Computational Cognitive Mechanisms of Stereotypes: Social Learning and Generalization

Authors: Wang Yingjie, Ruyuan Zhang, Ruyuan Zhang

Date: 2026-04-28T14:39:04+00:00

Abstract

Stereotypes are generalized beliefs that "a certain group possesses certain traits," which profoundly influence interpersonal interactions and intergroup relations; therefore, understanding the cognitive mechanisms underlying their formation and maintenance is of great significance. However, traditional theories mostly remain at the descriptive level and lack an integrative mechanistic explanation for the process of stereotypes from acquisition to application. Centering on the two stages of social learning and social generalization, this paper reviews the computational cognitive mechanisms of stereotypes by integrating reinforcement learning and Bayesian theory. At the social learning level, we first elaborate on how Bayesian structure learning infers latent group structures and distinguish two pathways for establishing group-trait associations: the experiential pathway primarily forms automated associative representations through reinforcement learning mechanisms, while the linguistic pathway primarily transmits probabilistic propositional representations through Bayesian inference. During the process of association updating and consolidation, prediction error is a crucial driver for updating; however, prior bias and asymmetric learning rates cause updates to favor the maintenance of existing beliefs. Meanwhile, the exploration-exploitation dilemma suggests that biased information sampling is a significant reason for the persistence of stereotypes. At the social generalization level, we analyze the group categorization process based on perceptual and functional similarity when encountering new individuals, and further discuss the retrieval mechanisms of learned associative knowledge. Finally, this paper proposes three future research directions: incorporating relational cues as a generalization path beyond feature similarity, developing social cognitive maps as an integrative framework for multi-cue representations, and utilizing large language models to simulate the consolidation process of stereotypes in linguistic transmission.

Full Text

Preamble

Computational Cognitive Mechanisms of Stereotypes: Social Learning and Generalization

Yingjie Wang [1], **Ruyuan Zhang** [2,3,4*]

[1] School of Psychology, Shanghai Jiao Tong University, and Brain Health Institute, National Center for Mental Health, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai 200030, China [2] School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing 100871, China [3] IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China

---

Abstract

Stereotypes are generalized beliefs about the characteristics and behaviors of social groups. While traditional social psychology has extensively documented the formation and impact of stereotypes, the underlying computational mechanisms remain less clear. This review explores stereotypes through the lens of computational cognitive science, focusing on two core processes: social learning and generalization. We examine how individuals integrate social information to form group-level representations and how these representations are generalized to novel group members. By synthesizing recent advances in machine learning and reinforcement learning, we propose a computational framework to understand the persistence and flexibility of stereotypes in dynamic social environments.

1. Introduction

Stereotypes serve as a cognitive heuristic that allows individuals to process complex social information efficiently. By categorizing individuals into groups based on shared attributes—such as ethnicity, gender, or occupation—people can make rapid inferences about others' traits and likely behaviors. However, these generalizations often lead to systematic biases and social prejudice. Understanding the cognitive architecture of stereotypes requires moving beyond descriptive accounts toward formal computational models that can explain how social data is sampled, processed, and applied.

2. The Computational Framework of Social Learning

Social learning is the process by which individuals acquire information about social groups from their environment, including direct interactions and indirect

observations (e.g., media, peer influence). From a computational perspective, this can be modeled as a Bayesian inference problem or a reinforcement learning process.

2.1 Bayesian Inference in Stereotype Formation In the Bayesian framework, stereotypes can be viewed as "priors" that are updated as new evidence is encountered. If an individual observes a member of Group G exhibiting behavior B, the posterior probability that Group G possesses trait T is updated according to:

$$P(T|B) = \frac{P(B|T)P(T)}{P(B)}$$

4 北京大学机器感知与智能教育部重点实验室，北京 100871

- Corresponding Author

Ying-Jie WANG1, Ru-Yuan ZHANG2, 3, 4*

Brain Health Institute, National Center for Mental Disorders, Shanghai Mental Health Center, Shanghai

Jiao Tong University School of Medicine and School of Psychology, Shanghai 200030, China.

School of Psychological and Cognitive Sciences and Beijng Key Laboratory of Behavior and Mental

Health, Peking University, Beijing 100871, China.

IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China.

Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871,

China.

Stereotypes are generalized beliefs that "a certain group possesses certain traits," which profoundly influence interpersonal interactions and intergroup relations.

Consequently, understanding the cognitive mechanisms underlying their formation and maintenance is of significant importance. However, traditional theories have largely remained at a descriptive level, lacking an integrated mechanistic explanation of the process from stereotype acquisition to application. This paper reviews the computational cognitive mechanisms of stereotyping by focusing on two key stages: social learning and social generalization, integrating reinforcement learning and Bayesian theory. At the social learning level, we first elucidate how Bayesian structural learning infers latent group structures.

We then distinguish between two pathways for establishing group-trait associations: the empirical pathway, which primarily forms automated associative representations through reinforcement learning mechanisms, and the linguistic pathway, which primarily transmits probabilistic propositional representations via Bayesian inference. During the updating and consolidation of these associations, prediction error serves as a critical driver for change; however, prior biases and asymmetric learning rates cause these updates to favor the maintenance of existing beliefs. Furthermore, the exploration-exploitation trade-off suggests that biased information sampling is a major factor contributing to the persistence of stereotypes.

At the social generalization level, we analyze the group categorization process when encountering new individuals based on perceptual and functional similarity, and further discuss the retrieval mechanisms of previously learned associative knowledge. Finally, this paper proposes three directions for future research: incorporating relational cues as a generalization path beyond feature similarity, developing social cognitive maps as an integrative framework for multi-cue representations, and utilizing Large Language Models (LLMs) to simulate the consolidation of stereotypes during linguistic transmission.

关键词

Stereotypes, Social Generalization, Reinforcement Learning, Bayesian Theory, Computational Modeling

1. 引言

When we think of a surgeon, the image of a calm and rational male often comes to mind; conversely, the mention of a kindergarten teacher more easily evokes the image of a gentle and patient female. These ubiquitous, automated, and rapid judgments in daily life are classic manifestations of stereotypes. A stereotype is generally defined as a set of overgeneralized and relatively fixed beliefs held by social members regarding a specific group [**?**, **?**]. As a pervasive social cognitive phenomenon, stereotyping profoundly influences interpersonal interactions, intergroup relations, and even broader social structures [**?**, **?**].

Since the inception of social psychology, investigating the causes and mechanisms of stereotypes has remained one of its core agendas. Early research focused primarily on motivational and emotional dimensions. Whether through the psychodynamic explanations of the Authoritarian Personality [**?**, **?**], the resource competition perspective of Realistic Group Conflict [**?**, **?**], or the emphasis on maintaining group self-esteem within Social Identity Theory [**?**, **?**, **?**], these

classical theories view stereotypes as specific phenomena driven by intense emotions, internal needs, or group dynamics. Subsequently, the rise of social cognition theory shifted the focus toward information processing. Researchers pro-

posed that stereotypes are essentially cognitive shortcuts or heuristic strategies adopted by the brain to simplify the processing of complex social information when cognitive resources are limited [?, ?, ?, ?]. For example, individuals automatically perform social categorization based on cues such as race and gender, invoking pre-existing schemas to make rapid inferences [?, ?, ?, ?, ?].

Despite the abundance of relevant theories, most remain at a descriptive level. To truly understand the ubiquity and persistence of stereotypes and to develop effective interventions, we must move from describing "what" they are to explaining "how" and "why" they function by delving into their cognitive computational mechanisms. In recent years, computational frameworks such as reinforcement learning and Bayesian theory have been introduced to the fields of social learning and decision-making. These frameworks use algorithmic models to describe how individuals acquire social knowledge and make judgments accordingly [?, ?, ?, ?], and their model predictions have been supported by neuroimaging research [?, ?, ?, ?, ?]. Within this framework, the formation and application of stereotypes can be decomposed into two core processes: social learning, which focuses on how individuals acquire knowledge about groups, and social generalization, which focuses on how individuals extend limited experience to new situations or unknown individuals. However, existing research often adopts a single perspective, and there is still a lack of integration regarding the mechanisms of the entire process. Consequently, this paper focuses on these two stages, combining reinforcement learning and Bayesian theory to systematically review the cognitive computational mechanisms underlying the formation, updating, and generalized application of stereotypes.

2. 社会学习：刻板印象的形成

The core of a stereotype lies in the belief structure that "a certain group possesses a specific trait." The formation of this structure can be decomposed into two fundamental questions: First, where does the concept of "a certain group" originate? That is, how does the brain discern which individuals belong to the same category from complex social observations? Second, how is the belief that they "possess a specific trait" acquired? This concerns how individuals learn the associations between a group and its corresponding traits or values. Building upon these foundations, the manner in which these associations are updated or solidified in the face of new evidence determines the plasticity or persistence of the stereotype.

2.1 群体的发现与构建

Stereotypes represent generalizations about specific groups; therefore, understanding how individuals identify social groups and represent social structures is of paramount importance. In reality, group boundaries are often not directly observable; instead, they must be inferred from patterns of individual behavior, social interactions, and shared characteristics.

The Social Structure Learning Model proposes that the brain utilizes Bayesian inference to discover latent, invisible group structures from observable social data—such as voting behaviors, clothing preferences, and social interactions. Specifically, this involves estimating the posterior probability of a latent group structure given the observed data, denoted as P(latent group structure | observed data) [**?**]. This process transcends simple dyadic similarity judgments (determining whether an individual is similar to oneself) and involves more complex multivariate relational reasoning. For instance, in triadic relationships characterized by logic such as "the enemy of my enemy is my friend" or "the friend of my friend is my friend," individual A and individual B may share no direct feature overlap. However, because both maintain specific relationships with individual C, A may still be inferred to belong to the same group as B [**?**, **?**, **?**].

Experiments by Gershman et al. (2017) provided direct evidence for this mechanism [**?**]. In a movie poster selection task, researchers controlled the dyadic choice overlap between the participant and agents A and B at 50%, making them equivalent from a binary similarity perspective. The critical manipulation involved the choice patterns of a third party, C: when C was highly consistent with both A and the participant (75% overlap), it suggested that all three belonged to the same latent group; conversely, when C was consistent with A but inconsistent with the participant, it suggested that A and C formed a separate group. The results demonstrated that participants' choices were significantly influenced by the latent group structure, and the goodness-of-fit for the Bayesian structure learning model was significantly higher than that of the dyadic similarity model. Lau et al. further discovered that this structure learning guides not only behavioral choices but also affective and cognitive generalization [**?**]. Using a similar paradigm where participants selected political preferences (e.g., support for the death penalty), researchers asked participants to provide social evaluations of each agent following the experiment. The results showed that if the model inferred an agent belonged to the same latent group as the participant, the participant was not only more likely to follow that person on unknown issues but also assigned them higher ratings for morality and popularity. Model-based functional magnetic resonance imaging (fMRI) studies have revealed a corresponding neural dissociation: dyadic similarity correlates with activity in the medial prefrontal cortex and pregenual anterior cingulate cortex (mPFC/pgACC), whereas the representation of latent social group structures correlates with activity in the right anterior insula (rAI) [**?**]. Furthermore, multivoxel pattern analysis (MVPA) research has found that neural patterns in the dorsal anterior cingulate cortex/midcingulate cortex (dACC/MCC) and the anterior insula (AI) can accurately distinguish between ingroups and outgroups [**?**]. This neuroscientific evidence suggests that the encoding of latent group structures has an independent neural basis that cannot be reduced to simple similarity computations.

2.2 群体-特质联结的建立

After identifying a group, another core task of social learning is acquiring knowledge regarding "what traits or social values this group possesses." The establishment of these group-trait associations is achieved through two qualitatively distinct pathways: the experiential pathway and the linguistic pathway. The experiential pathway relies on individuals acquiring associations through direct or indirect social interaction experiences. In contrast, the linguistic pathway involves receiving and transmitting generalized knowledge about groups through the medium of language.

2.2.1 经验通路与联想表征

The experiential pathway encompasses various forms of learning, characterized by the establishment of group-trait associations through direct or indirect interactions between individuals and their social environment. At the representational level, these learning forms primarily establish associative representations —the automatic activation of links between a group and specific traits [**?**, **?**]. At the computational level, reinforcement learning provides an algorithmic description of this process: social interactions are viewed as a series of decision-making processes in which individuals receive rewards or punishments through their actions and utilize this feedback to adjust their evaluations of the group.

The value estimation is updated as $V(\text{group}) \leftarrow V(\text{group}) + \alpha\delta$, where α represents the learning rate, which controls the extent to which new information updates existing value estimates. The term δ represents the prediction error, defined as the discrepancy between the actual outcome received and the initial expectation [**?**, **?**, **?**, **?**, **?**].

There are important computational distinctions among the three core forms of learning within the experiential pathway [**?**, **?**, **?**]. Evaluative Conditioning (EC) serves as the foundation for the formation of affective components in group-trait associations [**?**, **?**, **?**]. In this process, the individual does not need to perform any specific action; rather, an automatic association is established simply because members of a certain group (the conditional stimulus, CS) repeatedly co-occur in time and space with positive or negative stimuli (the unconditional stimulus, US) [**?**, **?**, **?**, **?**].

This type of learning relies primarily on the amygdala, allowing individuals to form automated affective biases toward specific groups even before acquiring concrete trait knowledge through interaction. This corresponds to model-free learning within the reinforcement learning framework [**?**, **?**].

Direct interactive learning depends on an individual' s behavioral decisions and environmental feedback, representing a typical instrumental learning process [**?**, **?**, **?**, **?**, **?**]. Observational learning, by contrast, involves value-shaping mechanisms; here, an observer' s preferences are influenced not only by the feedback received by the target object but also directly by the behavioral preferences of

the actor, leading the observer to unintentionally inherit the biased estimates of others [?, ?, ?, ?].

Through social interaction, individuals can generate value evaluations of others (such as the amount of a reciprocal payment) and infer abstract traits (such as degree of generosity) [?, ?]. At the neural level, social values (such as gaining the trust of others or an increase in social status) and material rewards (such as money) may be encoded as value signals within the same set of brain regions [?, ?], primarily including the ventromedial prefrontal cortex (vmPFC) and the striatum [?, ?, ?]. Compared to the encoding of value, the inference of others' traits and intentions further involves regions such as the posterior cingulate cortex (PCC), precuneus, superior temporal sulcus (STS), and the temporoparietal junction (TPJ) [?, ?, ?, ?].

2.2.2 语言通路与命题表征

The linguistic pathway transmits and forms knowledge about groups through the symbolic system of language. Its uniqueness lies in its primary capacity to carry propositional representations—semantic content capable of truth-value judgment (De Houwer et al., 2021). For instance, the statement "girls are kind" is a proposition that can be evaluated as true or false, rather than a simple associative activation between a group and a trait word (Gelman et al., 2004). As a uniquely human and most widely utilized form of associative acquisition (K. A. Collins & Clément, 2012; Maass, 1999; Martin et al., 2014), language enables individuals to transcend one-off observed behavioral events and generalize information across different individuals and specific contexts. However, highly generalized communication may deviate from the actual behavior of specific individuals in concrete situations (Beukeboom, 2025).

Research on transmission chains reveals how linguistic communication fosters the emergence of stereotypes. Initially unstructured information, when passed through multiple generations, spontaneously evolves into a simplified stereotype system with a clear categorical structure (Hutchison et al., 2018). The inherent tendency toward information compression in linguistic transmission—generalizing specific behaviors into abstract traits, such as paraphrasing "he hit someone" as "he is violent" —further maintains and reinforces these stereotypes (Beukeboom & Burgers, 2019).

The propositional representations formed through the linguistic pathway can be characterized at the computational level using conditional probabilities within a Bayesian framework. Specifically, stereotypes can be formalized as $p(\text{trait} \mid \text{group})$, representing the conditional probability that an individual possesses a certain characteristic given their membership in a specific group.

This conditional probability (McCauley & Stitt, 1978) can be further decomposed according to Bayes' theorem as $p(\text{trait} \mid \text{group}) = p(\text{trait}) \times p(\text{group} \mid \text{trait})/p(\text{group})$. In this equation, $p(\text{trait})$ represents the trait prior (the base rate of the trait in the general population), $p(\text{group} \mid \text{trait})$ is the likelihood (the

probability that a person with that trait belongs to the group), and p(group) is the group prior (the proportion of the group in the total population). Research by Solanki and Cesario (2025) provides empirical evidence for this probabilistic structure. They designed a stereotyping questionnaire testing these four components across eight social categories (e.g., male/female, Asian/Black) with ten traits per category (5 stereotypical/5 non-stereotypical). The results demonstrated that participants' direct posterior estimates were highly correlated with theoretical posteriors, and individuals with higher cognitive ability were better at differentiating information according to Bayesian rules. This suggests that the propositional representation of stereotypes exhibits a probabilistic structure consistent with the Bayesian framework. It should be emphasized that while Bayesian inference provides a possible computational model for propositional representation, there remain significant theoretical differences between the two: propositional representation defines the representational form of cognitive content, whereas Bayesian inference is a computational mechanism describing information integration.

This probabilistic propositional representation has also been validated in Large Language Models (LLMs). The fill-mask association test (FMAT) developed by Bao (2024) provides a means to directly quantify propositional representations at the corpus level. This method utilizes the masked language modeling functionality of the BERT model to calculate $p(\text{group} \mid \text{trait})$ by comparing the predicted probabilities of different group words within trait-descriptive contexts, which is formally consistent with likelihood estimation in the Bayesian framework. This suggests that although LLMs can replicate human implicit biases (Bai, Wang, et al., 2025; Garg et al., 2018; Hagendorff et al., 2023), they reflect group-trait association patterns already present in the corpus rather than independently discovering or constructing these associations.

2.2.3 联想表征与命题表征的交互

Although the experiential and linguistic pathways emphasize different primary types of representation, these two forms of representation continuously interact during actual cognitive processes. The Associative-Propositional Evaluation (APE) model, proposed by Gawronski and Bodenhausen (2006, 2011), elucidates this mechanism. On one hand, associative activation can serve as input for propositional representations; for instance, an automatic negative affective reaction (associative activation) resulting from an interaction with a group member may be captured by consciousness and subsequently transformed into a propositional representation (e.g., "this group might be unfriendly"). On the other hand, propositional representations can constrain associative ones. For example, after learning that a certain stereotype is statistically inaccurate, an individual may suppress behavioral expressions through propositional-level negation, even if the group label still automatically activates a negative valence. De Houwer et al. (2021) further suggested that many psychological phenomena traditionally classified as associative processes may, in essence, be propositional in nature.

A meta-analysis by Kurdi et al. (2023) provides an empirical response to this debate. They found that both the repeated pairing of stimuli (typical input for associative learning) and explicit semantic relationships (such as "A helped B," typical input for propositional learning) can alter evaluations of target objects. However, the efficacy of these two inputs differs in their impact on explicit versus implicit attitudes (Kurdi et al., 2023). these results suggest that in actual cognitive processes, the boundary between associative and propositional representations is more blurred than traditional dual-process theories assume, and that stereotypes may be shaped by both learning mechanisms simultaneously.

2.3 联结的更新与固化

Once a group-trait association is established, it is neither immutable nor easily altered. The following discussion examines the dynamics of these associations from three perspectives: how prediction errors drive updates, how existing beliefs resist such updates, and how the exploration-exploitation dilemma maintains stereotypes at the level of information sampling.

2.3.1 预测误差驱动的更新

The core principle of reinforcement learning is that individuals adjust the weights of existing associations through prediction error (PE)—the discrepancy between actual outcomes and expectations [**?**, **?**]. In the context of stereotyping, the high prediction error generated by counter-stereotypical information may serve as a critical driver for stereotype updating. Falbén et al. employed a probabilistic selection task in which participants were presented with pairs of male and female faces and were required to determine, through trial-and-error learning, which individual in each pair was more likely to enjoy ballet or boxing (interests associated with gender stereotypes) [**?**, **?**]. Computational modeling using the Reinforcement Learning Drift-Diffusion Model (RL-DDM) revealed that participants exhibited faster learning rates when processing counter-stereotypical information (e.g., a male preferring ballet) compared to stereotype-consistent information (e.g., a female preferring ballet or a male preferring boxing).

Golubickis et al. further demonstrated that the degree of congruence between facial features and stereotypes significantly influences learning rates [**?**, **?**]. In counter-stereotypical learning contexts, individuals with high gender-typicality faces (e.g., a highly feminine woman working in construction) are learned more rapidly. Conversely, in stereotype-consistent learning contexts, individuals with low gender-typicality faces (e.g., a less masculine man working in construction) exhibit faster learning rates.

Functional magnetic resonance imaging (fMRI) studies have found that the process of impression updating activates a network of brain regions associated with conflict monitoring and social reasoning, including the rostrolateral prefrontal

cortex (rlPFC), the superior temporal sulcus, the right inferior parietal lobule (rIPL), and the posterior cingulate cortex (PCC) [?, ?, ?, ?].

However, when confronted with individuals who do not conform to stereotypes, the brain does not always update its associative knowledge; it can also flexibly adjust its perception of group structure through two primary mechanisms: subtyping and subgrouping [?, ?, ?, ?]. Subtyping refers to the process of isolating individuals who deviate extremely from typical group characteristics as "exceptions," thereby forming a new category that contains only those individuals and protecting the original stereotype from revision. Subgrouping occurs when multiple individuals collectively exhibit a pattern of deviation, leading the brain to create new nested subcategories within the original broad category (e.g., forming a "feminist" subgroup under the category of "women"). This allows for the refinement of the original stereotype while maintaining its core structure. Bayesian model simulations indicate that subgrouping—and subsequent stereotype modification—is most likely to be triggered when disconfirming information is moderately dispersed and the degree of deviation is moderate (e.g., anomalous behaviors distributed across three agents). In contrast, concentrated and extreme counter-examples (e.g., all anomalous behaviors concentrated in a single individual) are most likely to trigger subtyping, which paradoxically leads to the solidification of original prejudices [?, ?]. This computational finding offers counter-intuitive suggestions for social interventions: rather than promoting a few perfect, extreme counter-stereotypical role models, presenting a large number of diverse group members may be more effective in altering implicit cognitive structures.

2.3.2 先验偏差对更新的抵抗

Even when prediction errors are capable of driving updates, pre-existing stereotypes distort this process at multiple stages, biasing it toward the maintenance rather than the revision of existing beliefs. Computational modeling analyses have revealed two primary mechanisms underlying this phenomenon [?, ?, ?, ?, ?, ?]. First, stereotypes function as biased prior beliefs, establishing divergent initial reward expectations for different groups. For instance, a white participant holding negative stereotypes may preset a lower probability of cooperation (sharing) when initially interacting with a Black player. Second, stereotypes lead to asymmetric learning rates. For example, individuals may learn more rapidly from behaviors that confirm negative expectations while learning more slowly from positive behaviors that challenge those expectations. The synergy of these two mechanisms ensures that even when faced with identical objective feedback, the internal value representations learned by the individual remain biased.

Hedrich et al. [?] provide an explanation for this asymmetry at a more fundamental level: the human reinforcement learning system exhibits a preference for processing features that change slowly, such as personality traits or group characteristics [?, ?]. Rather than focusing on rapidly fluctuating information,

the brain is more inclined

to encode and rely on features perceived as stable. This may serve as one of the cognitive foundations for why stereotypes are so difficult to overturn with a single piece of disconfirming evidence. This is particularly true for moral stereotypes; when participants are first exposed to moral stereotypes regarding a group (e.g., honesty or untrustworthiness), they not only set more extreme initial expectations but also demonstrate a greater reluctance to adjust these expectations in light of counter-evidence [**?**, **?**].

2.3.3 探索-利用困境与顽固性

Why are stereotypes so difficult to change once they are formed? Beyond prior bias and asymmetric learning rates, a deeper reason lies in the bias of information sampling itself—the explore-exploit dilemma (Bai et al., 2022). A rational agent seeking to maximize long-term rewards may tend to continue interacting with a specific group (“exploitation”) if their initial interaction (“exploration”) yielded sufficient returns, rather than exploring other groups that might be equally or even more rewarding. While this locally adaptive exploration strategy secures stable returns in the short term with minimal trial-and-error costs, the long-term cost is substantial: accidental early positive or negative experiences cause individuals to prematurely cease exploration at a local optimum.

This process results in the global solidification of inaccurate impressions. Bai et al. (2022) adapted the multi-armed bandit paradigm to validate this phenomenon: participants engaged in multiple rounds of interaction with four social groups in a fictional city, where they could either choose to cooperate with members of different groups to earn rewards or be passively paired under random assignment conditions. The results showed that even when there were no actual differences between groups, participants who could choose freely tended to concentrate their selections on a single group and significantly overestimated intergroup differences, forming inaccurate global impressions (Bai et al., 2022). Similarly, Allidina and Cunningham (2021) found that individuals holding negative stereotypes about a group tend to avoid interacting with its members; this avoidance behavior reduces opportunities to acquire counter-stereotypical information, creating a similar vicious cycle (Allidina & Cunningham, 2021).

This framework also explains why stereotypes possess the classic two-dimensional structure of warmth and competence (Bai, Griffiths, et al., 2025). Utilizing a contextual multi-armed bandit paradigm, researchers tasked participants with acting as recruiters to assign jobs—distinguished across two dimensions—to members of four virtual groups. Compared to random exploration, participants engaged in autonomous exploration developed more pronounced hierarchical group selections, resulting in greater psychological distance within the two-dimensional space. Three interventions designed to reduce exploration costs (exploration bonuses, reduced reward probabilities, and random constraints) effectively reduced stratification and stereotyping.

This suggests that when exploration is costly and decision-makers must generalize based on shared features such as occupation or educational background, feature-based exploration mechanisms lead decision-makers to separate different groups across multiple dimensions, thereby spontaneously reproducing the two-dimensional stereotype space of warmth and competence.

From the perspective of computational integration, the explore-exploit dilemma may represent the intersection of reinforcement learning and Bayesian theory. Prior beliefs within a Bayesian framework influence the exploration strategies of reinforcement learning: strong prior expectations incline individuals to select interaction partners consistent with those expectations, leading to biased information sampling. In turn, this biased sampling reinforces the original priors, creating a self-perpetuating cycle (Villiger, 2025).

3 社会泛化：刻板印象的应用

Social learning explains how stereotypes are formed; however, acquisition alone is insufficient to account for their pervasive influence. The power of stereotypes relies heavily on generalization—the process of extending limited experiences to novel individuals and contexts [**?**, **?**, **?**, **?**]. A fundamental challenge in this generalization process is determining how the brain, when encountering a previously unknown individual, matches them to a known social category in order to retrieve and apply previously learned associative knowledge.

3.1 泛化路径

Generalization within social contexts also adheres to the universal law of generalization: generalization is highly dependent on similarity, meaning that the more similar two contexts are, the more readily knowledge transfers between them [**?**, **?**, **?**, **?**]. In the application of stereotypes, this similarity-based matching is primarily realized through two distinct pathways: perceptual cues and functional cues.

3.1.1 基于知觉线索

Perceptual-based generalization is the most intuitive pathway for social inference, relying on immediately accessible physical characteristics. These include facial features (such as skin tone, facial structure, and expression), body posture, clothing, and vocal tone (Hu & O' Toole, 2023; Krahé et al., 2021; Todorov et al., 2015). At the computational level, this process can be understood as the brain comparing the perceptual feature vector of a new individual with stored group prototypes within a multidimensional feature space to perform classification. Research has demonstrated that after participants learn that certain individuals are untrustworthy, they exhibit distrust toward strangers with similar facial features; notably, this effect diminishes as facial similarity decreases (FeldmanHall et al., 2018).

From the perspective of neural mechanisms, perceptual generalization is initiated during the early stages of visual processing. When an individual learns that a specific face is associated with an aversive outcome, the tuning curves of neurons in the visual cortex responsible for facial identity cues become sharpened. This sharpening increases the brain' s sensitivity to threat-related facial features, thereby facilitating the generalization of aversive responses to perceptually similar faces (Stegmann et al., 2020).

As the most frequently utilized perceptual cue, the face maintains a systematic association with the content of social stereotypes (Sutherland & Young, 2022). For example, faces judged to belong to a lower social class ("poor") are typically wider, shorter, and have flatter facial contours, often accompanied by downturned mouth corners and darker, cooler skin tones. These specific features overlap significantly with facial characteristics judged as indicating low competence, coldness, and untrustworthiness (Bjornsdottir et al., 2024). A meta-analysis of fMRI studies indicates that the neural computational hubs for facial social evaluation are located within emotion and value assessment circuits. Specifically, the evaluation of negative faces (e.g., those rated as untrustworthy) consistently activates the bilateral amygdala. Conversely, the evaluation of positive faces (e.g., those rated as attractive or trustworthy) activates brain regions associated with reward and value computation, including the medial prefrontal cortex (mPFC), the medial orbitofrontal cortex (mOFC), and the nucleus accumbens (NAcc) (Mende-Siedlecki, Said, et al., 2013).

3.1.2 基于功能线索

Unlike perceptual generalization, functional-based generalization relies on abstract properties that cannot be obtained through immediate observation. At the computational level, functional similarity can be understood as an indirect matching process mediated by latent variables. When two or more individuals share the same function, predict the same outcome, or require the same behavioral response, knowledge transfer can occur even if they are entirely distinct in terms of their perceptual features [**?**, **?**, **?**].

Social roles, particularly occupational labels, serve as the most prevalent functional cues. According to Social Role Theory, stereotypes regarding large-scale social groups (such as different genders or ethnicities) do not originate from an inherent essence, but rather from observations of the social roles that members of these groups typically occupy [**?**, **?**]. For instance, the role of a nurse requires caring traits, while the role of a police officer requires decisiveness. Due to historical and social factors, women have disproportionately worked as nurses and men as police officers; over time, these role-specific traits have been generalized to the corresponding groups, forming gender stereotypes. Numerous studies have confirmed the critical role of social roles in social generalization. Occupational labels not only drive the induction of personality traits and skills but also influence the generalization of rights and obligations. In most cases, the efficacy of these labels takes precedence over cues such as race or gender

[?, ?].

When social role information and gender information are presented simultaneously (e.g., a "male nurse"), the former (e.g., "caring," a trait associated with the nursing role) can override the latter (e.g., "ambitious," a traditional masculine trait), thereby dominating the inferences people make about an individual' s characteristics [?, ?].

3.1.3 知觉与功能线索的整合

Perceptual cues and functional cues do not operate independently; rather, they undergo complex integration within the brain. The Dynamic Interactive Theory proposed by Freeman and Johnson (2016) suggests that recurrent connections and overlapping representations exist between the visual perceptual system (such as the fusiform face area, FFA) and higher-order social conceptual systems (such as the anterior temporal lobe, ATL, and the orbitofrontal cortex, OFC) [?, ?]. For example, if a stereotype that "Black individuals are hostile" is prevalent within a culture, the neural activation patterns for Black faces in the orbitofrontal cortex and visual cortex will spontaneously shift toward the activation patterns associated with angry expressions, even if the face is actually neutral. This top-down modulation explains why stereotypes can distort fundamental visual perception (for instance, making it more likely to misidentify a tool held by a Black person as a weapon).

Transcranial magnetic stimulation (TMS) studies have confirmed that the dorsomedial prefrontal cortex (dmPFC) plays a central role in this integration process, responsible for synthesizing information from various channels—such as facial features and verbal descriptions—into a final social impression [?].

3.2 联结知识的检索与应用

Once a new individual is identified as a member of a specific group, the application of stereotypes transforms into the retrieval of acquired associative knowledge. Just as the social learning phase involves two distinct pathways, retrieval during the generalization phase may also operate along two routes. Associative representations established via the experiential pathway support rapid, automated retrieval: group labels directly activate pre-existing affective valence, driving approach or avoidance tendencies without the need for conscious reasoning. Conversely, propositional representations established via the linguistic pathway support more deliberate retrieval. In this case, individuals make probabilistic inferences about a new person' s characteristics based on existing group-trait beliefs—for instance, judging the credibility of the statement that "a member of this group likely possesses a certain trait." The retrieval of associative knowledge is not merely a simple invocation of a fixed group mean; rather, it relies on a more sophisticated structure of social knowledge.

Frolichs et al. (2022) found that when people apply existing knowledge to new

individuals, they utilize two structured strategies: first, a reference point strategy, where the new individual is compared to the "average" person of their group, using the group stereotype as a starting point for adjustment; and second, a granularity strategy, which utilizes the correlational structure between personality traits to perform cross-trait generalization (e.g., inferring "talkative" from "extroverted") (Frolichs et al., 2022). A substantial body of research has demonstrated that people can leverage the internal relational structure of social knowledge for generalization. Examples include inferring occupational competence from impressions of ability (Hackel, Mende-Siedlecki, et al., 2022), predicting cooperative behavior from impressions of honesty (Bellucci et al., 2019; Rösler et al., 2025), and deducing cross-situational social strategies from underlying motives (van Baar et al., 2022).

Social generalization is not a unidirectional process; it encompasses both deductive inference from the group to the individual ("He belongs to Group A, so he is likely friendly") and inductive inference from the individual to the group ("This member of Group A is friendly, so Group A is likely friendly"). Social concepts play a critical role in cognitive simplification throughout this bidirectional process: abstract social labels (such as occupation or race) compress complex, multi-dimensional social information into concise signals, thereby significantly reducing the computational costs of social decision-making (Hackel et al., 2024; Hackel & Kalkstein, 2023).

4 总结

This paper focuses on the two core processes of social learning and social generalization, reviewing research progress within the computational frameworks of reinforcement learning and Bayesian theory in the study of stereotypes. It also provides a preliminary overview of the neural substrates underlying each stage [Figure 1: see original paper]. At the level of social learning, the discovery of a "specific group"relies on Bayesian structural learning, with neural foundations involving regions such as the right anterior insula and the anterior cingulate cortex. The acquisition of "possessing a specific trait" is achieved through two pathways: the experiential pathway and the linguistic pathway. The experiential pathway establishes associative representations through reinforcement learning mechanisms, primarily involving value evaluation and emotional learning circuits such as the amygdala, striatum, and ventromedial prefrontal cortex (vmPFC). The linguistic pathway carries the formation and transmission of propositional representations, which may be associated with brain regions involved in social reasoning and mentalizing (dorsomedial prefrontal cortex, temporoparietal junction) as well as social knowledge storage and integration (anterior temporal lobe); however, direct research in this area remains scarce. Regarding the update and maintenance of associations, prediction error serves as a critical driver for updating, yet prior biases and asymmetric learning rates cause the updating process to favor the maintenance of existing beliefs. Furthermore, the exploration-exploitation dilemma suggests that biased information sampling is a significant

reason for the persistence of stereotypes. At the level of social generalization, the brain matches new individuals to known group categories through perceptual similarity (relying on early processing pathways such as the visual cortex and amygdala) and functional similarity (relying on high-level conceptual systems such as the anterior temporal lobe or orbitofrontal cortex).

These two pathways are integrated under the regulation of the dorsomedial prefrontal cortex (dmPFC). Current research still faces several limitations. First, most studies treat group discovery and value learning as separate entities; it remains unclear how the brain integrates these two types of computations during a single learning process. Bayesian Reinforcement Learning (Bayesian RL) may provide a more unified descriptive framework. Second, the ecological validity of existing computational models needs improvement. Simplified laboratory paradigms struggle to fully capture the complexity of real-world social interactions. The simultaneous input of multi-dimensional information, the interference of emotional factors, and the constraints of social norms in reality have not been sufficiently addressed in current models [**?**, **?**]. Additionally, inter-individual variation in model parameters and its sources (e.g., cognitive ability, cultural background) require further investigation [**?**, **?**]. Third, while the distinction between associative and propositional representations is conceptually clear, the interaction mechanisms between them in actual cognitive processes —particularly the conditions for the transformation from association to proposition and the boundaries of propositional inhibition of associations—require more refined experimental paradigms to be revealed. Fourth, there is a lack of research on the computational mechanisms and neural substrates of deductive inference (from group to individual) and inductive inference (from individual to group) during the generalization process.

dACC/MCC Social structural learning: P(latent group structure | observed data)

Associative representation: V(group)

dmPFC

Striatum

Propositional Representation: P(Trait | Group) Group-Trait Association

vmPFC Amygdala

dmPFC

Visual Cortex

Linking Retrieval and Application

Introduction

In the current landscape of artificial intelligence, the integration of information retrieval with downstream applications has become a pivotal area of research. As large-scale models continue to evolve, the ability to effectively bridge the

gap between vast data repositories and practical task execution is essential for enhancing model performance and reliability. This section explores the mechanisms and methodologies involved in linking retrieval systems with various application domains.

1.1 Retrieval-Augmented Generation (RAG)

One of the most prominent applications of linking retrieval to practical use is Retrieval-Augmented Generation (RAG). By combining the generative capabilities of large language models (LLMs) with external knowledge bases, RAG systems can mitigate common issues such as "hallucinations" and outdated training data. In this framework, the retrieval component identifies relevant documents or data points, which are then provided as context to the generative model to produce more accurate and grounded responses.

1.2 Knowledge Integration and Application

The process of linking retrieval to application extends beyond simple text generation. It involves the sophisticated integration of structured and unstructured knowledge into specialized workflows. For instance, in scientific research, retrieval systems can extract specific experimental parameters or chemical properties from massive datasets, which are then directly applied to predictive modeling or simulation tasks. This seamless transition from data acquisition to application ensures that the most relevant information is utilized in decision-making processes.

1.3 Challenges in Retrieval-Application Linkage

Despite the potential benefits, several challenges remain in effectively linking retrieval with application. These include:

- **Latency and Efficiency**: Real-time applications require low-latency retrieval to maintain a fluid user experience. Optimizing the search and ranking algorithms is critical to reducing the time between a query and its application.
- **Contextual Relevance**: Ensuring that the retrieved information is not only accurate but also contextually appropriate for the specific application task is a significant hurdle. This requires advanced semantic understanding and ranking mechanisms.
- **Data Privacy and Security**: When linking retrieval systems to sensitive applications, such as in healthcare or finance, maintaining data privacy and ensuring secure access to information is paramount.

1.4 Future Directions

The future of linking retrieval and application lies in the development of more autonomous and adaptive systems. We anticipate a shift toward "agentic"

retrieval, where AI agents can independently determine when to retrieve information, which sources to query, and how to best apply the findings to achieve a given objective. Furthermore, the integration of multi-modal

OFC Amygdala

"Group discovery" corresponds to the Bayesian structure learning process (indicated by the red box). In this process, the brain infers latent group structures from social observation data—specifically by calculating P(latent group structure | observed data). This involves the anterior insula (AI), which encodes abstract latent social structures, and the dorsal anterior cingulate cortex/midcingulate cortex (dACC/MCC). The "group-trait association" component illustrates a dual-pathway mechanism for establishing associations and their dual representations. Associative representations are established via the experiential pathway (including conditioning, direct interaction, and observational learning) through reinforcement learning (blue box), where V(group) serves as the formal description of this representation. Key brain regions involved include the amygdala, which supports affective association learning in evaluative conditioning, as well as the striatum and ventromedial prefrontal cortex (vmPFC), which encode both social and material value signals. Propositional representations are primarily carried by the linguistic pathway (green dashed box), and their probabilistic structure can be described by P(trait | group).

Candidate neural substrates for these processes likely involve brain regions associated with belief attribution and social reasoning, such as the temporoparietal junction (TPJ) and the superior temporal sulcus (STS). Additionally, the dorsomedial prefrontal cortex (dmPFC) may participate in the integration of beliefs at the propositional level. It should be noted that the localization of brain regions for propositional representation is based on theoretical speculation derived from literature on mentalizing and Bayesian inference in social cognition; this remains to be verified through direct empirical evidence. "Association retrieval and application" illustrates the mechanisms of group identification within social generalization. The perceptual generalization pathway (orange box) relies on the sharpening of tuning curves during early processing stages in the visual cortex, as well as the amygdala' s automated assessment of threatening faces. The functional generalization pathway (purple box) depends on high-level conceptual systems, such as the anterior temporal lobe (ATL) and the orbitofrontal cortex (OFC). The bidirectional circular arrows represent the dual nature of social generalization: it encompasses both deductive inference from the group to the individual and inductive inference from the individual to the group. Together, these processes constitute a continuously updating cycle of social cognition.

5 未来展望

Based on current limitations and recent advances in computational cognitive science, future research can advance the understanding of stereotyping mechanisms from the following three perspectives, thereby constructing a more integrated theoretical framework.

First, the role of relational cues in the formation and generalization of stereotypes warrants in-depth investigation. Existing research on generalization primarily focuses on feature similarity; however, relational cues in social life—such as an individual' s position and connections within a social network—may represent another critical pathway for generalization (Wang et al., 2025). Studies have demonstrated that the brain can encode transitive relationships within social networks and use this information to perform trust evaluations or trait inferences regarding strangers (Son et al., 2021, 2023). Neural evidence further indicates that individuals with similar relational values are encoded more similarly in the brain (Babür et al., 2024). Future research could design experimental paradigms incorporating social network information to

examine how relational and feature-based cues interact and mutually regulate one another to influence the generalization of stereotypes. For instance, by manipulating participants' positions and connection patterns within virtual social networks, researchers could test whether generalization based on network structure operates independently of generalization based on feature similarity.

Second, the concept of social cognitive maps may provide a theoretical framework for understanding the representation and integration of multi-cue information. Existing research has found neural evidence of social cognitive maps within two-dimensional trait spaces, such as generosity-competence (Gao et al., 2026) and morality-competence (Liu et al., 2025), suggesting that the hippocampal-entorhinal cortex can encode social information in a manner analogous to spatial navigation (Liang et al., 2024; Park et al., 2020; Schafer & Schiller, 2018). However, these studies have primarily focused on individual trait evaluations and have not yet addressed group-level representations. From this perspective, the social learning of stereotypes can be viewed as a process of marking group positions on a map. Bayesian structure learning determines the distribution and boundaries of groups, reinforcement learning assigns value signals to each location, and generalization involves inferring the attributes of a new individual based on their coordinates on the map (Tavares et al., 2015). Future work could combine group-level learning tasks with high-resolution neuroimaging

to investigate whether group-trait associations are encoded spatially within the hippocampal-entorhinal cortex, and to test the integration mechanisms of different generalization cues within this map space.

Third, Large Language Models (LLMs) provide a new research tool for elucidating how linguistic transmission shapes and solidifies stereotypes. Language is the primary pathway for the formation of propositional representations, and

LLMs can serve as computational simulators of this pathway. By constructing multi-agent interaction networks based on LLMs, researchers can manipulate parameters such as network structure, initial bias intensity, and transmission rules to observe how minor biases are amplified and solidified into group consensus under specific conditions (Gelpí et al., 2025; Stewart & Raihani, 2023). Combining LLM simulations with neuroscience techniques—for example, by comparing the internal representations of LLMs with stereotype encoding patterns in human neural data—will help validate the contribution of linguistic transmission to stereotype formation at both the computational and neural levels. However, one must remain cautious of the limitations of LLMs; they reflect the statistical characteristics of a corpus rather than human cognitive processes, and LLM outputs should not be equated with human cognition.

In summary, this paper attempts to provide a coherent explanatory framework from a computational cognitive perspective for the seemingly disparate phenomena in stereotype research. With the cross-disciplinary integration of computational modeling, neuroimaging, and large language models, the key for future research lies in bridging the current framework—which remains largely at the level of algorithmic description—with neural implementation, and testing its predictions within social contexts that possess greater ecological validity.

参考文献

Abstract and Generalization of Social Decision-Making Information

Abstract

In complex social environments, individuals must not only learn from direct experience but also extract abstract rules and structures from social information to guide behavior in novel situations. This process, known as the abstraction and generalization of social decision-making information, is fundamental to human social intelligence. This paper reviews recent advances in cognitive neuroscience and computational modeling regarding how the brain represents social hierarchies, social networks, and social norms. We discuss how the brain utilizes cognitive maps—specifically those supported by the hippocampal-entorhinal system —to organize social information into structured representations. Furthermore, we examine the role of the prefrontal cortex in generalizing these abstract structures to facilitate flexible decision-making across different social contexts. By integrating perspectives from reinforcement learning and representation learning, this review provides a comprehensive framework for understanding the neural mechanisms underlying social abstraction and generalization.

---

1. Introduction

Human social life is characterized by its high degree of complexity and fluidity. To navigate this landscape effectively, individuals cannot rely solely on memorizing specific interactions; instead, they must possess the ability to abstract general principles from limited observations and generalize these principles to unfamiliar social scenarios. For instance, understanding the latent power dynamics within a new professional organization often requires generalizing prior knowledge about social hierarchies [**?**].

Recent research in cognitive science and machine learning has begun to converge on the idea that the brain constructs "cognitive maps" of social spaces. These maps allow for the relational organization of social entities, enabling inferences that go beyond direct experience. This review aims to synthesize current findings on how social information is transformed from raw sensory input into abstract representations and how these representations are deployed to guide social decision-making.

2. The Representation of Social Structures

2.1 Social Hierarchies and Linear Abstraction

Social hierarchy is one of the most ubiquitous structures in human and animal societies. Research suggests that the brain represents social status not just as a collection of individual traits, but as positions along a continuous, abstract dimension. Studies using functional Magnetic Resonance Imaging (fMRI) have demonstrated that the hippocampus and the medial prefrontal cortex (mPFC) are involved in encoding the relative ranks of individuals within a group, even when those ranks must be inferred through transitive inference (e.g., if A $>$ B and B $>$ C, then A $>$ C).

2.2 Social Networks and Multidimensional Spaces

Beyond linear hierarchies, social relationships form complex networks.

Allidina, S., & Cunningham, W. A. (2021). Avoidance begets avoidance: A computational account of negative stereotype persistence. Journal of Experimental Psychology: General, 150(10), 2078–2099. https://doi.org/10.1037/xge0001037 Allidina, S., & Cunningham, W. A. (2023). Motivated categories: Social structures shape the construction of social categories through attentional mechanisms. Personality and Social Psychology Review, 27(4), 393–413. https://doi.org/10.1177/10888683231172255 Allport, G. W. (1954). The nature of prejudice. Addison-Wesley.

Amodio, D. M. (2019). Social cognition 2.0: An interactive memory systems account. Trends in Cognitive Sciences, 23(1), 21–33. https://doi.org/10.1016/j.tics.2018.10.002

Amodio, D. M. (2025). A learning and memory account of impression formation and updating. Nature Reviews Psychology, 4(6), 417–432.

https://doi.org/10.1038/s44159-025-00445-x Amodio, D. M., & Cikara, M. (2021). The social neuroscience of prejudice. Annual Review of Psychology, 72, 439–469. https://doi.org/10.1146/annurev-psych-010419-050928 Babür, B. G., Leong, Y. C., Pan, C. X., & Hackel, L. M. (2024). Neural responses to social rejection reflect dissociable learning about relational value and reward. Proceedings of the National Academy of Sciences, 121(49), e2400022121. https://doi.org/10.1073/pnas.2400022121 Bai, X., Fiske, S. T., & Griffiths, T. L. (2022). Globally inaccurate stereotypes can result from locally adaptive exploration.

Psychological Science, 33(5), 671–684. https://doi.org/10.1177/09567976211045929 Bai, X., Griffiths, T. L., & Fiske, S. T. (2025). Costly exploration produces stereotypes with dimensions of warmth and competence. Journal of Experimental Psychology. General, 154(2), 347–357. https://doi.org/10.1037/xge0001694 Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. Proceedings of the National Academy of Sciences, 122(8), e2416228122. https://doi.org/10.1073/pnas.2416228122 Bao, H.-W.-S. (2024). The Fill-Mask Association Test (FMAT): Measuring propositions in natural language. Journal of Personality and Social Psychology, 127(3), 537–561. https://doi.org/10.1037/pspa0000396 Bellucci, G., Molter, F., & Park, S. Q. (2019). Neural representations of honesty predict future trust behavior. Nature Communications, 10(1), 5184. https://doi.org/10.1038/s41467-019-13261-8 Beukeboom, C. J. (2025). Linguistic stereotyping in natural language: How and when we generalize in communication about people. Atlantic Journal of Communication, 33(5), 750–765. https://doi.org/10.1080/15456870.2025.2525799 Beukeboom, C. J., & Burgers, C. B. C. (2019). How stereotypes become shared knowledge: An integrative review on the role of biased language use in communication about categorized individuals. Review of Communication Research, 7, 1–37. https://doi.org/10.12840/issn.2255-4165.017 Bjornsdottir, R. T., Hensel, L. B., Zhan, J., Garrod, O. G. B., Schyns, P. G., & Jack, R. E. (2024). Social class perception is driven by stereotype-related facial features. Journal of Experimental Psychology. General, 153(3), 742–753. https://doi.org/10.1037/xge0001519 Cikara, M., Van Bavel, J. J., Ingbretsen, Z. A., & Lau, T. (2017). Decoding "us" and "them" : Neural representations of generalized group concepts. Journal of Experimental Psychology. General, 146(5), 621–631. https://doi.org/10.1037/xge0000287 Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering and generalizing task-set structure. Psychological Review, 120(1), 190–229. https://doi.org/10.1037/a0030852 Collins, K. A., & Clément, R. (2012). Language and prejudice: Direct and moderated effects. Journal of Language and

Social Psychology, 31(4), 376–396. https://doi.org/10.1177/0261927X12446611 Cushman, F. (2024). Computational social psychology. Annual Review of Psychology, 75, 625–652. https://doi.org/10.1146/annurev-psych-021323-040420 De Houwer, J., Dessel, P. V., & Moran, T. (2021). Attitudes as

propositional representations. Trends in Cognitive Sciences, 25(10), 870–882. https://doi.org/10.1016/j.tics.2021.07.003 De Houwer, J., Thomas, S., & Baeyens, F. (2001). Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. Psychological Bulletin, 127(6), 853–869. https://doi.org/10.1037/0033-2909.127.6.853 Eckstein, M. K., Wilbrecht, L., & Collins, A. G. (2021). What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. Current Opinion in Behavioral Sciences, Value Based DecisionMaking, 41, 128–137. https://doi.org/10.1016/j.cobeha.2021.06.004 Falbén, J. K., Golubickis, M., Tsamadi, D., Persson, L. M., & Macrae, C. N. (2023). The power of the unexpected:

Prediction errors enhance stereotype-based learning. Cognition, 235, 105386.

https://doi.org/10.1016/j.cognition.2023.105386 FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. Proceedings of the National Academy of Sciences, 115(7), E1690–E1697. https://doi.org/10.1073/pnas.1715227115 FeldmanHall, O., & Nassar, M. R. (2021). The computational challenge of social learning. Trends in Cognitive Sciences, 25(12), 1045–1057. https://doi.org/10.1016/j.tics.2021.09.002 Ferrari, C., Lega, C., Vernice, M., Tamietto, M., Mende-Siedlecki, P., Vecchi, T., Todorov, A., & Cattaneo, Z. (2016). The dorsomedial prefrontal cortex plays a causal role in integrating social impressions from faces and verbal descriptions. Cerebral Cortex, 26(1), 156–165. https://doi.org/10.1093/cercor/bhu186 Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), The handbook of social psychology (4th ed., Vols. 1–2, pp. 357–411). McGraw-Hill.

Fiske, S. T., & Taylor, S. E. (1991). Social cognition. McGraw-Hill.

Frank, S. A. (2018). Measurement invariance explains the universal law of generalization for psychological perception.

Proceedings of the National Academy of Sciences, 115(39), 9803–9806. https://doi.org/10.1073/pnas.1809787115 Freeman, J. B., & Johnson, K. L. (2016). More than meets the eye: Split-second social perception. Trends in Cognitive Sciences, 20(5), 362–374. https://doi.org/10.1016/j.tics.2016.03.003 Frolichs, K. M. M., Rosenblau, G., & Korn, C. W. (2022). Incorporating social knowledge structures into computational models. Nature Communications, 13(1), 6205. https://doi.org/10.1038/s41467-022-33418-2 Gao, T., Deng, Y., & Han, S. (2026). Construction of individual-specific social cognitive maps in the human brain. Cell Reports, 45, 116890. https://doi.org/10.1016/j.celrep.2025.116890 Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115 Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation:

An integrative review of implicit and explicit attitude change. Psychological Bulletin, 132(5), 692–731. https://doi.org/10.1037/00332909.132.5.692 Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. In J. M. Olson & M. P. Zanna (Eds.), Advances in experimental social psychology (Vol. 44, pp. 59–127). Academic Press. https://doi.org/10.1016/B978-0-12-385522-0.00002-0 Gelman, S. A., Taylor, M. G., & Nguyen, S. P. (2004). Mother-child conversations about gender: Understanding the acquisition of essentialist beliefs. Monographs of the Society for Research in Child Development, 69(1), vii, 116–127. https://doi.org/10.1111/j.1540-5834.2004.06901001.x

Gelpí, R. A., Tang, Y., Jackson, E. C., & Cunningham, W. A. (2025). Social coordination perpetuates stereotypic expectations and behaviors across generations in deep multiagent reinforcement learning. PNAS Nexus, 4(3), pgaf076. https://doi.org/10.1093/pnasnexus/pgaf076 Gershman, S. J., & Cikara, M. (2020). Social-structure learning. Current Directions in Psychological Science, 29(5), 460–466. https://doi.org/10.1177/0963721420924481 Gershman, S. J., & Cikara, M. (2023). Structure learning principles of stereotype change. Psychonomic Bulletin & Review, 30(4), 1273–1293. https://doi.org/10.3758/s13423-023-02252-y Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. Current Opinion in Neurobiology, Cognitive Neuroscience, 20(2), 251–256. https://doi.org/10.1016/j.conb.2010.02.008 Gershman, S. J., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. Current Opinion in Behavioral Sciences, 5, 43–50. https://doi.org/10.1016/j.cobeha.2015.07.007 Gershman, S. J., Pouncy, H. T., & Gweon, H. (2017). Learning the structure of social influence. Cognitive Science, 41(S3), 545–575. https://doi.org/10.1111/cogs.12480

Golubickis, M., Persson, L. M., Falbén, J. K., Seow, S. H., Jalalian, P., Sharma, Y., Ivanova, M., & Macrae, C. N. (2024).

Facial misfits accelerate stereotype-based associative learning. Scientific Reports, 14(1), 19320. https://doi.org/10.1038/s41598-024-67770-8 Gustafsson Sendén, M., Eagly, A., & Sczesny, S. (2020). Of caring nurses and assertive police officers: Social role information overrides gender stereotypes in linguistic behavior. Social Psychological and Personality Science, 11(6), 743–751. https://doi.org/10.1177/1948550619876636 Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. Nature Neuroscience, 18(9), 1233–1235. https://doi.org/10.1038/nn.4080 Hackel, L. M., & Kalkstein, D. A. (2023). Social concepts simplify complex reinforcement learning. Psychological Science, 34(9), 968–983. https://doi.org/10.1177/09567976231180587 Hackel, L. M., Kalkstein, D. A., & Mende-Siedlecki, P. (2024). Simplifying social learning. Trends in Cognitive Sciences, 28(5), 428–440. https://doi.org/10.1016/j.tics.2024.01.004 Hackel, L. M., Kogon, D., Amodio, D. M., & Wood, W. (2022). Group value learned through interactions with members: A reinforcement learning account. Journal of Experimental Social Psy-

chology, 99, 104267. https://doi.org/10.1016/j.jesp.2021.104267 Hackel, L. M., Mende-Siedlecki, P., & Amodio, D. M. (2020). Reinforcement learning in social interaction: The distinguishing role of trait inference. Journal of Experimental Social Psychology, 88, 103948. https://doi.org/10.1016/j.jesp.2019.103948 Hackel, L. M., Mende-Siedlecki, P., Loken, S., & Amodio, D. M. (2022). Context-dependent learning in social interaction:

Trait impressions support flexible social choices. Journal of Personality and Social Psychology, 123(4), 655–675. https://doi.org/10.1037/pspa0000296 Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nature Computational Science, 3(10), 833–838. https://doi.org/10.1038/s43588-023-00527-x Hamilton, D. L., & Rose, T. L. (1980). Illusory correlation and the maintenance of stereotypic beliefs. Journal of Personality and Social Psychology, 39(5), 832–845. https://doi.org/10.1037/0022-3514.39.5.832 Hedrich, N. L., Schulz, E., Hall-McMaster, S., & Schuck, N. W. (2024). An inductive bias for slowly changing features in human reinforcement learning. PLOS Computational Biology, 20(11), e1012568. https://doi.org/10.1371/journal.pcbi.1012568 Hu, Y., & O' Toole, A. J. (2023). First impressions: Integrating faces and bodies in personality trait perception. Cognition, 231, 105309. https://doi.org/10.1016/j.cognition.2022.105309 Hutchison, J., Cunningham, S. J., Slessor, G., Urquhart, J., Smith, K., & Martin, D. (2018). Context and perceptual salience

influence the formation of novel stereotypes via cumulative cultural evolution. Cognitive Science, 42(Suppl. 1), 186–212. https://doi.org/10.1111

Note: Figure translations are in progress. See original paper for figures.