are explicitly manipulated. This allows us to demonstrate the model's breadth of explanatory power without parameter tuning. We used the following parameter values, excepted where noted otherwise: $\sigma_0^2 = 2, \sigma^2 = 0.1, \alpha = 10$. The qualitative results were not highly sensitive to these parameter values, although extreme values of these parameters will (as expected) change the results qualitatively.

When and how do people subtype targets?

One of the most important and well-documented findings in the subtyping literature is the effect of deviance dispersion on stereotype change: stereotype change is greater when stereotype-inconsistent information is dispersed across multiple individuals, compared to when it is concentrated in a single individual (Weber and Crocker, 1983; Johnston and Hewstone, 1992; Hewstone and Hamberger, 2000; Hantzi, 1995; Johnston et al., 1994). This finding has traditionally been interpreted as evidence that strongly deviant individuals in the concentrated condition are mentally segregated into subtypes, which allows the existing stereotype associated with the rest of the group to remain intact. By contrast, weakly deviant individuals in the dispersed condition are assimilated into the group, driving change in the stereotype. Here we present simulations based on the latent groups model to demonstrate when and how people subtype targets.

The latent groups model reproduces the structure that is hypothesized to underlie the dispersion effect on stereotype change (Figure 1). Each simulation took as input data from either 6 or 30 individuals, each with 2 binary traits (other simulations reported in this paper used the same setup except for specific changes detailed below). We included two individuals each with one counter-stereotypical trait to model the dispersed condition, or a single individual with two counter-stereotypical traits to model the concentrated condition. In the dispersed condition, the model places more probability on a single latent group for all individuals (including the deviants) compared to in the concentrated condition, where it favors segregating the individual deviant into a separate group.

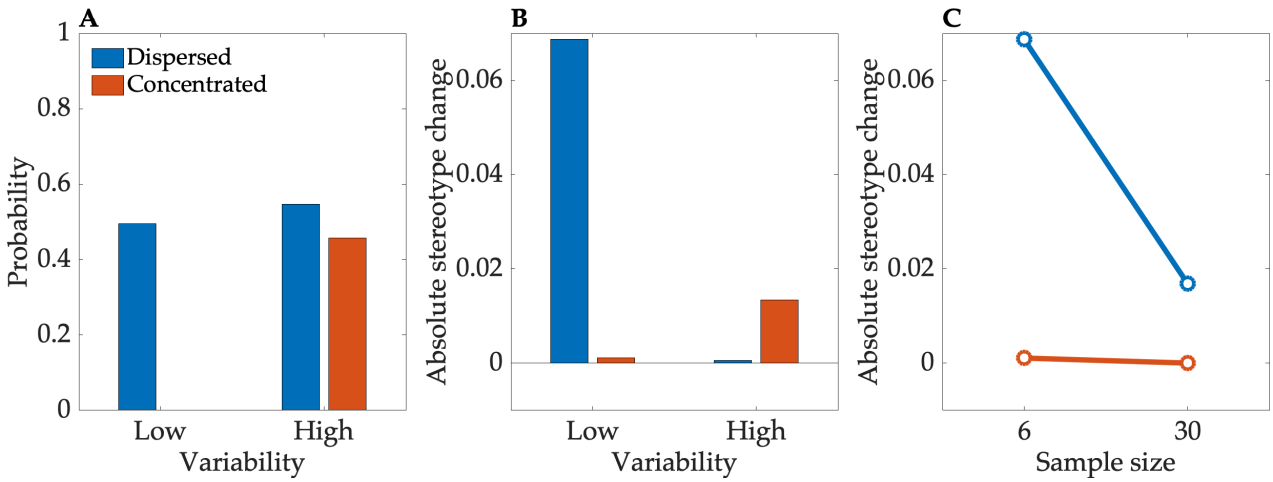

Figure 1: Simulation of dispersion, variability, and sample size effects on subtyping. (A) Probability that a deviant individual is assigned to the group as a function of whether perceived variability is high or low and whether the counter-stereotypical trait are concentrated in a single individual or dispersed across multiple individuals. (B) Degree of change in the prediction of trait values for the group after observing the deviant. (C) Degree of change as a function of sample size and dispersion.

Importantly, the model also identifies and explains several moderating factors of the dispersion effect. First, the dispersion effect is much stronger when variability is low (Hewstone and Hamberger, 2000); see panel A of Figure 1. We manipulated dispersion by stretching out the range of attribute values in the observed population. The interaction between dispersion and variability occurs in the model because high variability makes the stereotype more "tolerant" of the deviant, even in the concentrated condition, thereby diluting the contrast between dispersed and concentrated conditions.

We now turn to stereotype change, which we measure as the change in the inferred average trait for a group before vs. after a set of observations. As predicted, we observe the most stereotype change in the low-variability/dispersed condition (Figure 1, panel B). Second, the dispersion effect is weaker when the sample size is larger (Weber and Crocker, 1983); see panel C of Figure 1. In the model, a larger sample size increases the probability that the deviants will be subtyped even in the dispersed condition, because there is greater certainty about the trait distribution and hence the stereotype is less tolerant of the deviants.

Another determinant of subtyping is deviants' typicality; individuals who are counter-stereotypical

on one dimension but are otherwise more typical of their group produce more stereotype change than more atypical individuals (Hewstone and Hamberger, 2000; Hewstone et al., 2000; Johnston and Hewstone, 1992; Hantzi, 1995; Rothbart and John, 1985; Wilder, 1984a; Weber and Crocker, 1983). For example, white, middle-class, high-earning, but introverted lawyers bring about more lawyer-related stereotype change (i.e., the expectation that lawyers are extroverted is weakened) compared to introverted black lawyers (because black lawyers are less prototypical of the category 'lawyers') (Weber and Crocker, 1983). There is some controversy around whether typicality mediates stereotype change or is simply correlated with it (see Richards and Hewstone, 2001, for discussion). For our purposes, the important thing to note is that according to our model, higher typicality makes it more likely that a deviant will be assimilated into the group and hence drive stereotype change. In fact, even a "neutral" (i.e., unrelated) trait can contribute to perceived atypicality (Kunda and Oleson, 1995). In one of their studies, Kunda and Oleson presented subjects with information about an introverted lawyer. One set of subjects was given additional information (that the lawyer worked for a large or small firm) which separate pretests had established to be neutral with respect to introversion. Subjects who received no additional information updated their stereotypes about lawyers more (i.e., they reported viewing lawyers in general as being higher in introversion) compared to subjects who also received the neutral information. Our model explains this finding in terms of the same mechanism producing typicality effects in subtyping: anything that makes an individual unique relative to the group also reduces the probability of their membership, "fencing" them off, thereby suppressing stereotype change. Figure 2 presents the results of simulations that capture both findings. We manipulated typicality by changing the proportion of individuals with a particular trait value (either half of the individuals or all of them). Probability of group membership increases with a typical trait and decreases with a neutral trait, relative to no trait. Stereotype change tracks these differences in probability.

Kunda and Oleson (1997) investigated the impact of deviance magnitude (along a single trait) on stereotype change, finding that moderate deviants produced more stereotype change than extreme deviants (see also Dannals and Miller, 2017). Our model recapitulates this finding (Figure

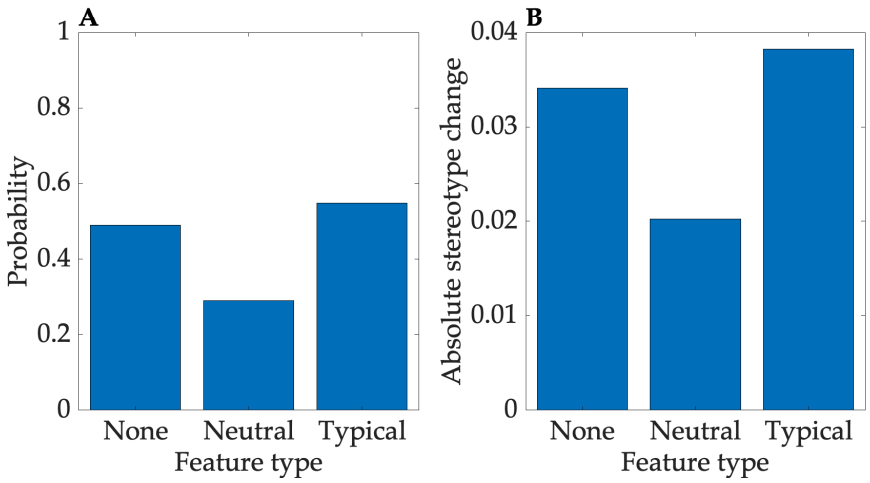

Figure 2: Simulation of typicality effects on subtyping. (A) Probability that a deviant individual is assigned to the group as a function of whether the deviant trait type is absent, neutral, or typical. (B) Degree of change in the prediction of trait values for the group after observing the deviant.

3).[10] We manipulated deviance magnitude by setting one of the traits to a range of values varying between -1 and 1. The probability of assimilation into the group decreases monotonically with deviance magnitude, but this has a non-monotonic effect on stereotype change. At very low deviance, membership probability is high but since deviance is also small there is little change in the stereotype. At very high deviance, membership probability is low, weakening the impact of the deviant. Stereotype change is maximized in between these two extremes, at moderate deviance magnitudes. This explanation mirrors structure learning accounts that have been proposed in other domains (Gershman et al., 2013, 2017a).

Finally, we explore the effects of cognitive load. Several studies have found reduced subtyping, and therefore greater stereotype change in the direction of the disconfirming individual, under cognitive load (Moreno and Bodenhausen, 1999; Yzerbyt et al., 1999). These studies were motivated by the idea that stereotypes serve an economizing function by simplifying mental computation (Allport, 1954; Tajfel, 1969; Macrae et al., 1994; Bodenhausen and Lichtenstein, 1987). The effect of load on subtyping was interpreted as evidence that constructing richer representations of social groups demands greater cognitive resources. Our model sheds light on the mechanistic

[10]Note that if deviance is measured by absolute distance, a large deviance along one trait is equivalent to the same degree of deviance divided across multiple traits. We are not aware of studies directly examining this distinction.

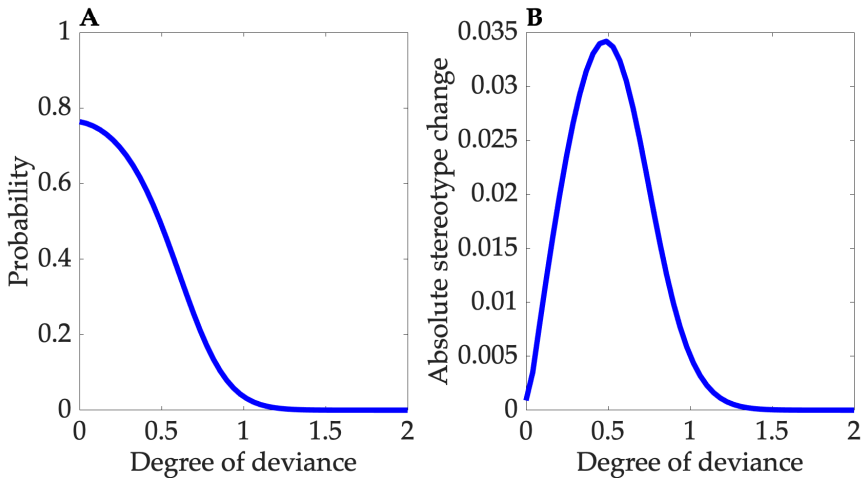

Figure 3: Simulation of deviance effect on subtyping. (A) Probability that a deviant individual is assigned to the group as a function of deviance magnitude. (B) Degree of change in the prediction of trait values for the group after observing the deviant.

nature of the economizing function of stereotypes. In particular, fewer latent groups place smaller demands on mental computation because they require the observer to marginalize over a smaller number of hypotheses. We can operationalize cognitive demand in terms of the concentration parameter α, which controls the expected complexity of stereotype representations (Figure 4). Indeed, we see higher probability that the deviant is assigned to the group, and therefore greater stereotype change, at lower levels of α. Recently, Dasgupta and Griffiths (2022) have suggested a specific link between the concentration parameter and cognitive demand measured information-theoretically. The quantitative predictions of this model for stereotyping remain untested.

When and how do people subgroup?

We now describe a generalization of the latent groups model to hierarchically structured groups, which we represent as a tree structure (Figure 5). Each node in the tree picks out a collection of individuals, defining a subgroup relative to the superordinate node. In other words, a subgroup is a subset of individuals drawn from a larger group. A subtype is a special case of a subgroup consisting of a single individual (an outlier) that is forked off from a group without sharing its statistical properties (i.e., it is not subordinate to the group in the tree structure). Each individual

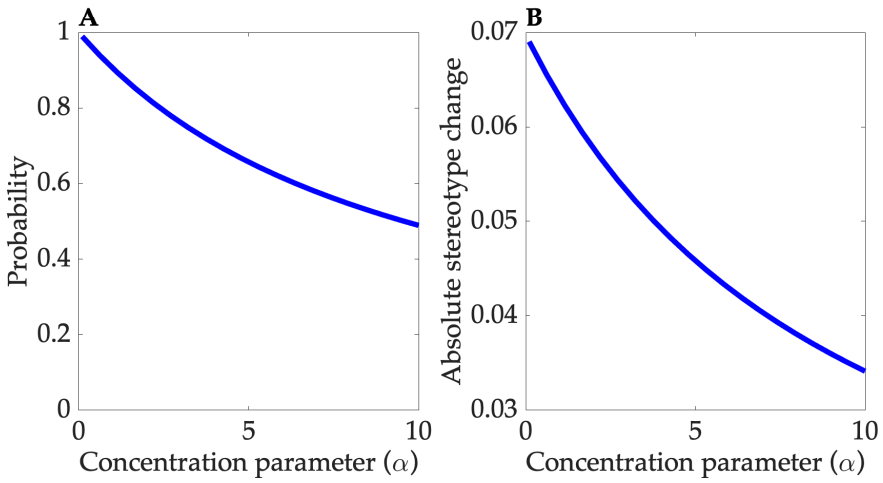

Figure 4: Simulation of cognitive load effects on subtyping. (A) Probability that a deviant individual is assigned to the group as a function of the concentration parameter α. Larger values of alpha tend to produce more complex group structures and hence hypothetically demand more cognitive resources. Cognitive load is hypothesized to push the concentration parameter lower, thereby increasing the probability of assigning the deviant to the group. (B) Degree of change in the prediction of trait values for the group after observing the deviant. If the concentration parameter decreases under cognitive load, then stereotype change will be greater.

is identified by a path through the tree. For example, the root node could correspond to "all people," and below it are nodes corresponding to "men" and "women." Below each of these are nodes corresponding to "old" and "young." An old man would then correspond to the path people→men→old.

Each node indexes a trait distribution, just as in our treatment of flat clusters in the previous section. Let pa(k) denote the parent of node k in the tree structure. The mean for group k is sampled from a Gaussian with mean $\mu_{\text{pa}(k)}$ and covariance $\sigma_0^2\mathbf{I}$. Thus, subgroups will tend to share traits with their superordinate groups, and this similarity will decrease as a function of distance in the tree. We will assume that there is a root node in the tree $(k = 0)$ that defines the prior for any groups that are not subgroups, with mean $m_k = \mathbf{0}$ and covariance $\sigma_0^2\mathbf{I}$. Unlike in the flat clustering model, where each row of $\mathbf{Z}$ has a single 1 (corresponding to the group label), in the hierarchical clustering model each row contains multiple 1s (corresponding to the group labels at each level of the hierarchy).

It will be convenient for us to describe the inferences in terms of the *relative* group-level mean,

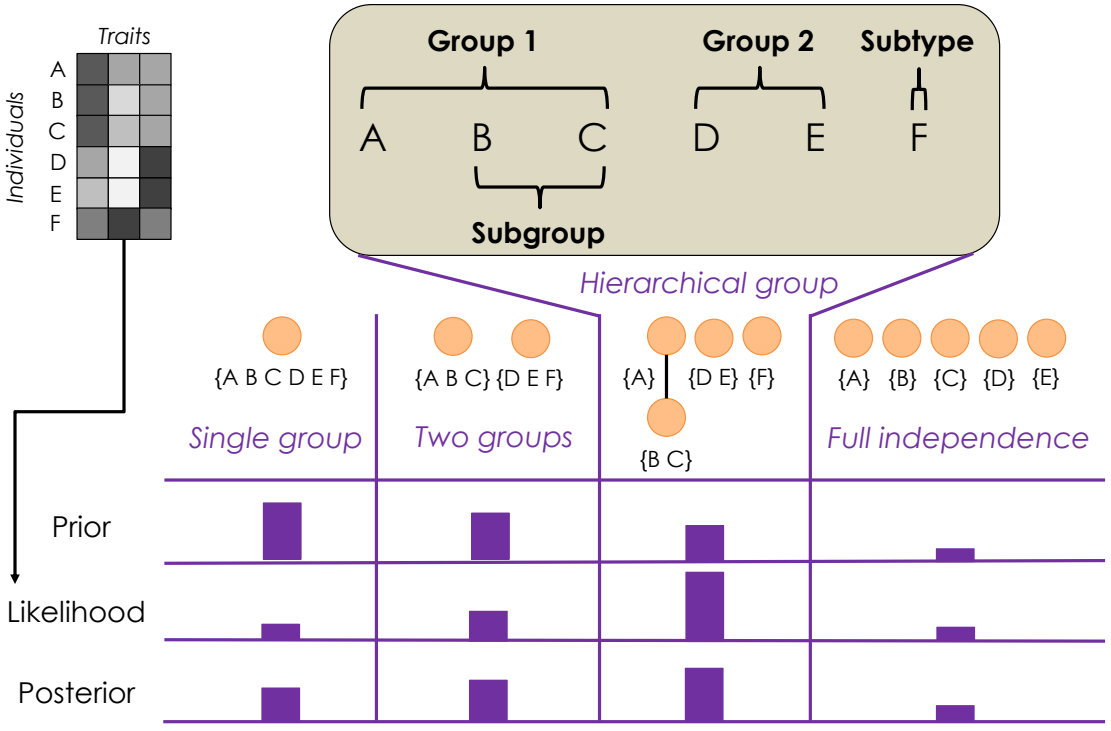

Figure 5: **Schematic of the hierarchical latent groups model.** (Top left) The observation matrix, where grayscale values denote the magnitude of a trait for a given individual. (Top right) Illustration of a grouping structure, where one subgroup is nested within a superordinate group, and one individual is fenced off as a subtype. (Bottom) Application of Bayes' rule to structure learning. The posterior distribution over structures is obtained by multiplying the prior probability of each structure with the likelihood of the observations under that structure and then renormalizing. The height of each bar denotes the probability of the corresponding structural hypothesis, shown as a set of trees. Some of these trees are "degenerate" (only a single level deep), in which case they correspond to flat (non-hierarchical) groups. The individuals assigned to each node are shown in brackets. Note that although the prior on the hierarchically organized group is low, the likelihood given the data is highest, giving rise to a posterior that favors the nested group structure.

$\tilde{\mu}_k = \mu_k - \mu_{\text{pa}(k)}$ (i.e., how much more or less of each trait is a subgroup associated with relative to its superordinate group). For a given tree structure, the posterior expectation of the relative mean for group k is given by:

$$\mathbb{E}[\tilde{\mu}|\mathbf{X}, \mathbf{Z}] = \left(\mathbf{Z}^\top \mathbf{Z} + \frac{\sigma^2}{\sigma_0^2}\mathbf{I}\right)^{-1} \mathbf{Z}^\top \mathbf{X}. \tag{15}$$

The group-level means μ_k can be obtained by summing the relative means along the path in the tree leading to group k.

We define a prior distribution over trees known as the *nested Chinese restaurant process* (nCRP; Blei et al., 2010), which has previously been applied to modeling hierarchical category learning (Canini and Griffiths, 2011) and perception (Gershman et al., 2016). The basic building block is the Chinese restaurant process, defined in Eq. 12. In the nCRP, this process recurses to some depth d_n, so that each individual is assigned to d_n groups, defining a path through the tree. We place a uniform distribution over depths up to 2 (in principle, there is nothing stopping us from considering deeper depths, but this was not necessary for our simulations).

In the first study to directly compare subtyping and subgrouping, Maurer et al. (1995) showed that subjects given subtyping instructions (distinguishing stereotype-consistent individuals from stereotype-inconsistent individuals) viewed the group as more stereotypical and homogeneous compared to subjects given subgrouping instructions (sorting individuals into multiple groups based on their similarities and differences). Subtyping instructions also led subjects to perceive a greater difference in typicality between confirming and disconfirming individuals. In a subsequent study using the Weber and Crocker (1983) dispersed vs. concentrated manipulation, Hewstone and Hamberger (2000) showed that subgrouping instructions eliminated the difference between dispersed and concentrated conditions, instead showing increased stereotype change in both conditions. Figure 6 shows a simulation of this finding, which the model captures by allowing deviants to be assimilated into subgroups (thereby leading to change in the superordinate stereotype) rather than segregated into subtypes (which prevents superordinate stereotype change).[11]

[11] This prediction will in general depend on the value of α, though it holds for a range of values.

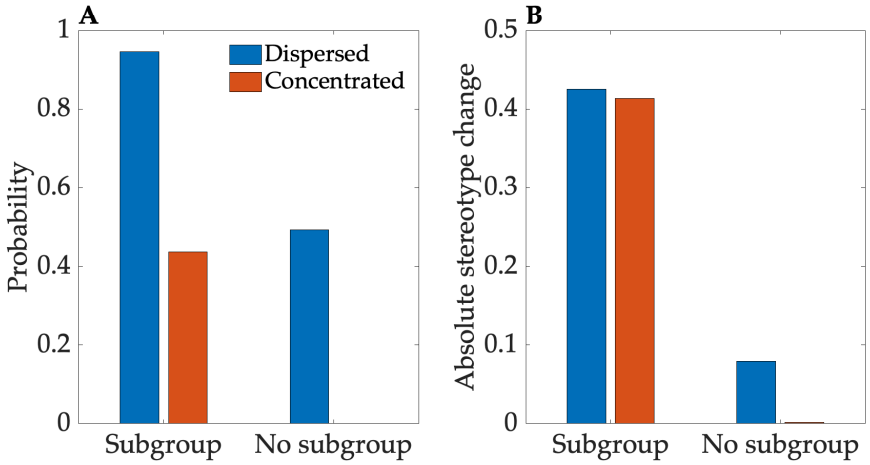

Figure 6: Simulation of dispersion effects on subgrouping. Each simulation took as input data from 10 individuals, each with 2 binary traits. As in the earlier simulation of dispersion, we included two individuals each with one counter-stereotypical trait to model the dispersed condition, or a single individual with two counter-stereotypical traits to model the concentrated condition. (A) Probability that a deviant individual is assigned to the group as a function of whether subgroups were allowed by the model during the categorization of group members. (B) Degree of change in the prediction of trait values for the superordinate group after observing the deviant.

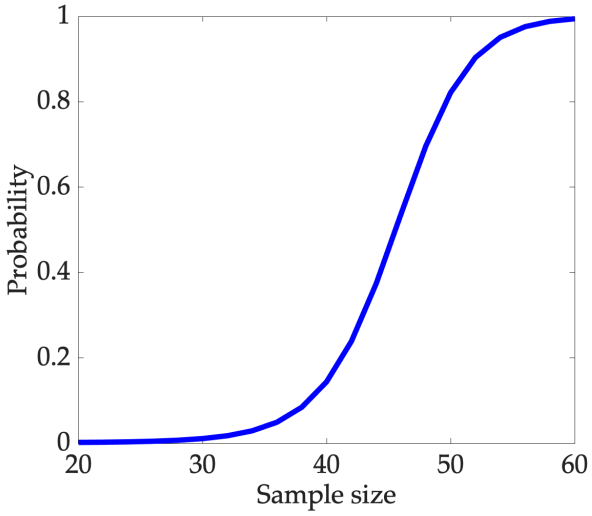

Figure 7: Probability that a deviant individual is assigned to the group as a function of sample size.

The study of subgrouping led to an important insight into the origin of the outgroup homogeneity effect. Park et al. (1992) showed that perceived variability is closely linked to the number of subgroups that subjects generate, and that more subgroups were generated for ingroups compared to outgroups (see also Kraus et al., 1993). Indeed, the outgroup homogeneity effect was eliminated entirely when the difference in number of generated subgroups was controlled for. If we start from the assumption that a key difference between ingroups and outgroups is sample size (people are exposed to more individuals in the ingroup than in the outgroup), we can ask how sample size contributes to the formation of subgroups. Our model provides an answer to this question: the number of subgroups increases with sample size. A corollary of this property is that the probability of assigning a new individual to the superordinate group increases with sample size (Figure 7), because as subgroups proliferate it becomes increasingly easier to assimilate a new individual into one of them. The expanding diversity of the superordinate group renders it more tolerant of deviants.

Our model makes a novel prediction about grouping as a function of deviance. The simulation shown in Figure 8 exhibits three regimes. For small levels of deviance, the deviant is assimilated into the group. For moderate levels of deviance, a deviant is assigned to a subgroup. Only for large levels of deviance is the deviant assigned to a subtype. Intuitively, small levels of deviance are expected within the normal range of variation for a group. If the deviance is too large to be accomodated by this normal range, the model tries to accommodate it by capturing the "residual variation" in a subgroup. This allows the deviant to share statistical properties with the superordinate group while also deviating from it. When the deviance is sufficiently large, this doesn't work anymore; it is more plausible to place the deviant in an entirely separate group (i.e., a subtype). This measure could be tested by directly assaying inferences about subgroups and subtypes while manipulating deviance.

General discussion

We have presented a sequence of increasingly sophisticated stereotyping models, culminating in an account of hierarchically organized latent groups. This model captures many important phe-

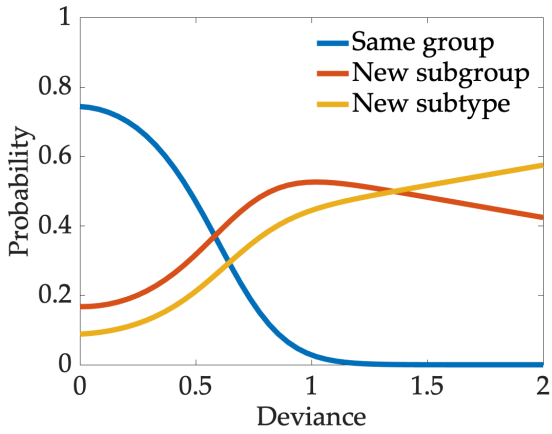

Figure 8: Probability of different grouping structures under different levels of deviance. In this simulation, 10 individuals were given a feature value of 1, and then an 11th individual was given a feature value of 1-d, where d is the deviance magnitude.

nomena in the literature on stereotypes and stereotype change: illusory correlation, accentuation, outgroup homogeneity, subtyping, and subgrouping, as well as the effects of moderating factors such as sample size, variability, and cognitive load. The latter two phenomena are explained by the model as the result of structure learning mechanisms, a distinctive feature of our account. Subtyping arises from the segregation of deviant individuals from the group, preventing them from driving stereotype change. This is a formalization of "refencing" as proposed by Allport (1954) and many subsequent researchers. Subgrouping arises from the assimilation of deviant individuals into a subordinate category that shares some features with the superordinate category. This allows deviants to drive stereotype change, because the subordinate category exerts a pull on the overall stereotype representation.

Comparison to other models

Early models of stereotyping can be categorized into one of two classes: exemplar models and abstraction models (see Linville and Fischer, 1993, for a review). Exemplar models take a similarity function defined over individual trait vectors as their basic primitive (Linville et al., 1989; Fiedler, 1996; Smith and Zárate, 1992). Inductive reasoning and generalization operate through similarity computation. For example, predicting whether an individual has a particular trait is computed by

generalizing from other similar individuals. Likewise, inferences at the group level are computed by generalizing from individuals that belong to a particular group (i.e., the group identity enters into the similarity computation). Abstraction models, by contrast, typically take trait distributions as their basic primitive, abstracting over individual exemplars (Kraus et al., 1993; Park and Judd, 1990; Nisbett and Kunda, 1985; Bordalo et al., 2016).

A second wave of stereotyping models adopted the connectionist framework, where inductive reasoning and generalization emerge from the dynamics of spreading activation between elementary processing units. In some connectionist models (Van Rooy et al., 2003; Vanhoomissen and Van Overwalle, 2010; Kunda and Thagard, 1996) the units are "localist" in the sense that they represent distinct symbolic variables (e.g., particular traits), while in other models the representation of variables is distributed across multiple units (Smith and Decoster, 1998; Queller and Smith, 2002; Kashima et al., 2000), so that the collection of traits for an individual is represented algebraically as a matrix. All of these models have in common the property that learning is driven by some form of mutual constraint satisfaction, although the constraints, learning rules, and dynamics differ between the models.

All of the connectionist models also have in common the property that memory for individual exemplars is discarded, in contrast to exemplar models in which the trait vectors for all exemplars are stored in memory (possibly corrupted by noise). So in that respect the connectionist models are similar to abstraction models, except that they do not explicitly represent trait distributions. Rather, they store some internal encoding of the observed trait vectors, for example through error-driven (Van Rooy et al., 2003) or associative (Kashima et al., 2000) learning. This means that the connectionist models cannot directly answer queries about trait distributions (e.g., perceived variability), which figure prominently in the experimental literature. Bayesian models specify a "natively probabilistic" representation that can directly answer queries about trait distributions.

Exemplar models can also answer queries about trait distributions, by computing statistics on the distribution of exemplars stored in memory. However, it has been argued that certain experimental observations are problematic for exemplar models. First, an exemplar model can only learn from information about exemplars, yet people are able to learn about groups from abstract

statements (Park and Hastie, 1987). Moreover, Park and Hastie argued that subjects in their studies were reporting perceived variability based on abstract information about feature frequency rather than specific exemplars, because their reports were unaffected by making memory for some exemplars more or less accessible. Second, an outgroup homogeneity effect can sometimes be observed even when subjects are exposed to equivalent numbers of exemplars from both the ingroup and the outgroup, simply by invoking a competitive context (Judd and Park, 1988). This finding is problematic for the Bayesian model we presented as well, and speaks to a fundamental limitation of models that are based on purely statistical information (see next section).

Our work is most closely related to abstraction models that have been studied in the categorization literature, especially Anderson's rational model of categorization (Anderson, 1991; Sanborn et al., 2010), which also uses the Chinese restaurant process prior over a set of underlying "clusters." Spicer and Sanborn (2017) were the first to apply this kind of model to the stereotyping literature, and subtyping in particular. We have expanded upon this approach, applying it more broadly and extending it to hierarchically organized groups. A key innovation of this kind of model relative to the other non-Bayesian models of stereotyping reviewed above is that there is provision for creation of novel social groups and subgroups. The other models generally assume a fixed number of groups and lack a mechanism for creating new ones. We have argued that group creation occurs through Bayesian structure learning.

As noted in Spicer and Sanborn (2017), subtyping and subgrouping may be considered the result of a common process of partitioning groups into subsets, with subtype representing the limit case of subgroups with only a single member. This partitioning mechanism is already sufficient to formalize Allport's "refencing" process. We have argued that a complete understanding of stereotype change requires a hierarchical representation of groups, in order to capture the statistical structure shared by a subgroup and its superordinate group. There is currently a dearth of evidence directly supporting this claim, which we see as an opportunity for future research. The simulation shown in Figure 8 suggests one way that hierarchical and non-hierarchical model predictions could be pulled apart.

Limitations and future directions

Following a venerable tradition in social psychology (e.g., Taylor, 1981; Brewer et al., 1981), our treatment of group stereotypes conceptualized them as a form of category knowledge. Consequently, our modeling assumptions, explanations and predictions mirrored those of models applied to non-social domains (Anderson, 1991; Sanborn et al., 2010; Fried and Holyoak, 1984; Feldman et al., 2009). However, we should not lose sight of the fact that social groups are special, in that they serve functions that do not exist in non-social domains. For example, as recently emphasized by Cikara (2021), group identification—figuring out who belongs to "us" versus "them"—is critical to intra-group cooperation and inter-group competition. Individuals form coalitions, in part, based on their beliefs about group identity, which are in turn strengthened by the resulting coalitions. These identities may or may not map onto the categories (e.g., race, gender) people sometimes use to organize their beliefs about social groups. Thus, flexible beliefs about group identity are both the input and the output of affiliative social behavior. This implies that our explanations of social structure learning will be incomplete as long as they are based on purely non-social categorization processes.

A related issue concerns stastistical vs. non-statistical explanations of stereotyping. The approach developed in this paper is purely statistical: all learning is driven by patterns of observed data (though, again, that could be first-hand observation, from media sources, from close others). We demonstrated that many stereotyping phenomena can be explained by statistical models—including some that have historically been conceived in terms of erroneous information processing or some form of motivated cognition (e.g., illusory correlation and the outgroup homogeneity effect; see Hilton and Von Hippel, 1996, for a review). However, we do not mean to suggest that motivational factors are irrelevant, and here we reiterate that social groups are special in ways not captured by a purely statistical account. Coalition formation, for example, is fundamentally motivated in the sense that individuals affiliate in an effort to achieve some goal. That said, we also find it quite remarkable how far we can get in recapitulating and explaining the *why* of past empirical observations with a bare-bones statistical approach.

The most challenging examples for our statistical approach come from studies where moti-

vated cognition seems to operate in opposition to rational information processing. If people use the probability calculus to activate and update their stereotypes based on observed data, then these processes should be immune to influence from putatively irrelevant motivational factors. There is evidence that this postulate is false. Negative stereotypes about a group are activated when a group member disparages the observer, and positive stereotypes are activated when a group member praises the observer (Sinclair and Kunda, 1999, 2000). People are more likely to apply stereotypes when their self-worth is low (Fein and Spencer, 1997) or their mortality is salient (Greenberg et al., 1990). When given the opportunity to seek information about group members, people selectively seek information that preserves their stereotypes (Johnston and Macrae, 1994; Johnston, 1996). People are also less likely to update stereotypes when their accuracy motivation is low (Moreno and Bodenhausen, 1999).

While we consider these compelling sources of evidence for motivational factors in stereotyping, the strong claim that motivation always works in opposition to rational information processing may require some nuance. The critical hinge is the "putatively irrelevant" descriptor. In some cases these motivational factors may actually be relevant to information processing (Gershman, 2019; Kim et al., 2020). For example, confirmation bias in information-seeking is rational under certain assumptions about the data-generating process (Austerweil and Griffiths, 2011; Oaksford and Chater, 1994; Navarro and Perfors, 2011). Greater stereotype activation under low accuracy motivation may likewise be rationalized under assumptions about strategic allocation of cognitive resources (Gershman et al., 2015; Lieder and Griffiths, 2020). The benefits of these models are that they can easily be extended to incorporate motivation. This also allows us to quantify just how much more explanatory value we gain by incorporating motivational factors.

To expand on this last point, it is well-known that the computational intractability of Bayesian structure learning necessitates approximations. We have not committed to a particular process model of human approximate inference. This is a rich topic of active research (see Gershman and Beck, 2017), and several possibilities are viable. For example, people might try to identify a single high probability point estimate of the latent structure, and thus ignore their uncertainty. Alternatively, people might use a Monte Carlo strategy in which they sample structures in proportion to

their posterior probability. The empirical literature at present does not seem to place strong constraints on which of these approximate inference strategies is most plausible. This presents yet another exciting opportunity for discovery at the intersection of structure learning and stereotyping. The issue of intractability is exacerbated by the fact that the trait space used by people in naturalistic settings is presumably enormous. One speculative possibility is that people apply some kind of selection or compression of this high-dimensional space into a more tractable low-dimensional space. For example, existing empirical work on stereotype content indicates that many social traits and features collapse into broader dimensions—e.g., warmth, competence, moral character, conservative/ progressive beliefs (Fiske et al., 2007; Goodwin et al., 2014; Koch et al., 2016). Studies combining differential weights on these fundamental dimensions of stereotypes with the structure provided in our paper is an exciting avenue for future research.

Finally, another limitation of our model is that the structure learning mechanism does not take into account explicit group labels. These explicit labels may provide constraints on the structure of stereotypes that get integrated with unsupervised structure learning, in a manner similar to models of "semi-supervised" learning (Gibson et al., 2013; Vong et al., 2016). In particular, we conjecture that people may rely more strongly on explicit labels when the observed data driving unsupervised learning is sparse or ambiguous. Learning about a group for the first time, one may update beliefs attached to the explicit group label (supervised learning), but further experience with that group may reveal a finer-grained differentiation of individuals that updates beliefs about latent groups or sub-groups (unsupervised learning). We see the integration of distinctively social structures and processes into our theoretical framework as an important direction for future work.

Implications

Despite the limitations reviewed above, we think there are several exciting implications of the model we've presented here. As we note in the introduction, the major innovation of our approach is that we specifically consider the setting in which group membership is unobservable and hence must be inferred. This is very different from the approaches past research has taken in which researchers choose, a priori, categories and traits that are assumed to be counter-stereotypical. There

are a number of limitations to the category-based approach, particularly for making predictions about novel contexts or about how groups and stereotypes will change over time. First, social categories are not fixed, homogenous entities (Cikara, 2021; Zárate et al., 2019). What is counter-stereotypic today may not be so tomorrow, and not in equal measure for all members of a given category (e.g., it would be much more surprising to see Kamala Harris than Condoleezza Rice appear at the Republican National Convention, though both are Black women in politics). Furthermore, studies based on social categories make it difficult to infer from them anything about generalized group processes. For example, some but not all social categories have strong stereotypes associated with them; perceivers' familiarity with the groups in question will vary; and so on. Our approach sidesteps these challenges because we can design experiments knowing precisely what observed data perceivers possess that then gives rise to different group structures, including the identification of subtypes and subgroups. Finally, the category-based approach is limited because it is context insensitive; it breaks down as agents' feature spaces become more complex. As we have demonstrated, our approach can accommodate this challenge because the model takes a vector of features as an input to generate group structure.

In the decades since work on impression formation and updating began, dozens of papers have documented the conditions under which updating, subgrouping, and subtyping occur; however, there is still no unified theory to answer the following questions: When does a collective of people become a group? When do our representations of one group cleave into two versus allow for two subgroups within a higher-order superordinate group? How do these different structures affect our beliefs about said groups (Hamilton et al., 2009)? And what happens when explicit categories intersect with alternative cues to social group structure? Incorporating structure learning begins to address major gaps in knowledge regarding how the mind solves the problems of social categorization, subgrouping versus subtyping, and cross-categorization. The model laid out here can begin to advance our understanding of how people decide what "counts" as a group. It makes specific predictions about (i) the mechanisms by which social structures influence individuals' beliefs and behaviors and (ii) the temporal dynamics underlying the group discovery and updating process. Suitably generalized, it can also make predictions about (iii) how people balance explicit

and latent groupings as inputs to their social structures, and (iv) how people solve the problem of context sensitivity and cross-cutting social categories (Gershman and Cikara, 2020).

Turning now to a consideration of more applied implications, being able to specify just how 'atypical' agents need to be in order to shift stereotypes would be illuminating in efforts to correct overly negative stereotypes (see Figure 8 above). Too little atypicality will result in too small a shift; too much atypicality will result in subtyping and therefore zero stereotype shift. For example, this approach could be used to rehabilitate perceptions of immigrants, who are often characterized as criminal (Stephan et al., 1999) despite data indicating either no relationship or a small negative relationship between immigration inflows and local crime rates (Ousey and Kubrin, 2018). In line with this idea, American participants who read stories about counter-stereotypic, i.e., high-achieving Syrian and Mexican immigrants, along with high-achieving German and Russian immigrants exhibited more positive and similar (across nationalities) evaluations of those exemplars' nationality groups relative to pre-story evaluations (Martinez et al., 2021). But again, being able to quantify the *degree* of atypicality for maximal updating impact would confer a major benefit to any such corrective effort.

Conclusion

Our model seeks to answer the why, when, and how questions of stereotype change as follows. Stereotypes change in order to track the distribution of traits in a group. This happens when the underlying parameters governing the trait distribution are inferred to have changed. In some cases, exposure to stereotype-inconsistent traits does not drive stereotype change because deviant individuals are assigned to new groups. The theoretical framework underlying both parameter and structure learning is Bayesian inference. Though a number of questions remain, understanding the interplay between stereotype content (parameters) and stereotype organization (structure) is an important step towards a complete theory of stereotype change.

References

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII*, pages 1–198. Springer.

Allport, G. W. (1954). *The Nature of Prejudice*. Addison-Wesley.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Lawrence Erlbaum Associates.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98:409–429.

Ashmore, R. D. (1981). Sex stereotypes and implicit personality theory. In Hamilton, D., editor, *Cognitive Processes in Stereotyping and Intergroup Behavior*, pages 37–81. Erlbaum.

Ashmore, R. D. and Del Boca, F. K. (1981). Conceptual approaches to stereotypes and stereotyping. In Hamilton, D., editor, *Cognitive Processes in Stereotyping and Intergroup Behavior*, pages 1–35. Erlbaum.

Austerweil, J. L., Gershman, S. J., Tenenbaum, J. B., and Griffiths, T. L. (2015). Structure and flexibility in Bayesian models of cognition. *Oxford Handbook of Computational and Mathematical Psychology*, pages 187–208.

Austerweil, J. L. and Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, 35:499–526.

Banaji, M. R. (2002). Stereotypes, social psychology of. In Smelser, N. J. and Baltes, P. B., editors, *International Encyclopedia of the Social and Behavioral Sciences*, pages 15100–15104. Pergamon New York.

Bergsieker, H., Leslie, L., Constantine, V., and Fiske, S. (2012). Stereotyping by omission: eliminate the negative, accentuate the positive. *Journal of Personality and Social Psychology*, 102:1214–1238.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57:1–30.

Bodenhausen, G. and Lichtenstein, M. (1987). Social stereotypes and information-processing strategies: The impact of task complexity. *Journal of Personality and Social Psychology*, 52:871–880.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131:1753–1794.

Bott, F. M., Kellen, D., and Klauer, K. C. (2021). Normative accounts of illusory correlations. *Psychological Review*, 128:856–878.

Brewer, M., Dull, V., and Lui, L. (1981). Perceptions of the elderly: Stereotypes as prototypes. *Journal of Personality and Social Psychology*, 41:656–670.

Brigham, J. (1971). Ethnic stereotypes. *Psychological Bulletin*, 76:15–38.

Canini, K. R. and Griffiths, T. L. (2011). A nonparametric Bayesian model of multi-level category learning. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Chan, W., McCrae, R. R., De Fruyt, F., Jussim, L., Löckenhoff, C. E., De Bolle, M., Costa Jr, P. T., Sutin, A. R., Realo, A., Allik, J., et al. (2012). Stereotypes of age differences in personality traits: Universal and accurate? *Journal of Personality and Social Psychology*, 103:1050–1066.

Charlesworth, T. E. and Banaji, M. R. (2021). Patterns of implicit and explicit stereotypes III: Long-term change in gender stereotypes. *Social Psychological and Personality Science*, pages 1–13.

Cikara, M. (2021). Causes and consequences of coalitional cognition. *Advances in Experimental Social Psychology*.

Corneille, O., Klein, O., Lambert, S., and Judd, C. M. (2002). On the role of familiarity with units of measurement in categorical accentuation: Tajfel and Wilkes (1963) revisited and replicated. *Psychological Science*, 13:380–383.

Costello, F. and Watts, P. (2019). The rationality of illusory correlation. *Psychological Review*, 126:437–450.

Crawford, J. T., Jussim, L., Madon, S., Cain, T. R., and Stevens, S. T. (2011). The use of stereotypes and individuating information in political person perception. *Personality and Social Psychology Bulletin*, 37:529–542.

Dannals, J. E. and Miller, D. T. (2017). Social norm perception in groups with outliers. *Journal of Experimental Psychology. General*, 146:1342–1359.

Dasgupta, I. and Griffiths, T. L. (2022). Clustering and the efficient use of cognitive resources. *Journal of Mathematical Psychology*, page 102675.

Diekman, A. B., Eagly, A. H., and Kulesa, P. (2002). Accuracy and bias in stereotypes about the social and political attitudes of women and men. *Journal of Experimental Social Psychology*, 38:268–282.

Eagly, A., Nater, C., Miller, D., Kaufmann, M., and Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of us public opinion polls from 1946 to 2018. *American Psychologist*, 75:301–315.

Fein, S. and Spencer, S. J. (1997). Prejudice as self-image maintenance: Affirming the self through derogating others. *Journal of personality and Social Psychology*, 73:31–44.

Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116:752–782.

Fiedler, K. (1996). Explaining and simulating judgment biases as an aggregation phenomenon in probabilistic, multiple-cue environments. *Psychological Review*, 103:193–214.

Fiedler, K., Freytag, P., and Meiser, T. (2009). Pseudocontingencies: An integrative account of an intriguing cognitive illusion. *Psychological Review*, 116:187–206.

Fiske, S. (1993). Controlling other people: The impact of power on stereotyping. *American Psychologist*, 48:621–628.

Fiske, S. T., Cuddy, A. J., and Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, 11:77–83.

Fiske, S. T. and Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology*, 23:1–74.

Fried, L. S. and Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10:234—257.

Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin & Review*, 26:13–28.

Gershman, S. J. (2021). *What Makes Us Smart*. Princeton University Press.

Gershman, S. J. and Beck, J. M. (2017). Complex probabilistic inference. In Moustafa, A., editor, *Computational Models of Brain and Behavior*, pages 453—464. Wiley Hoboken, NJ.

Gershman, S. J. and Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56:1–12.

Gershman, S. J. and Cikara, M. (2020). Social-structure learning. *Current Directions in Psychological Science*, 29:460–466.

Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349:273–278.

Gershman, S. J., Jones, C. E., Norman, K. A., Monfils, M.-H., and Niv, Y. (2013). Gradual extinction prevents the return of fear: implications for the discovery of state. *Frontiers in Behavioral Neuroscience*, 7:164.

Gershman, S. J., Monfils, M.-H., Norman, K. A., and Niv, Y. (2017a). The computational nature of memory modification. *Elife*, 6:e23763.

Gershman, S. J., Pouncy, H. T., and Gweon, H. (2017b). Learning the structure of social influence. *Cognitive Science*, 41:545–575.

Gershman, S. J., Tenenbaum, J. B., and Jäkel, F. (2016). Discovering hierarchical motion structure. *Vision Research*, 126:232–241.

Gibson, B. R., Rogers, T. T., and Zhu, X. (2013). Human semi-supervised learning. *Topics in Cognitive Science*, 5:132–172.

Goodwin, G., Piazza, J., and Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, 106:148–168.

Greenberg, J., Pyszczynski, T., and Solomon, S. (1990). Evidence for terror management theory: the effects of mortality salience on reactions to those who threaten or bolster the cultural worldview. *Journal of Personality and Social Psychology*, 58:308–318.

Greenwald, A. G. and Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102:4–27.

Hamilton, D. and Rose, T. (1980). Illusory correlation and the maintenance of stereotypic beliefs. *Journal of Personality and Social Psychology*, 39:832–845.

Hamilton, D. L. and Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology*, 12:392–407.

Hamilton, D. L., Sherman, S. J., Crump, S. A., and Spencer-Rodgers, J. (2009). The role of entitativity in stereotyping. In Nelson, T. D., editor, *Handbook of Prejudice, Stereotyping, and Discrimination*, pages 179–198. Taylor & Francis.

Hantzi, A. (1995). Change in stereotypic perceptions of familiar and unfamiliar groups: The pervasiveness of the subtyping model. *British Journal of Social Psychology*, 34:463–477.

Harnad, S. (1987). *Categorical Perception*. Cambridge University Press.

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

Hewstone, M. and Hamberger, J. (2000). Perceived variability and stereotype change. *Journal of Experimental Social Psychology*, 36:103–124.

Hewstone, M., Hassebrauck, M., Wirth, A., and Waenke, M. (2000). Pattern of disconfirming information and processing instructions as determinants of stereotype change. *British Journal of Social Psychology*, 39:399–411.

Hilton, J. L. and Von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, 47:237–271.

Johnston, L. (1996). Resisting change: information-seeking and stereotype change. *European Journal of social psychology*, 26:799–825.

Johnston, L. and Hewstone, M. (1992). Cognitive models of stereotype change: 3. subtyping and the perceived typicality of disconfirming group members. *Journal of Experimental Social Psychology*, 28:360–386.

Johnston, L., Hewstone, M., Pendry, L., and Frankish, C. (1994). Cognitive models of stereotype change (4): Motivational and cognitive influences. *European Journal of Social Psychology*, 24:237–265.

Johnston, L. C. and Macrae, C. N. (1994). Changing social stereotypes: The case of the information seeker. *European Journal of Social Psychology*, 24:581–592.

Jost, J. T. and Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology*, 33:1–27.

Judd, C. and Park, B. (1988). Out-group homogeneity: Judgments of variability at the individual and group levels. *Journal of Personality and Social Psychology*, 54:778–788.

Jussim, L., Cain, T. R., Crawford, J. T., Harber, K., and Cohen, F. (2009). The unbearable accuracy of stereotypes. In Nelson, T. D., editor, *Handbook of Prejudice, Stereotyping, and Discrimination*, pages 199–227. Psychology Press.

Kashima, Y., Woolcock, J., and Kashima, E. (2000). Group impressions as dynamic configurations: The tensor product model of group impression formation and change. *Psychological Review*, 107:914–942.

Kemp, C. and Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105:10687–10692.

Kim, M., Park, B., and Young, L. (2020). The psychology of motivated versus rational impression updating. *Trends in Cognitive Sciences*, 24:101–111.

Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., and Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative-progressive beliefs, and communion. *Journal of Personality and Social Psychology*, 110:675–709.

Koenig, A. M. and Eagly, A. H. (2014). Evidence for the social role theory of stereotype content: Observations of groups' roles shape stereotypes. *Journal of Personality and Social Psychology*, 107:371–392.

Konovalova, E. and Le Mens, G. (2020). An information sampling explanation for the in-group heterogeneity effect. *Psychological Review*, 127:47–73.

Kraus, S., Ryan, C. S., Judd, C. M., Hastie, R., and Park, B. (1993). Use of mental frequency distributions to represent variability among members of social categories. *Social Cognition*, 11:22–43.

Krueger, J. (1992). On the overestimation of between-group differences. *European Review of Social Psychology*, 3:31–56.

Krueger, J. (1996). Probabilistic national stereotypes. *European Journal of Social Psychology*, 26:961–980.

Krueger, J. and Rothbart, M. (1988). Use of categorical and individuating information in making inferences about personality. *Journal of Personality and Social Psychology*, 55:187–195.

Krueger, J. and Rothbart, M. (1990). Contrast and accentuation effects in category learning. *Journal of Personality and Social Psychology*, 59:651–663.

Kunda, Z. and Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: Constructing grounds for subtyping deviants. *Journal of Personality and Social Psychology*, 68:565–579.

Kunda, Z. and 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*, 72:965–979.

Kunda, Z. and Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, 103:284–308.

Lau, T., Gershman, S. J., and Cikara, M. (2020). Social structure learning in human anterior insula. *Elife*, 9:e53162.

Lau, T., Pouncy, H. T., Gershman, S. J., and Cikara, M. (2018). Discovering social groups via latent structure learning. *Journal of Experimental Psychology: General*, 147:1881–1891.

Lee, Y.-T., McCauley, C., and Jussim, L. (2013). Stereotypes as valid categories of knowledge and human perceptions of group differences. *Social and Personality Psychology Compass*, 7:470–486.

Lieder, F. and Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43.

Linville, P., Fischer, G., and Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: empirical evidence and a computer simulation. *Journal of personality and social psychology*, 57:165–188.

Linville, P. W. and Fischer, G. W. (1993). Exemplar and abstraction models of perceived group variability and stereotypicality. *Social Cognition*, 11:92–125.

Lippmann, W. (1922). *Public Opinion*. Harcourt, Brace, and Company.

Locksley, A., Borgida, E., Brekke, N., and Hepburn, C. (1980). Sex stereotypes and social judgment. *Journal of Personality and Social psychology*, 39:821–831.

Macrae, C., Milne, A., and Bodenhausen, G. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology*, 66:37–47.

Martinez, J. E., Feldman, L. A., Feldman, M. J., and Cikara, M. (2021). Narratives shape cognitive representations of immigrants and immigration-policy preferences. *Psychological Science*, 32:135–152.

Maurer, K., Park, B., and Rothbart, M. (1995). Subtyping versus subgrouping processes in stereotype representation. *Journal of Personality and Social Psychology*, 69:812–824.

McCauley, C. and Stitt, C. (1978). An individual and quantitative measure of stereotypes. *Journal of Personality and Social Psychology*, 36:929–940.

McCauley, C., Stitt, C., and Segal, M. (1980). Stereotyping: From prejudice to prediction. *Psychological Bulletin*, 87:195–208.

McGarty, C. and Penny, R. (1988). Categorization, accentuation and social judgement. *British Journal of Social Psychology*, 27:147–157.

McGarty, C. and Turner, J. C. (1992). The effects of categorization on social judgement. *British Journal of Social Psychology*, 31:253–268.

Miller, D. I., Eagly, A. H., and Linn, M. C. (2015). Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology*, 107:631–644.

Moreno, K. N. and Bodenhausen, G. V. (1999). Resisting stereotype change: The role of motivation and attentional capacity in defending social beliefs. *Group Processes & Intergroup Relations*, 2:5–16.

Mullen, B. and Hu, L.-T. (1989). Perceptions of ingroup and outgroup variability: A meta-analytic integration. *Basic and Applied Social Psychology*, 10:233–252.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Murphy, R. A., Schmeer, S., Vallee-Tourangeau, F., Mondragon, E., and Hilton, D. (2011). Making the illusory correlation effect appear and then disappear: The effects of increased learning. *Quarterly Journal of Experimental Psychology*, 64:24–40.

Navarro, D. and Perfors, A. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, 118:120–134.

Nelson, T. E., Biernat, M. R., and Manis, M. (1990). Everyday base rates (sex stereotypes): Potent and resilient. *Journal of Personality and Social Psychology*, 59:664–675.

Nisbett, R. and Kunda, Z. (1985). Perception of social distributions. *Journal of Personality and Social Psychology*, 48:297–311.

Oaksford, M. and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101:608–31.

Ostrom, T. and Sedikides, C. (1992). Out-group homogeneity effects in natural and minimal groups. *Psychological Bulletin*, 112:536–552.

Ousey, G. C. and Kubrin, C. E. (2018). Immigration and crime: Assessing a contentious issue. *Annual Review of Criminology*, 1:63–84.

Park, B. and Hastie, R. (1987). Perception of variability in category development: Instance-versus abstraction-based stereotypes. *Journal of Personality and Social Psychology*, 53:621–635.

Park, B. and Judd, C. (1990). Measures and models of perceived group variability. *Journal of Personality and Social Psychology*, 59:173–191.

Park, B. and 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*, 42:1051–1068.

Park, B., Ryan, C., and Judd, C. (1992). Role of meaningful subgroups in explaining differences in perceived variability for in-groups and out-groups. *Journal of Personality and Social Psychology*, 63:553–567.

Phillips, J. and Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*, 114:4649–4654.

Quattrone, G. and Jones, E. (1980). The perception of variability within in-groups and out-groups: Implications for the law of small numbers. *Journal of Personality and Social Psychology*, 38:141–152.

Queller, S. and Smith, E. (2002). Subtyping versus bookkeeping in stereotype learning and change: Connectionist simulations and empirical findings. *Journal of Personality and Social Psychology*, 82:300–313.

Richards, Z. and Hewstone, M. (2001). Subtyping and subgrouping: Processes for the prevention and promotion of stereotype change. *Personality and Social Psychology Review*, 5:52–73.

Robert, C. P. (2007). *The Bayesian Choice: from Decision-theoretic Foundations to Computational Implementation*. Springer.

Rogers, K. H. and Wood, D. (2010). Accuracy of united states regional personality stereotypes. *Journal of Research in Personality*, 44:704–713.

Rothbart, M. (1981). Memory processes and social beliefs. In Hamilton, D., editor, *Cognitive Processes in Stereotyping and Intergroup Behavior*, pages 145–181. Erlbaum.

Rothbart, M. and John, O. P. (1985). Social categorization and behavioral episodes: A cognitive analysis of the effects of intergroup contact. *Journal of Social Issues*, 41:81–104.

Ryan, C. S., Judd, C. M., and Park, B. (1996). Effects of racial stereotypes on judgments of individuals: The moderating role of perceived group variability. *Journal of Experimental Social Psychology*, 32:71–103.

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, 117:1144–1167.

Simon, B. and Brown, R. (1987). Perceived intragroup homogeneity in minority–majority contexts. *Journal of Personality and Social Psychology*, 53:703–711.

Simon, B. and Pettigrew, T. F. (1990). Social identity and perceived group homogeneity: Evidence for the ingroup homogeneity effect. *European Journal of Social Psychology*, 20:269–286.

Sinclair, L. and Kunda, Z. (1999). Reactions to a black professional: motivated inhibition and activation of conflicting stereotypes. *Journal of Personality and Social Psychology*, 77:885–904.

Sinclair, L. and Kunda, Z. (2000). Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me. *Personality and Social Psychology Bulletin*, 26:1329–1342.

Smith, E. and Decoster, J. (1998). Knowledge acquisition, accessibility, and use in person perception and stereotyping: Simulation with a recurrent connectionist network. *Journal of Personality and Social Psychology*, 74:21–35.

Smith, E. R. and Zárate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review*, 99:3–21.

Snyder, M. and Miene, P. (1994). On the functions of stereotypes and prejudice. In Zanna M., O. J., editor, *The Psychology of Prejudice*, pages 33–42. Psychology Press.

Spears, R., Eiser, J. R., and Van Der Pligt, J. (1987). Further evidence for expectation-based illusory correlations. *European Journal of Social Psychology*, 17:253–258.

Spicer, J. and Sanborn, A. (2017). A rational approach to stereotype change. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Stephan, W. G., Ybarra, O., and Bachman, G. (1999). Prejudice toward immigrants. *Journal of Applied Social Psychology*, 29:2221–2237.

Tajfel, H. (1969). Cognitive aspects of prejudice. *Journal of Biosocial Science*, 1:173–191.

Tajfel, H. and Wilkes, A. L. (1963). Classification and quantitative judgement. *British Journal of Psychology*, 54:101–114.

Taylor, S. E. (1981). A categorization approach to stereotyping. In Hamilton, D., editor, *Cognitive Processes in Stereotyping and Intergroup Behavior*, pages 83–115. Erlbaum.

Van Rooy, D., Van Overwalle, F., Vanhoomissen, T., Labiouse, C., and French, R. (2003). A recurrent connectionist model of group biases. *Psychological Review*, 110:536–563.

Vanhoomissen, T. and Van Overwalle, F. (2010). Me or not me as source of ingroup favoritism and outgroup derogation: A connectionist perspective. *Social Cognition*, 28:84–109.

Vong, W. K., Navarro, D. J., and Perfors, A. (2016). The helpfulness of category labels in semi-supervised learning depends on category structure. *Psychonomic Bulletin & Review*, 23:230–238.

Weber, R. and Crocker, J. (1983). Cognitive processes in the revision of stereotypic beliefs. *Journal of Personality and Social Psychology*, 45:961–977.

Wilder, D. A. (1984a). Intergroup contact: The typical member and the exception to the rule. *Journal of Experimental Social Psychology*, 20:177–194.

Wilder, D. A. (1984b). Predictions of belief homogeneity and similarity following social categorization. *British Journal of Social Psychology*, 23:323–333.

Yzerbyt, V. Y., Coull, A., and Rocher, S. J. (1999). Fencing off the deviant: The role of cognitive resources in the maintenance of stereotypes. *Journal of Personality and Social Psychology*, 77:449–462.

Zárate, M. A., Reyna, C., and Alvarez, M. J. (2019). Cultural inertia, identity, and intergroup dynamics in a changing context. In *Advances in Experimental Social Psychology*, volume 59, pages 175–233. Elsevier.