刻板印象的计算认知机制：社会学习与泛化

王滢洁[1]，张洳源[2, 3, 4*]

1 上海交通大学心理学院和上海交通大学医学院附属精神卫生中心国家精神健康研究中心脑健康研究院，上海 200030

2 北京大学心理与认知科学学院和行为与心理健康北京市重点实验室，北京 100871

3 北京大学 IDG 麦戈文脑科学研究所，北京 100871

4 北京大学机器感知与智能教育部重点实验室，北京 100871

* 通讯作者

Ying-Jie WANG[1], Ru-Yuan ZHANG[2, 3, 4*]

[1]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.

[2] School of Psychological and Cognitive Sciences and Beijng Key Laboratory of Behavior and Mental Health, Peking University, Beijing 100871, China.

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

[4] Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, China.

摘　要 刻板印象是关于“某群体具有某特质”的概括性信念，深刻影响着人际互动与群体关系，因此理解其形成与维持的认知机制具有重要意义。然而，传统理论多停留在描述层面，对刻板印象从习得到应用的过程缺乏整合性的机制解释。本文围绕社会学习与社会泛化两个环节，结合强化学习与贝叶斯理论，对刻板印象的计算认知机制进行梳理。在社会学习层面，首先阐述贝叶斯结构学习如何推断潜在的群体结构，并区分群体-特质联结建立的两条通路：经验通路主要通过强化学习机制形成自动化的联想表征，语言通路主要通过贝叶斯推断传递概率性的命题表征。在联结的更新与固化过程中，预测误差是驱动更新的重要动力，但先验偏差和不对称学习率使更新倾向于维持已有信念；同时，探索-利用困境表明，有偏的信息采样是刻板印象顽固性的重要原因。在社会泛化层面，分析面对新个体时人们基于知觉相似性与功能相似性的群体归类过程，并进一步讨论已学联结知识的检索机制。最后，本文提出三个未来研究方向：将关系线索作为特征相似性之外的泛化路径，发展社会认知地图作为多线索表征的整合框架，以及利用大语言模型模拟语言传播中的刻板印象固化过程。

关键词 刻板印象，社会泛化，强化学习，贝叶斯理论，计算建模

1. 引言

当我们提到外科医生时，脑海中往往浮现出一个冷静理性的男性形象；而提到幼儿园老师时，则更容易联想到温柔耐心的女性形象。这种在日常生活中无处不在、自动化的快速判断，正是刻板印象(Stereotype)的典型体现。刻板印象通常被定义为社会成员对特定群体所持有的、一套过度概括且相对固化的信念 (Allport, 1954)。它作为一种普遍存在的社会认知现象，深刻影响着人际互动、群体关系乃至社会结构(Fiske, 1998)。

自社会心理学诞生以来，探究刻板印象的成因与机制始终是其核心议题之一。早期研究多聚焦于动机与情感层面。无论是威权人格 (Authoritarian Personality)的心理动力学解释 (Adorno, 1950)、现实群体冲突 (Realistic Group Conflict)中资源竞争视角 (Sherif, 1936)，还是社会认同理论(Social Identity Theory)对维护群体自尊的强调 (Sherif & Sherif, 1953; Tajfel & Turner, 2004)，这些经典理论都将刻板印象视为由强烈情感、内在需求或群体动态所驱动的特定现象。随后，社会认知理论的兴起将视角转向信息加工过程。研究者提出，刻板印象本质上是，认知资源有限时，大脑为简化复杂社会信息加工而采取的一种认知捷径(cognitive shortcut)或启发式(heuristics)策略(Hamilton & Rose, 1980; Kahneman & Tversky, 1973; Tversky & Kahneman, 1974)。例如，个体会自动依据种族、性别等线索进行社会分类，并调用预存图式快速进行推断 (Allidina & Cunningham, 2023; Fiske & Taylor, 1991; Sherman et al., 2009; Taylor, 1981)。

尽管相关理论极其丰富，但是多停留在描述层面。如要真正理解刻板印象的普遍性和顽固性，并据此开发有效的干预措施，就必须从描述“是什么”转向解释“如何”与“为何”，深入探究其认知计算机制。近年来，强化学习和贝叶斯理论等计算框架被引入社会学习与决策领域，借助算法模型描述个体如何习得社会知识并据此做出判断(Amodio, 2025; Cushman, 2024; Gershman & Cikara, 2023)，其模型预测也已得到神经影像学研究的支持(Lockwood & Klein-Flügge, 2021; Mahmoodi & Rushworth, 2026; Olsson et al., 2020; Zhang et al., 2020)。在这一框架下，刻板印象的形成与应用可以被分解为两个核心环节：社会学习(social learning)，关注个体如何习得关于群体的知识；社会泛化(social generalization)，关注个体如何将有限经验推广到新情境或未知个体。然而，现有研究多从单一视角出发，对上述全过程机制的整合仍显不足。基于此，本文围绕这两个环节，结合强化学习与贝叶斯理论，系统梳理刻板印象形成、更新与泛化应用的认知计算机制。

2. 社会学习：刻板印象的形成

刻板印象的核心在于“某群体具有某特质”这一信念结构，其形成可以分解为两个基本问题：第一，“某群体”从何而来，即大脑如何从复杂的社会观察中辨别出哪些个体属于同一类别；第二，“具有某特质”如何习得，即个体如何学习到群体与特质或价值之间的联结。在此基础上，联结如何在新证据面前更新或固化，则决定了刻板印象的可塑性与顽固性。

2.1 群体的发现与构建

刻板印象是对某一个群体的概括，因此理解人们如何辨别社会群体、表征社会结构显得尤为重要。现实中的群体边界往往不能直接观察到，而需要从个体的行为模式、社会互动和共享特征中推断得出。

社会结构学习模型(Social Structure Learning Model)提出，大脑通过贝叶斯推断从可观察的社会数据（如人们的投票行为、着装偏好、社交互动）中发现不可见的潜在群体结构，即估计潜在群体的后验概率 P(潜在群体结构 | 观察数据) (Gershman & Cikara, 2020)。这一过程超越了简单的二元相似性判断（判断个体与自己是否相似），而是涉及更复杂的多元关系推理。例如，在“敌人的敌人是朋友”或“朋友的朋友是朋友”这类三元关系中，个体 A 与个体 B 可能没有任何直接的特征重叠，但因为二者都与个体 C 存在某种特定关系，A 仍可能被推断与 B 属于同一群体(Gershman et al., 2015; Gershman & Niv, 2010; Pietraszewski, 2022)。

Gershman 等 (2017)的实验为这一机制提供了直接证据(Gershman et al., 2017)。在电影海报选择任务中，研究者控制了代理人 A 和 B 与参与者的二元选择重叠率均为 50%，使两者从二元相似性角度看等价。关键操纵在于第三方 C 的选择模式：当 C 同时与 A 和参与者高度一致（75%重叠）时，暗示三者属于同一潜在群体；当 C 与 A 一致但与参与者不一致时，暗示 A 和 C 构成另一群体。结果显示，参与者的选择显著受潜在群体结构影响，贝叶斯结构学习模型的拟合优度显著高于二元相似性模型。Lau 等进一步发现，这种结构学习不仅指导行为选择，还导致情感和认知的泛化(Lau et al., 2018) 。研究利用类似范式，但参与者需要选择政治偏好(如，是否赞成死刑)，实验后要求参与者对各个代理人进行社会评价。结果显示，如果模型推断某代理人与参与者属于同一潜在群体，参与者不仅更有可能在未知议题上跟随该人，还会给该人打出更高的道德和受欢迎度评分。基于模型的功能性磁共振成像(fMRI)研究揭示了相关的神经分离，二元相似性与内侧前额叶皮层/膝前部前扣带皮层(medial prefrontal cortex, mPFC/pregenual anterior cingulate, pgACC)的活动相关，而潜在社会群体结构的表征则与右侧前脑岛(right anterior insula, rAI)的活动相关(Lau et al., 2020)。多体素模式分析(MVPA)研究进一步发现，背侧前扣带皮层/中扣带皮层 (dACC/MCC) 和前脑岛 (AI) 的神经模式可以准确区分内群体和外群体(Cikara et al., 2017)。这些神经证据表明，潜在群体结构的编码有其独立的神经基础，不能简化为简单的相似性计算。

2.2 群体-特质联结的建立

在发现群体之后，社会学习的另一核心任务是习得“这个群体具有什么特质或社会价值”。群体-特质联结的建立通过两条性质不同的通路实现：经验通路，依赖个体通过直接或间接的社会互动经验习得联结；语言通路，通过语言媒介接收和传递关于群体的概括性知识。

2.2.1 经验通路与联想表征

经验通路涵盖多种学习形式，其共同特征是通过个体与社会环境的直接或间接互动来建立群体-特质联结。在表征层面，这些学习形式主要建立联想表征，即群体与特质之间的自动化联结激活(Gawronski & Bodenhausen, 2006)。在计算层面，强化学习提供了对这一过程的算法描述：社会互动被视为一系列决策过程，个体通过行动获得奖励或惩罚，并利用这些反馈来调整其对群体的

价值评估 V(群体) ← V(群体) + αδ，其中 α 为学习率(learning rate)，控制新信息对已有价值估计的更新幅度；δ 为预测误差(prediction error)，即实际获得的结果与预期之间的差值 (Eckstein et al., 2021; Lockwood & Klein-Flügge, 2021; Sutton & Barto, 1998; Zhang et al., 2020)。

经验通路的三种核心学习形式在计算性质上存在重要区分(Amodio, 2025; Amodio & Cikara, 2021)。**评价性条件反射**(Evaluative Conditioning, EC)，是群体-特质联结中情感成分形成的基础 (De Houwer et al., 2001; Öhman, 2005)。个体并不需要执行任何动作，仅仅是因为某类群体成员(条件刺激, conditional stimulus, CS)反复与积极或消极的刺激(非条件刺激, unconditional stimulus, US)在时空上共现，便会自动建立两者之间的联结(Amodio & Cikara, 2021; Öhman, 2005; Olsson et al., 2005)。这种学习主要依赖于杏仁核(amygdala)，使得人们在通过互动获得具体的特质知识之前，就已经对特定群体形成了自动化的情感偏见，对应于强化学习中的无模型学习(model-free learning) (Amodio, 2019)。**直接互动学习**依赖于个体的行为决策与环境反馈，是典型的工具性学习(instrumental learning)过程(Amodio, 2025; Schultner, Stillerman, et al., 2024)。**观察学习**则涉及价值塑造(value shaping)机制，即观察者的偏好不仅受目标对象反馈的影响，还直接受行为者选择偏好的影响，无意中继承他人的有偏估计 (Schultner, Lindström, et al., 2024)。

个体通过社会互动可以产生对他人的价值评估(如回报金额)，也可以推断他人的抽象特质(如慷慨程度) (Hackel et al., 2020)。在神经层面，社会性价值(如获得他人的信任、社会地位的提升)和物质性奖赏(如金钱)可能在同一组脑区中被编码为价值信号 (Wake & Izuma, 2017)，主要包括腹内侧前额叶皮层(ventromedial prefrontal cortex, vmPFC)和纹状体(striatum) (Joiner et al., 2017; Kumaran et al., 2016)。相比于对价值的编码，对他人特质、意图的推断还涉及后扣带皮层(posterior cingulate cortex, PCC)、楔前叶(precuneus)、颞上沟(superior temporal sulcus, STS)和颞顶联合区(temporo-parietal junction, TPJ)等脑区(Bellucci et al., 2019; Hackel et al., 2015; Kobayashi et al., 2022)。

2.2.2 语言通路与命题表征

语言通路通过语言这一符号系统来传递和形成关于群体的知识，其独特之处在于主要承载命题表征，即具有真值判断功能的语义内容(De Houwer et al., 2021)。例如，“女孩很善良”是一个可以被评估为真或假的命题(proposition)，而非群体与特质词之间的简单联想激活 (Gelman et al., 2004)。作为人类特有且运用最广泛的联结习得形式 (K. A. Collins & Clément, 2012; Maass, 1999; Martin et al., 2014)，语言使人们能够超越一次性观察到的行为事件，并将信息概括到不同的个体和特定情境中，但高度概括性的交流可能偏离特定个体在具体情境的真实行为(Beukeboom, 2025)。传播链(transmission chain)研究揭示了语言传播如何催生刻板印象，最初无结构的信息经多代传递后，会自发演变为简化、有明确类别结构的刻板印象系统 (Hutchison et al., 2018)。语言传播过程中固有的信息压缩倾向，将具体行为概括为抽象特质，如将“他打人了”转述为“他很暴力”，进一步维持和强化刻板印象(Beukeboom & Burgers, 2019)。

语言通路所形成的命题表征，在计算层面可用贝叶斯框架中的条件概率进行刻画。具体而言，刻板印象可被形式化为 p(特质 | 群体)，即在给定个体属于某个群体，该个体拥有某种特征的

条件概率(McCauley & Stitt, 1978)。根据贝叶斯公式，该概率可进一步分解为 p(特质 | 群体) = p(特质) × p(群体 | 特质) / p(群体)，其中 p(特质)为特质先验（该特质在总人口中的基础概率），p(群体 | 特质)为似然（具有该特质的人属于该群体的概率），p(群体)为群体先验（该群体在总人口中的比例）。Solanki 和 Cesario (2025)的研究为这一概率结构提供了实证依据，他们设计了一套关于刻板印象的调查问卷，其中测试了 8 个社会类别(男/女，亚裔/黑人等)、每个类别 10 条特质(5 刻板/5 非刻板)的上述四个成分。结果显示，参与者直接估计的后验与理论后验呈高度相关，且高认知能力者更善于按贝叶斯规则区分信息。这表明，刻板印象的命题表征呈现出与贝叶斯框架一致的概率性结构。需要强调的是，虽然贝叶斯推断为命题表征提供了可能的计算模型，但两者在理论本质上仍有显著差异： 命题表征界定的是认知内容的表征形式，贝叶斯推断则是一种描述信息整合的计算机制。

这种概率性的命题表征，在大语言模型(Large Language Model, LLM)中也得到了印证。Bao (2024)开发的掩码填空联系测验(fill-mask association test, FMAT)提供了一种从语料层面直接量化命题表征的途径。该方法利用 BERT 语言模型的填充掩码功能，通过比较不同群体词在特质描述语境中的预测概率来计算 p(群体 | 特质)，与贝叶斯框架中的似然估计在形式上高度一致。提示，尽管 LLM 能复现人类的内隐偏见(Bai, Wang, et al., 2025; Garg et al., 2018; Hagendorff et al., 2023)，但其反映的是语料中已经存在的群体-特质联结模式，而非独立地发现和构建这些联结。

2.2.3 联想表征与命题表征的交互

经验通路与语言通路在主要的表征类型上虽各有侧重，但在实际认知过程中，两种表征会持续交互。Gawronski 和 Bodenhausen 提出的联想-命题评估模型(Associative-Propositional Evaluation Model, APE)对这一机制进行了阐述(Gawronski & Bodenhausen, 2006, 2011)。一方面，联想激活可以作为命题表征的输入，例如当个体与某群体成员互动后产生的自动化负面情感反应(联想激活)被意识捕获，进而转化为命题表征(“这个群体可能不友善”)。另一方面，命题表征可以约束联想表征，例如在获知某刻板印象在统计上不准确后，即便群体标签仍能自动激活负面效价，个体也可通过命题层面的否定来抑制行为表达。De Houwer 等 (2021)进一步提出，许多传统上归为联想过程的心理现象，本质上可能是命题性的(De Houwer et al., 2021)。

Kurdi 等 (2023)的元分析从实证角度回应了这一争论。他们发现，刺激间的反复配对（联想学习的典型输入）和明确的语义关系（如“A 帮助了 B”，命题学习的典型输入）都能够改变人们对目标对象的评价，但两者在影响外显态度和内隐态度上的效力有所不同(Kurdi et al., 2023)。这一结果提示，在实际认知过程中，联想表征与命题表征的区分边界比传统双过程理论所假设的更加模糊，刻板印象可能同时受到两种学习机制的塑造。

2.3 联结的更新与固化

群体-特质联结一旦建立，并非一成不变，但也绝非轻易改变。以下从三个方面讨论联结的变化：预测误差如何驱动更新，已有信念如何抵抗更新，以及探索-利用困境如何从信息采样层面维持刻板印象。

2.3.1 预测误差驱动的更新

强化学习的核心思想是，个体通过预测误差(Prediction Error, PE)，即实际结果与预期之间的差值，来调整已有联结的权重 (Sutton & Barto, 1998)。在刻板印象的语境下，反刻板印象信息产生的高预测误差，可能是促进刻板印象更新的重要动力。Falbén 等采用概率选择任务，让参与者面对配对的男女面孔，通过试错学习确定每对中谁更可能喜欢芭蕾舞或拳击(性别刻板印象相关的兴趣爱好)(Falbén et al., 2023)。计算模型(强化学习漂移扩散模型, RL-DDM)分析显示，相比与刻板印象一致(女性+芭蕾舞/男性+拳击)的信息，参与者在学习反刻板印象信息时表现出更快的学习速率。Golubickis 等进一步发现，面部特征与刻板印象的匹配度显著影响学习速率(Golubickis et al., 2024)。在反刻板印象学习情境中，高性别典型性面孔(如非常女性化的女性从事建筑工作)学习更快；在刻板印象学习情境中，低性别典型性面孔(如不太男性化的男性从事建筑工作)学习更快。fMRI 研究发现，印象更新过程激活了包括前外侧前额叶皮层(rostrolateral PFC)、颞上沟、右侧顶下小叶(right inferior parietal lobule, rIPL)和后扣带皮层(PCC)等与冲突监控和社会推理功能相关的脑区网络 (Mende-Siedlecki, Cai, et al., 2013)。

然而，当面对与刻板印象不符的个体时，大脑并非总是更新联结知识，也可以灵活调整对群体结构的知觉，主要通过亚型化和子群化两种机制实现 (Martinez et al., 2025; Maurer et al., 1995; Richards & Hewstone, 2001)。亚型化(Subtyping)指，将极端偏离群体典型特征的个体分离为“例外”，形成一个只包含其自身的新类别，从而保护原有刻板印象不受冲击。子群化(Subgrouping)则指，当多个个体共同表现出一种偏离模式时，大脑在原有大类别下创建新的嵌套子群体（如在“女性”下形成“女权主义者”子群），在保留原有刻板印象的同时对其进行细化。贝叶斯模型模拟显示，反证信息适度分散且偏离程度中等(如，反常行为分散在三位代理人上)时，最容易触发子群化，从而推动刻板印象的修正；而集中且极端的反例(如，所有反常行为都集中在 A 身上)则最容易触发亚型化，反而导致原有偏见的固化(Gershman & Cikara, 2023)。这一计算结论为刻板印象的社会干预提供了反直觉的建议：与其树立几个完美的、极端的反刻板印象模范，不如呈现大量多样化的群体成员形象，更能有效地改变隐性认知结构。

2.3.2 先验偏差对更新的抵抗

即便预测误差能够驱动更新，预先存在的刻板印象仍会从多个环节扭曲这一过程，使其倾向于维持而非修正已有信念。计算模型分析揭示了两种主要机制 (Schultner, Stillerman, et al., 2024; Traast et al., 2024, 2025)。第一，刻板印象充当有偏的先验信念(prior belief)，为不同群体设置了不同的初始奖赏预期。例如，一个持有负面刻板印象的白人参与者，在与黑人玩家互动之初，可能会预设较低的合作(分享)概率。第二，刻板印象导致不对称的学习率(learning rate)。例如，对证实负面预期的行为学习更快，对挑战负面预期的积极行为学习更慢。两种机制共同作用，导致即使在面对完全相同的客观反馈时，学习到的内部价值表征也会出现偏差。

Hedrich 等 (2024)从更基础的层面为这种不对称性提供了解释：人类的强化学习系统偏好加工变化缓慢的特征(如人格特质、群体特性) (Hedrich et al., 2024)。相比快速波动的信息，大脑更倾向

于编码和依赖那些被感知为稳定的特征，这可能是刻板印象难以被单次反证推翻的认知基础之一。道德性刻板印象尤其如此，当参与者先接触到关于某群体的道德性刻板印象(如诚实、不可信)时，不仅设定了更极端的初始期望值，且更不愿意根据反证信息调整预期 (Rösler et al., 2025)。

2.3.3 探索-利用困境与顽固性

刻板印象为何一旦形成就极难改变？除了先验偏差和不对称学习率，一个更深层的原因在于信息采样本身的偏差—— 探索-利用困境(explore-exploit dilemma) (Bai et al., 2022)。一个追求长期收益最大化的理性主体，如果与某个群体的初次互动(即“探索”)获得足够回报，就可能会倾向于继续与该群体互动(即“利用”)，而不再去探索其他可能同样甚至更具回报性的群体。这种局部适应性探索(locally adaptive exploration)策略虽然在短期内以最小的试错成本获得稳定的收益，但长远的代价是巨大的：早期偶然的积极或消极经历，导致个体在局部较优处过早停止探索行为，从而在全局上固化了不准确的印象。

Bai 等 (2022)的研究改编了多臂赌博机范式加以验证：参与者在虚构城市中与四个社会群体进行多轮互动，参与者可以选择与不同群体成员合作以获得奖励，或在随机分配条件下被动接受配对。结果显示，在群体间实际上没有差异的情况下，自主选择的参与者更倾向于集中选择某一群体，并显著高估群体间的差异，形成了不准确的全局印象(Bai et al., 2022)。Allidina 和 Cunningham (2021)发现，对某群体持负面刻板印象的个体倾向于回避与该群体成员的互动，回避行为减少了获取反证信息的机会，形成了类似的恶性循环 (Allidina & Cunningham, 2021)。

这一框架还能解释刻板印象为何具有经典的热情-能力二维结构 (Bai, Griffiths, et al., 2025)。研究利用情境化多臂赌博机范式，让参与者担任招聘者为四个虚拟群体的成员分配在两个维度上有所区别的工作。与随机探索相比，自主探索的参与者形成了更明显的群体分层选择，在二维空间中产生了更大的心理距离。三种降低探索成本的干预措施(探索奖励、降低奖励概率、随机限制)都有效减少了分层和刻板印象，说明当探索存在成本且决策者需要基于职业、教育背景等共享特征进行泛化时，基于特征的探索(feature-based exploration)机制会导致决策者将不同的群体在多个维度上进行分离，从而自发地重现热情-能力的二维刻板印象空间。

从计算整合的角度看，探索-利用困境可能是强化学习与贝叶斯理论的交汇点。贝叶斯框架中的先验信念影响了强化学习的探索策略：强烈的先验预期使个体倾向于选择与预期一致的互动对象，导致信息采样存在偏差；而有偏的采样反过来强化原有先验，形成自我强化的循环(Villiger, 2025)。

3 社会泛化：刻板印象的应用

社会学习解释了刻板印象如何形成，但仅有习得是不够的，刻板印象之所以能有如此大的影响，还依赖于泛化，即把这些有限经验推广到新的个体和情境中(Hackel, Kogon, et al., 2022)。泛化过程的挑战在于：面对一个从未接触过的新个体，大脑如何将其匹配到已知的群体类别，从而调用已学到的联结知识？

3.1 泛化路径

社会情境中的泛化也符合通用泛化法则(universal law of generalization)：泛化高度依赖于相似性，情境越相似，知识就越容易在它们之间迁移(Frank, 2018; Shepard, 1987; Verosky & Todorov, 2010)。在刻板印象的应用中，这种基于相似性的匹配主要通过知觉线索与功能线索两条路径来实现。

3.1.1 基于知觉线索

知觉泛化(Perceptual-based generalization)是最直观的路径，依赖于即时可得的物理特征，包括面孔特征(如肤色、面部结构、表情)、身体姿态、衣着服饰以及语音语调等 (Hu & O'Toole, 2023; Krahé et al., 2021; Todorov et al., 2015)。在计算层面，这一过程可理解为大脑将新个体的知觉特征向量与已存储的群体原型在特征空间中进行距离比较与分类。有研究显示，当参与者学习到部分人不值得信任后，面对面孔相似的陌生人会表现出不信任，且这种效应随面孔相似度降低而减弱(FeldmanHall et al., 2018)。从神经机制看，知觉泛化在视觉加工早期阶段已启动：当个体学习到某一面孔与厌恶性结果相关时，视觉皮层中神经元对面孔身份线索的调谐曲线会变得锐化，使大脑对威胁相关面孔特征更敏感，进而将厌恶反应泛化到相似面孔 (Stegmann et al., 2020)。

面孔作为最常用的知觉线索，与刻板印象内容存在系统性关联 (Sutherland & Young, 2022)。例如，被判断为社会阶层较低(“贫穷”)的面孔通常更宽、更短、面部轮廓更扁平，并伴有下撇的嘴角和更暗、更冷的肤色；而这些特征又恰好与被判断为能力低下、冷漠和不值得信任的面孔特征相重叠 (Bjornsdottir et al., 2024)。一项 fMRI 的元分析研究表明，面孔社会评估的神经计算中枢位于情感和价值评估回路：对负面面孔(如被评定为不值得信任)的评估，稳定激活双侧杏仁核(amygdala)；对正面面孔(如被评定为有吸引力、值得信任)的评估，则激活与奖赏和价值计算相关的脑区，包括内侧前额叶皮层(medial prefrontal cortex, mPFC)、内侧眶额皮层(medial orbitofrontal cortex, mOFC)以及伏隔核(nucleus accumbens, Nacc) (Mende-Siedlecki, Said, et al., 2013)。

3.1.2 基于功能线索

与知觉泛化不同，功能泛化(Functional-based generalization)依赖无法通过即时观察获得的抽象属性。在计算层面，功能相似性可以理解为通过潜变量(latent variable) 进行的间接匹配，当两个或多个个体因具有相同功能、预测相同结果或需要相同行为反应时，即使他们在知觉特征上完全不同，也能实现知识迁移(A. G. E. Collins & Frank, 2013; Shohamy & Wagner, 2008)。

社会角色(尤其是职业标签)是最普遍的功能线索。社会角色理论(Social Role Theory)指出，人们对大规模社会群体(如不同性别、种族)的刻板印象，并非源于其内在本质，而是对这些群体成员通常所扮演的社会角色的观察 (Koenig & Eagly, 2014)。例如，护士角色需关怀特质、警察角色需果断特质，因历史与社会因素，女性多从事护士、男性多从事警察，长期下来，角色特质便被泛化到对应群体，形成性别刻板印象。多项研究证实了社会角色在社会泛化中的重要作用。职业标签不仅能驱动人们对人格特质和技能的归纳，甚至能影响对权利和义务的泛化，且多数情况下其效力优先于种族、性别等线索 (Noyes et al., 2021)。例如，当社会角色信息与性别信息(如“男护士”)

同时出现时，前者(如“有爱心”，护士的角色特质)能够覆盖后者(如“有野心”，传统的男性特质)，主导人们的特质推断 (Gustafsson Sendén et al., 2020)。

3.1.3 知觉与功能线索的整合

知觉线索与功能线索并非独立运行，而是在大脑中进行着复杂的整合。Freeman 和 Johnson(2016)提出的动态交互模型(Dynamic Interactive Theory)认为，视觉知觉系统(如梭状回面孔区 Fusiform face area FFA)与高阶社会概念系统(如前颞叶 anterior temporal lobe ATL、眶额皮层 orbitofrontal cortex OFC)之间存在循环连接和重叠表征 (Freeman & Johnson, 2016; Stolier & Freeman, 2016)。例如，如果一种文化中普遍存在“黑人是敌对的”这一刻板印象，那么在眶额皮层和视觉皮层中，对黑人面孔的神经激活模式会自发地向对愤怒表情的激活模式靠拢，即便该面孔实际上是中性的。这种自上而下的调节解释了为什么刻板印象会扭曲最基本的视觉知觉(例如，更容易将黑人手中的工具误看成武器)。经颅磁刺激(TMS)研究证实，背内侧前额叶皮层(dorsomedial prefrontal cortex, dmPFC)在整合过程中起核心作用，负责将来自不同渠道的信息(如面孔和语言描述)整合成最终的社会印象 (Ferrari et al., 2016)。

3.2 联结知识的检索与应用

一旦新个体被识别为某群体成员，刻板印象的应用便转化为对已习得的联结知识的检索。正如社会学习阶段存在两条通路，泛化阶段的检索也可能沿两条路径运作。沿经验通路建立的联想表征支持快速、自动化的检索：群体标签直接激活预存的情感效价，驱动趋近或回避倾向，这一过程无需有意识的推理。沿语言通路建立的命题表征则支持更为审慎的检索：个体基于已有的群体-特质信念对新个体的特征进行概率性推断，例如判断“该群体成员可能具有某特质”的可信程度。

联结知识的检索并非简单地调用固定的群体均值，而是依赖于更为精细的社会知识结构。Frolichs 等 (2022)发现，人们在将已有知识应用于新个体时，会利用两种结构化策略：一是参考点策略，将新个体与其所在群体的平均人进行比较，以群体刻板印象作为推断的起点进行调整；二是粒度策略，利用人格特质之间的相关结构进行跨特质泛化(如从“外向”推断“健谈”)(Frolichs et al., 2022)。已有大量研究发现人们能够利用社会知识的内在关联结构进行泛化，例如从能力印象推断职业胜任力(Hackel, Mende-Siedlecki, et al., 2022)、从诚实印象推断合作行为(Bellucci et al., 2019; Rösler et al., 2025)、从潜在动机推断跨情境的社会策略(van Baar et al., 2022)。

社会泛化并非单向的，它既包含从群体到个体的演绎推断(“他属于 A 群体，所以他可能是友善的”)，也包含从个体到群体的归纳推断(“这个 A 群体成员很友善，所以 A 群体可能是友善的”)。社会概念在这一双向过程中起到了认知简化的关键作用：抽象的社会标签（如职业、种族标签）将复杂的多维度社会信息压缩为简洁的信号，极大降低了社会决策的计算成本(Hackel et al., 2024; Hackel & Kalkstein, 2023)。

4 总结

本文围绕社会学习与社会泛化两个核心环节，梳理了刻板印象研究中强化学习与贝叶斯理论两大计算框架下的研究进展，并对各环节的神经基础作了初步概述 (图 1)。在社会学习层面，“某群体”的发现依赖贝叶斯结构学习，其神经基础涉及右前脑岛、前扣带皮层等区域。“具有某特质”的习得通过经验通路与语言通路两条通路实现。其中，经验通路通过强化学习机制建立联想表征，主要涉及杏仁核、纹状体和腹内侧前额叶等价值评估和情感学习回路。语言通路承载命题表征的形成与传递，可能与社会推理和心智化加工(背内侧前额叶皮层、颞顶联合区)、社会知识存储与整合(前颞叶)的相关脑区相关，但尚缺乏直接研究。在联结的更新与维持方面，预测误差是更新的重要动力，但先验偏差和不对称学习率使更新过程倾向于维持已有信念；探索-利用困境则进一步表明有偏的信息采样是刻板印象顽固性的重要原因。在社会泛化层面，大脑通过知觉相似性(依赖视觉皮层和杏仁核等早期加工通路)和功能相似性(依赖前颞叶或眶额皮层等高级概念系统)将新个体匹配到已知群体类别，两条路径在背内侧前额叶的调控下实现整合。

当前研究仍存在若干局限。第一，多数研究将群体发现与价值学习割裂开来探讨，大脑如何在单一学习过程中整合这两类计算仍不明确，贝叶斯强化学习(Bayesian RL)可能提供更统一的描述框架。第二，现有计算模型的生态效度有待提升，简化的实验室范式难以完全捕捉现实社会互动的复杂性，现实中多维度信息的同时输入、情感因素的干扰以及社会规范的约束在现有模型中尚未得到充分研究(FeldmanHall & Nassar, 2021)，模型参数的个体间变异及其来源（如认知能力、文化背景）也需进一步考察(Eckstein et al., 2021)。第三，联想表征与命题表征的区分虽然在概念层面清晰，但两者在实际认知过程中的交互机制，特别是从联想到命题的转化条件以及命题对联想的抑制边界仍需更精细的实验范式来揭示。第四，泛化过程中从群体到个体的演绎推断与从个体到

群体的归纳推断的计算机制及相关神经基础也缺乏研究。

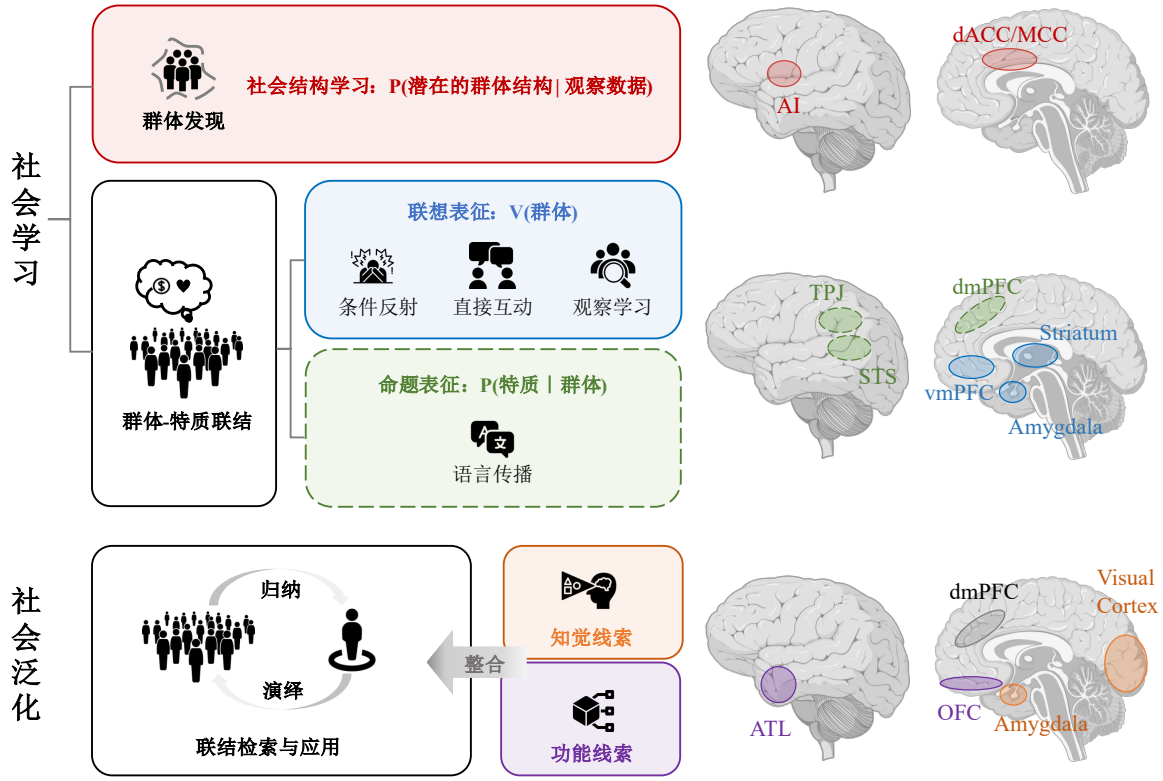

图 1 刻板印象从社会学习到泛化应用的计算框架与各环节关键脑区。
“群体发现”对应贝叶斯结构学习过程(红色框)：大脑从社会观察数据中推断潜在群体结构，即计算 P(潜在的群体结构| 观察数据)，涉及前脑岛(anterior insula, AI)编码抽象的潜在社会结构和背侧前扣带皮层/中扣带皮层(dorsal anterior cingulate cortex / midcingulate cortex, dACC/MCC)。
“群体-特质联结”展示联结建立的双通路机制及其双重表征。**联想表征**由经验通路(条件反射、直接互动、观察学习)通过强化学习建立(蓝色框)，其中 V(群体)是对这一表征的形式化描述。核心脑区包括：杏仁核(amygdala)支持评价性条件反射中的情感联结学习；纹状体(striatum)和腹内侧前额叶皮层(ventromedial prefrontal cortex, vmPFC)编码社会性与物质性价值信号。**命题表征**主要由语言通路承载(绿色虚线框)，其概率结构可通过 P(特征| 群体)描述。候选神经基础可能涉及颞顶联合区(temporo-parietal junction, TPJ)和颞上沟(superior temporal sulcus, STS)等与信念归因和社会推理相关的脑区；此外，背内侧前额叶皮层(dorsomedial prefrontal cortex, dmPFC)可能参与命题层面的信念整合。需要指出，命题表征的脑区定位基于社会认知中心智化与贝叶斯推断相关文献的理论推测，尚需直接实证检验。
“联结检索与应用”展示了社会泛化中的群体识别机制。**知觉泛化通路**(橙色框)依赖视觉皮层(visual cortex)早期加工阶段的调谐曲线锐化，以及杏仁核对威胁面孔的自动化评估；**功能泛化通路**(紫色框)依赖前颞叶(anterior temporal lobe, ATL)和眶额皮层(orbitofrontal cortex, OFC)等高级概念系统。双向循环箭头表示社会泛化的双向性：既包含从群体到个体的演绎推断，也包含从个体到群体的归纳推断，两者共同构成一个不断更新的社会认知循环。

5 未来展望

基于当前的局限性与计算认知科学的最新进展，未来研究可从以下三个方面推进对刻板印象机制的理解，从而构建更具整合性的理论框架。

第一，关系线索在刻板印象形成与泛化中的作用值得深入考察。现有泛化研究主要聚焦于特征相似性，然而社会生活中的关系线索——个体在社会网络中的位置与连接——可能是另一种关键的泛化路径 (汪涵 等, 2025)。已有研究表明，大脑能够编码社会网络中的传递性关系，并据此对陌生人进行信任评价或特质推断(Son et al., 2021, 2023)。神经层面的证据显示，关系价值相似的个体在大脑中会被更相似地编码(Babür et al., 2024)。未来研究可设计包含社会网络信息的实验范式，

考察关系线索与特征线索如何共同作用、相互调节，以影响刻板印象的泛化。例如，可以通过操纵参与者在虚拟社交网络中的位置与连接模式，检验基于网络结构的泛化是否独立于基于特征相似性的泛化。

第二，社会认知地图(cognitive map)可能为理解多线索信息的表征与整合提供理论框架。已有研究在慷慨-能力(Gao et al., 2026)和道德-能力(Liu et al., 2025) 等二维特质空间中发现了社会认知地图的神经证据，表明海马-内嗅皮层可以对社会信息进行类似空间导航的编码(Liang et al., 2024; Park et al., 2020; Schafer & Schiller, 2018)。然而，这些研究主要聚焦于个体的特质评价，尚未涉及群体层面的表征。在此视角下，刻板印象的社会学习可能是在地图上标记群体位置的过程。贝叶斯结构学习确定群体的分布和边界，强化学习为每个位置赋予价值信号，而泛化则是根据新个体在地图上的坐标推断其属性(Tavares et al., 2015)。未来可结合群体层面的学习任务与高分辨率神经影像技术，考察群体-特质联结是否以空间化的方式编码于海马-内嗅皮层，并检验不同泛化线索在地图空间中的整合机制。

第三，大语言模型为阐明语言传播如何塑造并固化刻板印象提供了新的研究工具。语言是命题表征形成的主要通路，而 LLM 可以作为这一通路的计算模拟器。通过构建基于 LLM 的多智能体交互网络，研究者可以操纵网络结构、初始偏见强度和传播规则等参数，观察微小偏见如何在特定条件下被放大并固化为群体共识(Gelpí et al., 2025; Stewart & Raihani, 2023)。将 LLM 模拟与神经科学技术相结合，例如通过比较 LLM 内部表征与人脑神经数据中的刻板印象编码模式，有助于在计算和神经两个层面同时验证语言传播对刻板印象形成的贡献。然而也需警惕 LLM 本身的局限，它反映的是语料统计特征而非人类认知过程，不能将 LLM 输出等同于人类认知。

综上，本文尝试从计算认知角度为刻板印象研究中看似分散的现象提供一个连贯的解释框架。随着计算建模、神经影像和大语言模型等方法的交叉融合，未来研究的关键在于将当前主要停留在算法描述层面的框架与神经实现对接，并在更具生态效度的社会情境中检验其预测。

参考文献

汪涵, 董昱林, 刘凝丰, & 朱露莎. (2025). 社会决策信息的抽象与泛化. *心理科学*, *48*(4), 962–971. https://doi.org/10.16719/j.cnki.1671-6981.20250416

Adorno, T. W. (1950). *The authoritarian personality*. John Wiley & Sons, Inc.

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 Decision-Making*, *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/10.1037/0033-2909.132.5.692

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 Psychology*, *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/cogs.12560

Joiner, J., Piva, M., Turrin, C., & Chang, S. W. C. (2017). Social learning through prediction error in the brain. *Npj Science of Learning*, *2*(1), 8. https://doi.org/10.1038/s41539-017-0009-2

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. https://doi.org/10.1037/h0034747

Kobayashi, K., Kable, J. W., Hsu, M., & Jenkins, A. C. (2022). Neural representations of others' traits predict social decisions. *Proceedings of the National Academy of Sciences*, *119*(22), e2116944119. https://doi.org/10.1073/pnas.2116944119

Koenig, A. M., & 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*(3), 371–392. https://doi.org/10.1037/a0037215

Krahé, B., Uhlmann, A., & Herzberg, M. (2021). The voice gives it away: Male and female pitch as a cue for gender stereotyping. *Social Psychology*, *52*(2), 101–113.

Kumaran, D., Banino, A., Blundell, C., Hassabis, D., & Dayan, P. (2016). Computations underlying social hierarchy learning: Distinct neural mechanisms for updating and representing self-relevant information. *Neuron*, *92*(5), 1135–1147. https://doi.org/10.1016/j.neuron.2016.10.052

Kurdi, B., Morehouse, K. N., & Dunham, Y. (2023). How do explicit and implicit evaluations shift? A preregistered meta-analysis of the effects of co-occurrence and relational information. *Journal of Personality and Social Psychology*, *124*(6), 1174–1202. https://doi.org/10.1037/pspa0000329

Lau, T., Gershman, S. J., & Cikara, M. (2020). Social structure learning in human anterior insula. *eLife*, *9*, e53162. https://doi.org/10.7554/eLife.53162

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

Liang, Z., Wu, S., Wu, J., Wang, W.-X., Qin, S., & Liu, C. (2024). Distance and grid-like codes support the navigation of abstract social space in the human brain. *eLife*, *12*, RP89025. https://doi.org/10.7554/eLife.89025

Liu, J., Zhou, Y., Wang, H., Yang, L.-Z., & Li, H. (2025). Rank labels scaffold social cognitive maps in the hippocampal-entorhinal system. *NeuroImage*, *317*, 121366. https://doi.org/10.1016/j.neuroimage.2025.121366

Lockwood, P. L., & Klein-Flügge, M. C. (2021). Computational modelling of social cognition and behaviour—A reinforcement learning primer. *Social Cognitive and Affective Neuroscience*, *16*(8), 761–771. https://doi.org/10.1093/scan/nsaa040

Maass, A. (1999). Linguistic intergroup bias: Stereotype perpetuation through language. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 31, pp. 79–121). Academic Press. https://doi.org/10.1016/S0065-2601(08)60272-5

Mahmoodi, A., & Rushworth, M. F. S. (2026). Computational origins of cortical brain circuits for social cognition. *Nature Reviews Neuroscience*, *27*, 345–356. https://doi.org/10.1038/s41583-026-01028-2

Martin, D., Hutchison, J., Slessor, G., Urquhart, J., Cunningham, S. J., & Smith, K. (2014). The spontaneous formation of stereotypes via cumulative cultural evolution. *Psychological Science*, *25*(9), 1777–1786. https://doi.org/10.1177/0956797614541129

Martinez, J. E., Krasner, R. H., Rosero, L., Gershman, S. J., & Cikara, M. (2025). Social group discovery, structure, and stereotype updating. *Journal of Experimental Psychology. General*, *154*(11), 3094–3113. https://doi.org/10.1037/xge0001830

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

McCauley, C., & Stitt, C. L. (1978). An individual and quantitative measure of stereotypes. *Journal of Personality and*

Social Psychology, *36*(9), 929–940. https://doi.org/10.1037/0022-3514.36.9.929

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience*, *8*(6), 623–631. https://doi.org/10.1093/scan/nss040

Mende-Siedlecki, P., Said, C. P., & Todorov, A. (2013). The social evaluation of faces: A meta-analysis of functional neuroimaging studies. *Social Cognitive and Affective Neuroscience*, *8*(3), 285–299. https://doi.org/10.1093/scan/nsr090

Noyes, A., Dunham, Y., Keil, F. C., & Ritchie, K. (2021). Evidence for multiple sources of inductive potential: Occupations and their relations to social institutions. *Cognitive Psychology*, *130*, 101422. https://doi.org/10.1016/j.cogpsych.2021.101422

Öhman, A. (2005). Conditioned fear of a face: A prelude to ethnic enmity? *Science*, *309*(5735), 711–713. https://doi.org/10.1126/science.1116710

Olsson, A., Ebert, J. P., Banaji, M. R., & Phelps, E. A. (2005). The role of social groups in the persistence of learned fear. *Science*, *309*(5735), 785–787. https://doi.org/10.1126/science.1113551

Olsson, A., Knapska, E., & Lindström, B. (2020). The neural and computational systems of social learning. *Nature Reviews Neuroscience*, *21*(4), 197–212. https://doi.org/10.1038/s41583-020-0276-4

Park, S. A., Miller, D. S., Nili, H., Ranganath, C., & Boorman, E. D. (2020). Map making: Constructing, combining, and inferring on abstract cognitive maps. *Neuron*, *107*(6), 1226-1238.e8. https://doi.org/10.1016/j.neuron.2020.06.030

Pietraszewski, D. (2022). Toward a computational theory of social groups: A finite set of cognitive primitives for representing any and all social groups in the context of conflict. *Behavioral and Brain Sciences*, *45*, e97. https://doi.org/10.1017/S0140525X21000583

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

Rösler, I. K., Kerber, I., & Amodio, D. M. (2025). Effects of moral stereotypes on the formation and persistence of group preferences. *Journal of Experimental Social Psychology*, *119*, 104750. https://doi.org/10.1016/j.jesp.2025.104750

Schafer, M., & Schiller, D. (2018). Navigating social space. *Neuron*, *100*(2), 476–489. https://doi.org/10.1016/j.neuron.2018.10.006

Schultner, D. T., Lindström, B. R., Cikara, M., & Amodio, D. M. (2024). Transmission of social bias through observational learning. *Science Advances*, *10*(26), eadk2030. https://doi.org/10.1126/sciadv.adk2030

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

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

Sherif, M. (1936). *The psychology of social norms*. Harper.

Sherif, M., & Sherif, C. W. (1953). *Groups in harmony and tension; an integration of studies of intergroup relations*. Harper & Brothers.

Sherman, J. W., Kruschke, J. K., Sherman, S. J., Percy, E. J., Petrocelli, J. V., & Conrey, F. R. (2009). Attentional processes in stereotype formation: A common model for category accentuation and illusory correlation. *Journal of Personality and Social Psychology*, *96*(2), 305–323. https://doi.org/10.1037/a0013778

Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: Hippocampal–midbrain encoding of overlapping events. *Neuron*, *60*(2), 378–389. https://doi.org/10.1016/j.neuron.2008.09.023

Solanki, P., & Cesario, J. (2025). Stereotypes as bayesian prediction of social groups. *The Journal of Social Psychology*, *165*(5), 640–662. https://doi.org/10.1080/00224545.2024.2368017

Son, J.-Y., Bhandari, A., & FeldmanHall, O. (2021). Cognitive maps of social features enable flexible inference in social

networks. *Proceedings of the National Academy of Sciences*, *118*(39), e2021699118. https://doi.org/10.1073/pnas.2021699118

Son, J.-Y., Bhandari, A., & FeldmanHall, O. (2023). Abstract cognitive maps of social network structure aid adaptive inference. *Proceedings of the National Academy of Sciences*, *120*(47), e2310801120. https://doi.org/10.1073/pnas.2310801120

Stegmann, Y., Ahrens, L., Pauli, P., Keil, A., & Wieser, M. J. (2020). Social aversive generalization learning sharpens the tuning of visuocortical neurons to facial identity cues. *eLife*, *9*, e55204. https://doi.org/10.7554/eLife.55204

Stewart, A. J., & Raihani, N. (2023). Group reciprocity and the evolution of stereotyping. *Proceedings of the Royal Society B*, *290*(1991), 20221834. https://doi.org/10.1098/rspb.2022.1834

Stolier, R. M., & Freeman, J. B. (2016). Neural pattern similarity reveals the inherent intersection of social categories. *Nature Neuroscience*, *19*(6), 795–797. https://doi.org/10.1038/nn.4296

Sutherland, C. A. M., & Young, A. W. (2022). Understanding trait impressions from faces. *British Journal of Psychology*, *113*(4), 1056–1078. https://doi.org/10.1111/bjop.12583

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). MIT press.

Tajfel, H., & Turner, J. C. (2004). *The social identity theory of intergroup behavior*. Psychology Press. https://doi.org/10.4324/9780203505984-16

Tavares, R. M., Mendelsohn, A., Grossman, Y., Williams, C. H., Shapiro, M., Trope, Y., & Schiller, D. (2015). A map for social navigation in the human brain. *Neuron*, *87*(1), 231–243. https://doi.org/10.1016/j.neuron.2015.06.011

Taylor, S. E. (1981). A categorization approach to stereotyping. In D. L. Hamilton (Ed.), *Cognitive Processes in Stereotyping and Intergroup Behavior* (1st ed., pp. 83–114). Psychology Press.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Traast, I. J., Doosje, B., & Amodio, D. M. (2025). Impression formation through social interaction: The effect of ethnicity in the Dutch context. *Group Processes & Intergroup Relations*, *28*(4), 831–855. https://doi.org/10.1177/13684302241305054

Traast, I. J., Schultner, D. T., Doosje, B., & Amodio, D. M. (2024). Race effects on impression formation in social interaction: An instrumental learning account. *Journal of Experimental Psychology. General*, *153*(12), 2985–3001. https://doi.org/10.1037/xge0001523

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

van Baar, J. M., Nassar, M. R., Deng, W., & FeldmanHall, O. (2022). Latent motives guide structure learning during adaptive social choice. *Nature Human Behaviour*, *6*(3), Article 3. https://doi.org/10.1038/s41562-021-01207-4

Verosky, S. C., & Todorov, A. (2010). Generalization of affective learning about faces to perceptually similar faces. *Psychological Science*, *21*(6), 779–785. https://doi.org/10.1177/0956797610371965

Villiger, D. (2025). Stereotypes and self-fulfilling prophecies in the Bayesian brain. *Inquiry*, *68*(10), 3317–3341. https://doi.org/10.1080/0020174X.2023.2166983

Wake, S. J., & Izuma, K. (2017). A common neural code for social and monetary rewards in the human striatum. *Social Cognitive and Affective Neuroscience*, *12*(10), 1558–1564. https://doi.org/10.1093/scan/nsx092

Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J., & Lamm, C. (2020). Using reinforcement learning models in social neuroscience: Frameworks, pitfalls and suggestions of best practices. *Social Cognitive and Affective Neuroscience*, *15*(6), 695–707. https://doi.org/10.1093/scan/nsaa089

Computational cognitive mechanism of stereotype: Social learning and generalization

Abstract: Stereotypes—generalized beliefs that members of a social group tend to possess certain traits—profoundly shape interpersonal interactions and intergroup relations, making it important to understand the cognitive mechanisms underlying their formation and maintenance. Traditional theories, however, have largely remained at the descriptive level, lacking an integrative, mechanistic account of how stereotypes are acquired and applied. This review addresses this gap by organizing the literature around two core processes—social learning and social generalization—and integrating reinforcement learning (RL) and Bayesian theory to elucidate the computational cognitive mechanisms of stereotyping.

At the level of social learning, we first describe how Bayesian structure learning enables the brain to infer latent group categories from observable social data. We then distinguish two pathways through which group–trait associations are established: the experiential pathway, which primarily forms automatic associative representations through RL mechanisms, and the linguistic pathway, which primarily conveys probabilistic propositional representations through Bayesian inference. Regarding the updating and consolidation of these associations, prediction error serves as the core signal driving revision, yet prior biases and asymmetric learning rates cause updating to favor existing beliefs. Furthermore, the explore–exploit dilemma reveals that biased information sampling is a key source of stereotype persistence. At the level of social generalization, we analyze how the brain categorizes novel individuals into known groups on the basis of perceptual similarity and functional similarity, and discuss the retrieval mechanisms through which learned associative knowledge is applied. Finally, we propose three directions for future research: relational cues as a generalization pathway beyond feature similarity, social cognitive maps as an integrative framework for multi-cue representation, and large language model as a tool to simulate stereotype consolidation through linguistic transmission.

Keywords: Stereotype, Social Generalization, Reinforcement Learning, Bayesian Theory, Computational modeling