Cognitive Systems Research 93 (2025) 101396

Contents lists available at ScienceDirect

Cognitive Systems Research

journal homepage: www.elsevier.com/locate/cogsys

Positivity bias in generalizing moral impressions from individuals to groups

Hui Huang [a,b,1], Jiecheng Huangliang [a,b,1], Qinglu Xiao [a,b], Ting Zhao [a,b], Qingqing Ye [a,b], Haokui Xu [c,*], Jun Yin [a,b,**]

[a] Department of Psychology, Ningbo University, Ningbo, China
[b] Center of Group Behavior and Social Psychological Service, Ningbo University, Ningbo, China
[c] Department of Psychology, College of Education, Zhejiang University of Technology, Hangzhou, China

ARTICLE INFO

Keywords:
Positivity bias
Moral impression
Morality generalization
Prior belief
Valence

ABSTRACT

Individuals should seek to become aware of the valence (positive/negative) of their moral impressions of others, as it determines whether the observed individuals may be helpful or harmful. Studies have documented the capacity to form moral impressions of groups by integrating individual moral characters, but whether and how moral valence influences the generalization of moral impressions from individuals to groups remain unknown, especially in dynamic learning contexts. In the present study, participants sequentially predicted and observed positive (helping) or negative (hindering) behaviors of two group members to form moral impressions about these individuals. They reported their moral impressions of an unknown group member with no trait-implying behavior and of the entire group. Experiments 1 and 2 demonstrated greater generalization of positive moral impressions than of negative moral impressions from individuals to groups, indicating a group-level positivity bias. Experiment 3 manipulated participants' prior beliefs through preexposure to group moral impressions. The results revealed that the group-level positivity bias persisted after the initial formation of positive group impressions, while the initial formation of negative impressions led to a corresponding negativity bias. The Bayesian findings suggested that the observed positivity bias in morality generalization within groups can be attributed to a prior belief in the positivity of groups. Thus, social groups not only influence the selection of shared moral characters among their members but also contribute to prior knowledge shaping group moral impressions. This prior belief functions as a default assumption in social evaluations.

1. Introduction

Effectively navigating social interaction involves learning about other individuals and using that information appropriately, such as seeking cooperation with potential allies (Henrich & Muthukrishna, 2021; Henrich & Henrich, 2007). The social context has shaped human cognitive capacities throughout evolution, informing assessments of whether an individual is good or bad (i.e., whether he or she is likely to help or harm us; Leach et al., 2007; Randell, 2021). Consequently, people possess a strong ability to form moral impressions of others (i.e., moral character),[2] which are believed to dominate other kinds of social impressions (Brambilla et al., 2021; Goodwin et al., 2014; Wojciszke et al., 1998). Understanding the valence of these impressions, whether positive or negative, is crucial for reaping the benefits of social interactions and avoiding exploitation. Therefore, the influence of valence on the formation of moral impressions has become a major focus of research.

Previous studies have consistently shown that morality is a distinct dimension and often is more influential than competence or warmth in shaping how we form impressions of others. Wojciszke et al. (1998) found that morality is a better predictor of our impressions than competence is. Notably, Goodwin et al. (2014) revealed that moral character information forms the core of our identity, largely independent of external influences. Furthermore, Ybarra et al. (2001) have

* Corresponding author at: Department of Psychology, College of Education, Zhejiang University of Technology, Hangzhou, China.
** Corresponding author at: Department of Psychology, Ningbo University, Ningbo, China.
E-mail addresses: haokuixu.psy@gmail.com (H. Xu), yinjun1@nbu.edu.cn (J. Yin).

[1] Hui Huang and Jiecheng Huangliang contribute equally to this study.

[2] The term moral trait is occasionally utilized to indicate moral impressions with more tangible attributes, such as kindness or amiability. These concrete attributes enable the formation of an overall moral impression of others, whether positive or negative. To maintain consistency with previous studies, we predominantly employ the term "moral impressions" when discussing this topic, particularly in relation to the measurement outlined in the Methods section.

https://doi.org/10.1016/j.cogsys.2025.101396
Received 26 June 2024; Received in revised form 8 May 2025; Accepted 25 August 2025
Available online 30 August 2025

shown that moral trait words are processed more rapidly than are words associated with competence. Additionally, impressions of trustworthiness (a kind of moral character) can form after a 100-ms exposure to a face (Willis & Todorov, 2006), and during the observation of dyadic social interactions, moral valence is rapidly encoded, influencing both N1 and N2 event-related potential components (Yoder & Decety, 2014). These processing capabilities enable us to effectively form moral impressions about others.

Research on impression formation for individuals widely demonstrates that moral impressions at the individual level do not equally weigh all attributes. Instead, they tend to assign greater importance to negative moral information than to positive moral information, a phenomenon known as negativity bias in individual impression formation (Martijn et al., 1992; Skowronski & Carlston, 1987). This tendency is stronger when immoral information follows moral information than vice versa (Reeder & Coovert, 1986). The cue-diagnosticity approach is typically used to explain this bias, positing that people perceive negative moral information as more indicative of a person's character than positive moral information (Alves et al., 2017b; Shin & Niv, 2021; Unkelbach et al., 2020). In social contexts, positive attributes are often considered the default (Alves et al., 2017a), while negative attributes are thought to be less common (Mende-Siedlecki et al., 2013; Shin & Niv, 2021). These beliefs are supported by research on conceptual similarities indicating that the distribution of negative concepts is sparser than that of positive concepts (Alves et al., 2015; Unkelbach et al., 2008). Consequently, negative information serves as a more diagnostic cue and is given greater weight in evaluations. Research also suggests that the diagnosticity of cues depends on their contextual relevance to a specific judgment task rather than on the mere presence of negative cues (Reeder & Brewer, 1979; Skowronski & Carlston, 1989; Wojciszke et al., 1993). For example, positivity bias was found in the domain of ability evaluation (Martijn et al., 1992; Skowronski & Carlston, 1987). In addition to demonstrating the negativity bias that helps prevent exploitation, studies have revealed that when people form impressions about the moral character of others, their impressions of individuals engaging in negative behaviors are more uncertain than those of individuals involved in positive behaviors (Siegel et al., 2018). This uncertainty allows for the rapid updating of initially negative impressions if there is a behavioral improvement, thereby contributing to the maintenance and development of relationships (Crockett et al., 2021).

The aforementioned studies focus predominantly on moral impressions of specific individuals. However, a significant portion of social activities occur within social groups and involve the behaviors of groups of individuals. Numerous studies have proposed that human cognitive mechanisms are finely tuned to group attributes and have even developed to make spontaneous trait inferences about groups, thereby forming moral impressions of groups (Hamilton et al., 2015; Hamilton & Thurston, 2022; Kashima et al., 2000; Stangor & Lange, 1994). Specifically, individuals tend to generalize individual moral impressions to the overall group and extend these impressions from known members of the group to unknown group members who have not exhibited trait-indicative behavior (Crawford et al., 2002; Hackel et al., 2022; Kunda & Thagard, 1996). This mechanism enables the rapid formation of moral impressions of entire groups to facilitate social interactions. Although studies have documented the capacity to form moral impressions of groups by integrating different moral traits (Lammers et al., 2018; Rozin & Royzman, 2001; Shin & Niv, 2021), whether and how moral valence influences the generalization of moral impressions from individuals to groups remain unknown.

In the broad context of generalization within groups, individuals commonly perceive members within a social group as sharing similar attributes. They seek evidence of these shared attributes from observed members and then extend these to the entire group (Liberman et al., 2017; Powell & Spelke, 2013; Ranganath & Nosek, 2008; Yin et al., 2022). When moral impressions are shared among certain group members, they serve as evidence when forming an overall group impression (Hackel et al., 2022). It would seem that the generalization of morality should rely on the cumulative evidence of shared moral characters among group members rather than on the moral valence itself. However, social groups not only establish boundaries regarding what shared attributes qualify as evidence but also shape prior beliefs regarding groups (Hu et al., 2023; Shutts & Kalish, 2021; Sidanius & Pratto, 2001). Studies indicate that people's mental representations of the social world tend toward positivity, where positive attributes are more common than negative attributes (Alves et al., 2017b; Feinberg et al., 2014; Unkelbach et al., 2020). This finding suggests that when generalizing moral impressions to groups, individuals may believe that positivity is predominantly shared among group members, which is known as the prior belief in the positivity of groups. This belief aligns with the view that prosociality is a fundamental aspect of human nature (Schroeder & Graziano, 2018). Moreover, this belief complements the diagnostic value typically assigned to negative behavior in individual trait inference, where negative impressions tend to be attributed specifically to individual actors rather than generalized to the group as a whole (Alves et al., 2017b; Unkelbach et al., 2020).

Accordingly, considering the aforementioned prior belief, we propose that positive moral impressions are generalized from individuals to groups to a greater extent than negative impressions are, revealing a positivity bias in generalizing moral impressions (i.e., group-level positivity bias). Even when the positive and negative moral impressions of individuals in two groups are assessed at a similar level, the generalization of morality to groups is stronger for individual positive impressions of group members than for individual negative impressions. In other words, positive moral impressions of known group members are generalized more readily to an unknown group member who has not exhibited trait-implying behavior than are negative moral impressions. Our proposed positivity bias results in placing increased emphasis on attributing attributes to social groups rather than to particular individuals across various situations (Baumeister et al., 2001; Pratto & John, 1991). This positivity bias at the group level is not inconsistent with the negativity bias found in individual-level impression formation. Rather, it reflects a distinct cognitive mechanism operating in different contexts of social inference. Specifically, negativity bias tends to emerge robustly in impression formation and updating at the individual level, particularly when detecting moral deviance (e.g., Fiske, 1980; Reeder & Coovert, 1986). In contrast, the positivity bias that we identified occurs during the generalization of moral impressions at the group level, especially when individuals' behaviors are used to infer the moral character of the broader group. This positivity bias is supported by a recent study that showed that helpful actions are more likely than harmful actions to be implicitly generalized to group members (Huangliang et al., 2024). However, the study did not explore why this bias occurs, specifically whether it is due to a prior belief in group positivity.

Instead of relying solely on statistical evidence regarding group members' moral characteristics, our hypothesis places greater emphasis on the role of prior beliefs in forming moral impressions of groups. While the weighted averaging model (Kashima et al., 2000) has been proposed to highlight how evidence about traits is proportionally integrated (Rozin & Royzman, 2001), it tends to overlook the significance of prior beliefs about evaluated targets. A recent study by Shin and Niv (2021) revealed a bias in the overall estimation of a group's generosity toward the value of rare events, irrespective of valence, stemming from the inference of hidden causes behind group members' behaviors. However, their study focused primarily on how statistical evidence supports the inference. Our hypothesis does not aim to refute these models but rather underscores the importance of considering prior beliefs about groups when generalizing moral impressions to groups. Moreover, the majority of studies investigating group impression formation have considered a limited number of behaviors, typically presented at no more than two time points (i.e., initially after being presented with one set of behavior descriptions and subsequently after exposure to another set) (Crockett

et al., 2021; but see Hackel et al., 2022). However, evaluations of individuals are often based on probabilistic inferences drawn from observed behaviors unfolding over time. Therefore, it is essential to address the presence of positivity bias in generalizing moral impressions and the impact of prior beliefs in dynamic learning contexts.

In this study, we investigated the impact of moral valence on the generalization of moral impressions from individuals to groups in dynamic learning contexts by employing a reinforcement learning task. We aimed to understand whether and why individuals tend to exhibit a positivity bias when making such generalizations. To address this question, it is crucial to design scenarios in which participants can form comparable intensities of positive and negative moral impressions of individuals engaging in valenced behaviors. By doing so, we sought to observe whether and how the evaluations of an unknown group member or the corresponding group manifest positivity bias. Adapting previous scenarios that utilized animated events to depict individuals with different valenced moral impressions, actors were portrayed as engaging in either positive behavior (helping) or negative behavior (hindering) in relation to another agent climbing a hill (Hamlin et al., 2007; Isik et al., 2017), a method demonstrated to elicit comparable strong impressions across conditions (Huangliang et al., 2024). To integrate these scenarios with groups, a group of three agents was formed by presenting groupness cues of a common fate (e.g., moving synchronously), similar to previous studies (Duan et al., 2021; Hamlin et al., 2007; Ludwig et al., 2020; Powell & Spelke, 2013). Participants consecutively predicted and observed the positive or negative behaviors of two group members (referred to as known members) to form moral impressions about these individuals. Moreover, participants reported their moral impressions of an unknown group member who did not exhibit trait-implying behavior, as well as of the entire group. The hypothesis positing a positivity bias in generalizing moral impressions from individuals to groups suggests that the moral impressions of group members engaging in helping behaviors exhibit stronger morality generalization than those of individuals engaging in hindering behaviors, even when individual-level valenced impressions are matched. When the prior belief is intervened in, the positivity bias changes accordingly. Furthermore, we anticipated that when decomposing such an effect into different variables and parameters using the Bayesian learning model, individuals hold a default prior belief in the positivity of groups.

This study comprised three experiments. Initially, in Experiment 1, we examined whether participants formed comparable strong moral impressions of individuals after observing group members engage in helping or hindering behaviors, and we tried to provide initial evidence that, although the intensities of individual impressions were comparable, positive moral impressions were more readily generalized to other group members than were negative moral impressions. Subsequently, in Experiment 2, which consisted of two sub-experiments, we sought to replicate the findings of Experiment 1 and performed assessments of moral impressions at the group level to further examine the occurrence of positivity bias at the group level. Finally, in Experiment 3, we explored potential shifts in positivity bias associated with morality generalization when prior default beliefs of positivity in groups were altered by preexposure to differently valenced moral impressions about groups. In all the experiments, the Bayesian learning model was employed to calculate the magnitude of the prior belief in the positivity of groups.

2. Experiment 1

Experiment 1 aimed to examine whether participants formed comparable strong moral impressions of individuals who engaged in helping versus hindering behaviors within a group-based reinforcement learning task. After observing two group members perform similar valenced actions, participants rated a third, unobserved member to assess moral generalization. This design provided preliminary evidence for a positivity bias in generalizing individual impressions to the group level.

2.1. Methods

2.1.1. Participants

A total of 36 participants with normal or corrected-to-normal vision (15 men and 21 women; mean age = 20.92 years, SD = 2.17) participated in this study. The participants were from the authors' university. All participants provided their demographic information, provided written informed consent and received financial rewards. To determine the minimum sample size needed to detect a significant effect, we conducted a power analysis using G*Power software (version 3.1; Faul et al., 2009). In our study, we introduced a bias index to assess biases in forming moral impressions. Since the bias index was derived from raw evaluations of moral impressions, determining the required effect size involved detecting the generalization bias from these raw evaluations. Specifically, with the raw evaluations with the dependent variable, we aimed to detect a 2 (moral valence: positive vs. negative) × 2 (rated target: known vs. unknown member) interaction effect. Assuming a conventional medium effect size of $f = 0.20$, with an alpha level of 0.05 and a power of 0.80, the minimum required sample size was determined to be 36 participants. Even when using the bias index as the dependent variable, comparing two measures (known member vs. unknown member or group) requires at least 34 participants to detect a conventional medium effect size ($f = 0.20$) with an alpha level of 0.05 and a power of 0.80. The final recruited sample size of 36 met these requirements. Since Experiments 2a, 2b, and 3 all examined potential biases in moral impression formation using similar designs, we maintained the same expected effect size as in Experiment 1 and recruited 36 valid participants for each between-subjects condition.

All the experiments were approved by the Research Ethics Board of the Department of Psychology at the authors' university and were performed in accordance with the relevant guidelines and regulations. We reported all the measures, manipulations, and exclusions for all the experiments. No study designs were preregistered. All the data, analysis codes and videos containing example trials for each condition have been made publicly available at the Open Science Framework (OSF) and can be accessed at https://osf.io/xuc83/?view_only=7aebad94347346fe92a30daac7e6b29b. ChatGPT-3.5 was used for reviewing and editing the text.

2.1.2. Apparatus and stimuli

All stimuli were presented on a 15.6-inch LED monitor (1,024 × 768-pixel resolution; 120 Hz refresh rate) on a laptop computer at a 60-cm viewing distance. We utilized custom software developed in PsychoPy (version 3; Peirce et al., 2019) to present the stimuli. The visual stimuli consisted of animations representing helping and hindering behaviors created using Blender (version 3.0.1). Each event had a background containing two hills, one at the left and the other at the right edge of the screen, and there was a platform between the hills. The two hills were identical in both size and slope. There was no actual difference in steepness or difficulty between the hill that the climber ascended in the helping scenario and that in the hindering scenario. At the center of the platform, there were three agents with the same color but different shapes. The agents are referred to by their shapes (in pinyin) during the experiment (i.e., "jiajia", cross; "fangfang", square; "yuanyuan", circle). These agents exhibited behaviors that involved either assisting another agent (referred to as the climber) in climbing a hill or impeding the climber from doing so (see Fig. 1a). Two climbers moved to the bottom of each hill and attempted to reach the top. However, they slid partway down the steep slope each time and were unable to successfully reach the summit. One of the climbers was either bumped up the hill by an actor (helping behavior) or bumped down the hill by an actor (hindering behavior), and then both agents exited the scene. The remaining climber, who did not engage in any interactions, slid down to the bottom of the hill and then exited the scene. Each agent was depicted with two eyes as facial features, occupying a space measuring approximately 1.5° × 1.5°.

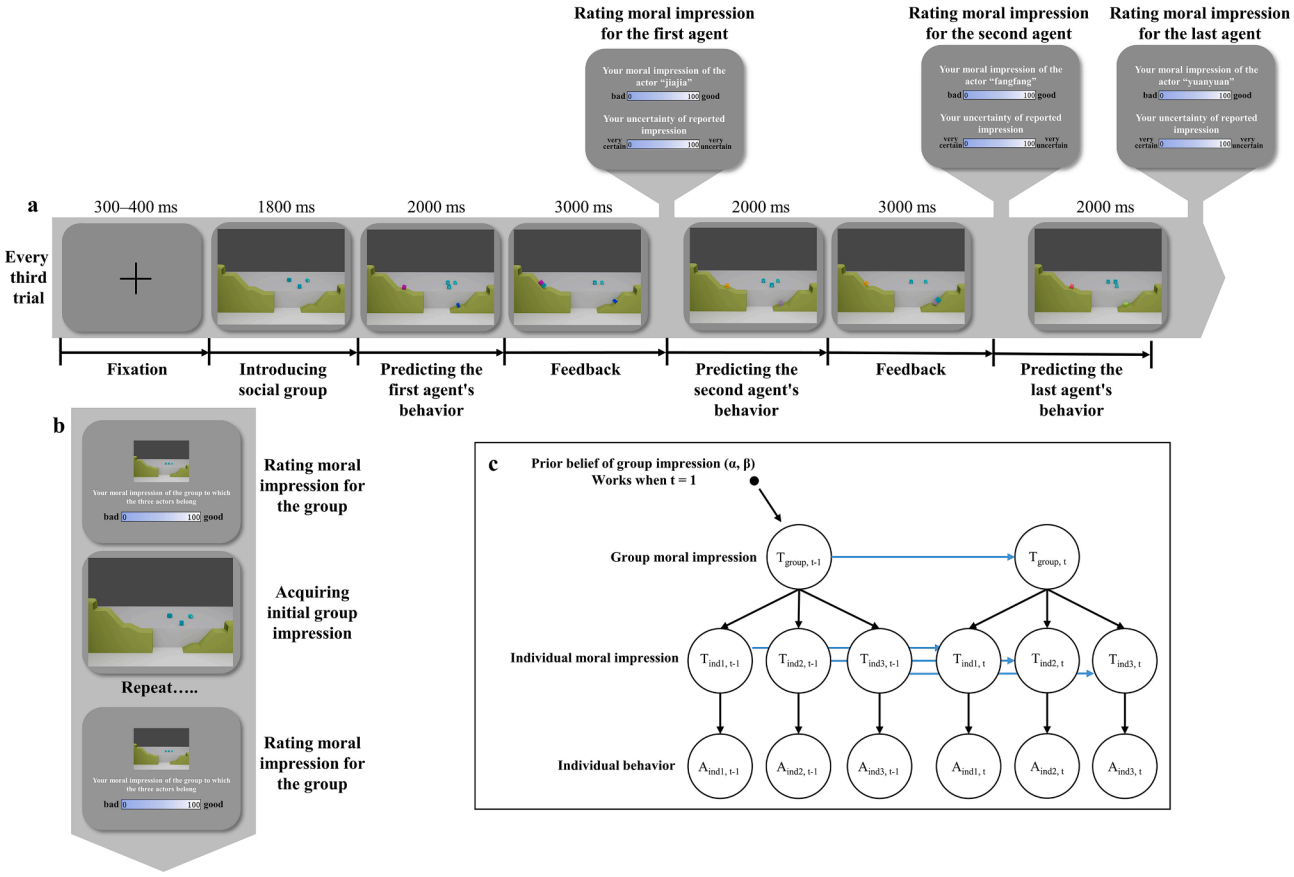

Fig. 1. Overview of the experimental procedure and generative model. (a) In each trial, participants predicted the behaviors of three agents from one social group, and the future behaviors of the first two agents were revealed as feedback. Agents exhibited either a helping behavior or a hindering behavior. After every third trial, participants indicated their subjective impressions of each agent's morality and the uncertainty of their reported impression (for the text shown in the pictures, please refer to the main text). (b) The formation of the initial group impression in Experiment 3. (c) We utilized a modified hierarchical generative model to replicate the process of generating individual moral impressions and behaviors and applied Bayesian methods to infer the potential priors related to groups. The prior distribution of group moral impressions was represented as a beta distribution with parameters α and β. The black arrow indicates the generative direction, and the blue arrow indicates the temporal dynamic direction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Regarding the helped/hindered targets (i.e., climbers), we ensured that they were visually distinct from the three agents in both shape and color (see Supplementary Fig. 7). There were six unique climbers, and each was characterized by a distinct color and shape. These climbers were organized into three pairs, with each pair containing one agent positioned above and one below the platform, resulting in six possible combinations that remained constant for each participant. In each trial, a pair of climbers was randomly selected, ensuring that across a block, each pair appeared an equal number of times and each individual climber had an equal probability of being helped or hindered. Thus, while there was variation in climber color and shape, this variation was systematically controlled for and randomized to avoid potential confounds.

2.1.3. Procedure and design

At the beginning of each trial, participants were presented with a fixation cross (1.5° × 1.5°) displayed in the center of the screen. Subsequently, a video was presented of the three agents performing synchronized movements along a circular path (see Fig. 1 and Video S1). We intentionally used the synchronous movement—a feature known to promote perceived groupness (Campbell, 1958)—to strengthen groupness. This cue was consistent with prior studies on group perception and has been widely used to operationalize groupness in similar contexts (Yin et al., 2022; 2023). After this introductory sequence, the agents were distributed horizontally across the platform. Notably, our focus was primarily the generalization of morality within groups based on different valenced impressions rather than whether generalization differences occur for individuals with varying degrees of groupness. Therefore, only grouped agents were present in our study.

There were six possible presentation orders for the three actors, and these orders were counterbalanced across participants. The first agent in the left position then moved to the center position, equidistant from both hills. Participants were instructed to predict whether this agent would help the climber climb the left hill or hinder the climber from climbing it by pressing the "J" or "F" button, respectively, on a standard keyboard. Once participants made their choice, the actual behavior of the agent (helping or hindering) was displayed as feedback, and the same sequence was presented for the second agent in the middle position. Finally, the third agent in the right position moved to the center position, equidistant from both hills, and participants predicted the agent's future behaviors. However, no feedback was provided for this agent. Hence, in each trial, participants were asked to predict the behaviors of three actors in a group, but only the future behaviors of the first two actors were revealed as feedback. If participants' prediction was correct, they received 1 point; otherwise, they lost 1 point. Participants were informed that the total points accumulated after completing the experiments would be exchanged for money. The first two actors were labeled "known" members since their action outcomes were observable, while the last actor was referred to as an "unknown" member because he or she was a newly encountered group member and his or her action outcome was unknown to the participants. Importantly, in every third trial, after revealing the behavior of the first or second actor or predicting the

behavior of the last actor, participants were asked to rate their moral impression of that specific actor on a scale from 0 (bad) to 100 (good) and the uncertainty of the reported impression from 0 (very certain) to 100 (very uncertain).

The experiment comprised two blocks with different moral valences, each consisting of 60 trials. The ratio of helping and hindering behaviors performed by agents in each block was determined based on previous research (Hackel et al., 2022). In the positive block, known group members acted as helpers and had a positive valence, with 70 % of the trials involving a helping behavior and 30 % involving a hindering behavior. In the negative block, known group members acted as hinderers and had a negative valence, with 70 % of the trials involving a hindering behavior and 30 % involving a helping behavior. The order of presentation of the two blocks was counterbalanced across participants, and at the beginning of each block, participants were informed that the actors' behaviors were reset. This meant that the agents in the second block were to be regarded as entirely new and unrelated to those in the first block, allowing us to manipulate different moral valences across blocks. In other words, participants were instructed not to associate impressions formed in the first block with those in the second. However, within each block, participants were told that the agents' identities remained consistent across all trials, and they were required to form impressions based on the agents' behaviors observed throughout the block.

2.1.4. Data analysis

To examine whether the middle value on the moral impression scale represented a neutral impression, a separate group of 36 participants (19 men and 17 women, mean age = 20.10 years, SD = 2.12) was recruited. As previously mentioned, participants viewed a picture of the first frame of the trial used for forming moral impressions. They were then asked to rate their moral impression of a randomly selected actor on a scale ranging from 0 (bad) to 100 (good). A one-sample *t* test revealed that the morality rating ($M = 51.6$, $SE = 1.3$) did not significantly differ from the middle value of the scale (i.e., 50; $t(35) = 1.20$, $p = 0.239$, $d = 0.20$, 95 % CI = [-0.13, 0.53]). Thus, the middle value of the scale was considered to indicate a neutral moral impression before a positive or negative impression of the individuals was formed.

To facilitate direct comparisons across positive and negative moral impressions and to compute a bias index that would symmetrically capture the pattern toward positivity or negativity, we rescaled the original ratings to a standardized −1 to + 1 scale during analysis. We calculated a bias index, which simply indicates the direction of the bias. Specifically, original morality values were scaled within the range of 0 (indicating a negative character) to 1 (indicating a positive character), aligning with the Bayesian modeling approach in our study. This scale directly corresponds to the probability of exhibiting positive behavior and represents the level of positive moral impression. Consequently, the level of negative moral impression was calculated by subtracting this value from 1. The positivity bias was then determined by subtracting the level of negative moral impression in the negative block from the level of positive moral impression in the positive block. Namely, the bias index was calculated as follows: Bias index = $M_{pos}/100 - (1 - M_{neg}/100)$, where M_{pos} represents the moral impression in the positive block and M_{neg} represents the moral impression in the negative block. This bias metric ranged from −1 to 1, where a positive value indicated a positivity bias in moral impression formation, and vice versa. Higher absolute values indicated more pronounced bias. To better illustrate bias in morality generalization, it is more appropriate to compare the bias index for the unknown member or the group against that of the known member than to compare simply against zero, as comparison against zero could be confounded by any inherent bias in the impressions of the observed known member. To provide a comprehensive overview of the original ratings, we conducted an analysis using the level of moral impression as the dependent variable, as detailed in the Supplementary Material, Fig. S2 and Fig. S3. The analysis of the original ratings yielded results consistent with those based on the bias index.

Behavioral prediction was used to provide participants with the opportunity to form individual impressions. Since feedback was provided for the two known members, their moral characters were gradually learned with exposure to prediction errors. To capture the learning rates for moral characters with different valences, we applied the Rescorla–Wagner reinforcement learning model to evaluate participants' learning behaviors during the task (Zhang et al., 2020). The learning rates for each known actor were computed using hierarchical Bayesian modeling of the 2-armed bandit task with the Rescorla–Wagner (delta) model, utilizing the R library hBayesDM (version 1.2.1) (Ahn et al., 2017). The learning rates for the two known members were then averaged for each participant.

The data were analyzed using R version 4.1.2 with the library bruceR (version 0.8.9) (Bao, 2022) and the package ggplot2 (version 3.4.2) (Wickham et al., 2016). The Greenhouse–Geisser epsilon was used to adjust the degrees of freedom when necessary (the uncorrected degrees of freedom are presented for simplicity), and post hoc contrasts were corrected with the Bonferroni method. For the analysis of variance, the partial eta squared (i.e., η_p^2) value was reported, and for *t* tests, Cohen's *d* was calculated. We provide the 95 % confidence intervals for the effect sizes.

2.1.5. Bayesian modeling

We used a computational model to simulate the mental process of generalizing individual moral impressions to the group level. We assumed that participants held representations of the causal relationships among the group moral impression, individual moral impressions and individual behaviors (Siegel et al., 2018; Baker et al., 2017). The mental process of morality generalization involved the reverse inference of group moral impressions based on this causal representation and the observed evidence of individual moral impressions. Specifically, we utilized a modified hierarchical generative model to replicate this causal representation and applied Bayesian methods to infer the potential priors related to group moral impressions.

An overview of the model is presented in Fig. 1c. The group moral impression at the very beginning was related to prior beliefs, whereas the group moral impression at each subsequent moment was affected only by the group moral impression at the previous moment. When analyzing the rating data, the impression values in the model were scaled within the range of 0 (indicating a negative character) to 1 (indicating a positive character). Therefore, higher values indicate stronger beliefs in positivity, with values greater than 0.5 implying a distinct inclination toward positivity. We represented the prior distribution of group moral impressions as a beta distribution with parameters α_{group} and β_{group}:

$$T_{group,0} \sim Beta(\alpha_{group}, \beta_{group})$$

To simulate the influence of previous group moral impressions on current group moral impressions, we employed a Gaussian process with variance parameter σ_{group}. Since the mean value in the Gaussian process spans the entire real number field and group moral impressions are constrained within the range of (0, 1), we utilized the sigmoid function and the inverse-sigmoid function to transform these two values:

$$T_{group,t} \sim F_{sigmoid}\left(Normal\left(F_{sigmoid}^{-1}\left(T_{group,t-1}\right), \sigma_{group}\right)\right)$$

$$F_{sigmoid}(x) = \frac{1}{1+e^{-x}},\ F_{sigmoid}^{-1}(x) = -\ln\left(1-\frac{1}{x}\right)$$

The second level of the model pertains to individual moral impressions. Individual moral impressions at the current moment were influenced by both individual moral impressions at the previous moment and group moral impressions at the current moment. We modeled this dynamic process using a Gaussian process with variance parameter σ_{ind}. Importantly, both current group moral impressions and previous indi-

vidual moral impressions affect the positive or negative moral impressions formed by individuals at the current moment but not the variance in the impressions. The current group moral impression and previous individual moral impression jointly contribute to the mean component of the Gaussian process with the weight parameter w_{ind}:

$$T_{ind,t} \sim F_{sigmoid}\left(Normal\left(F^{-1}_{sigmoid}\left(w_{ind} \times T_{ind,t-1} + (1 - w_{ind}) \times T_{group,t-1}\right), \sigma_{ind}\right)\right)$$

As with the approach used for determining group moral impressions, we described the prior distribution (for the initial value) of individual moral impressions as a beta distribution with parameters α_{ind} and β_{ind}.

The bottom level of the model illustrates the relationship between individual moral character and behavior. Potential behaviors were categorized as either positive (helping) or negative (hindering) and exhibited a binomial distribution, with the individual moral impression serving as the parameter:

$$A_t \sim Binomial\left(T_{ind,t}, 1 - T_{ind,t}\right)$$

To estimate participants' priors for group moral impressions, we employed importance sampling. Initially, for each participant, we independently sampled values for α_{group} and β_{group} from a uniform distribution. Since the initial group moral impression of any group should be the same before any behaviors have been observed, the same prior parameters were used when simulating the tasks in two blocks. Subsequently, for each individual the participants observed, we generated an individual moral impression based on these two prior parameters using the generative model. In a Gaussian distribution with the individual moral impressions reported by participants as the mean, the probability of the generated individual moral impressions was used as the weight for this sampling. We repeated this sampling procedure 1,000 times and computed the weighted mean of all the sampling results to estimate the prior group impressions of each participant. The value of prior group impressions ranged from 0 to 1. A value significantly different from 0.5 indicated that the group moral impression held by the participant before observing any behaviors was biased toward either positivity (greater than 0.5) or negativity (less than 0.5).

2.2. Results and discussion

We observed that participants effectively learned the morality of known group members. This finding was evident in the fact that the subjective rating of the helper's morality ($M = 64.6$, $SE = 1.1$; $t(35) = 13.08$, $p < 0.001$, $d = 2.18$, 95 % CI = [1.84, 2.52]) significantly exceeded the middle value (i.e., 50). In contrast, the subjective rating of the hinderer's morality was significantly lower than the middle value ($M = 36.3$, $SE = 1.1$; $t(35) = 11.97$, $p < 0.001$, $d = 1.99$, 95 % CI = [1.66, 2.33]).

Furthermore, when participants rated the known members ($M = 0.013$, $SE = 0.013$), the bias index did not significantly differ from 0 ($t(35) = 1.02$, $p = 0.316$, $d = 0.17$, 95 % CI = [-0.17, 0.51]), indicating that participants formed comparable intensities of moral impressions for differently valenced individuals. However, when the rated target was an unknown member ($M = 0.095$, $SE = 0.030$), the bias index exceeded 0 ($t(35) = 3.19$, $p = 0.003$, $d = 0.53$, 95 % CI = [0.19, 0.87]) and was greater than that calculated for the known members ($t(35) = 3.09$, $p = 0.004$, $d = 0.51$, 95 % CI = [0.18, 0.85]; Fig. 2a). This result suggests that participants exhibit a positivity bias in generalizing moral impressions to unknown group members, rather than demonstrating bias at the individual level. Based on Bayesian modeling, considering a neutral value of 0.5 indicating neutrality, the estimated prior leaned in a positive direction ($M = 0.533$; $SE = 0.012$; $t(35) = 2.70$, $p = 0.011$, $d = 0.45$, 95 % CI = [0.11, 0.79]; Fig. 3a), showing prior beliefs in positivity regarding groups.

Positivity bias in morality generalization cannot be attributed to a lack of opportunities to learn about negative moral characters or to uncertainty about learned moral impressions, as the rate of learning morality for hinderers ($M = 0.37$, $SE = 0.01$) was faster than that for helpers ($M = 0.07$, $SE = 0.01$; $t(35) = 25.62$, $p < 0.001$, $d = 4.27$, 95 % CI = [3.93, 4.61]; Fig. S5a), and the uncertainty for positive moral impressions of known members ($M = 34.8$, $SE = 2.9$) was not significantly different from that for negative moral impressions ($M = 36.2$, $SE = 2.9$; $t(35) = 0.82$, $p = 0.419$, $d = 0.14$, 95 % CI = [-0.20, 0.48]; Fig. S6a).

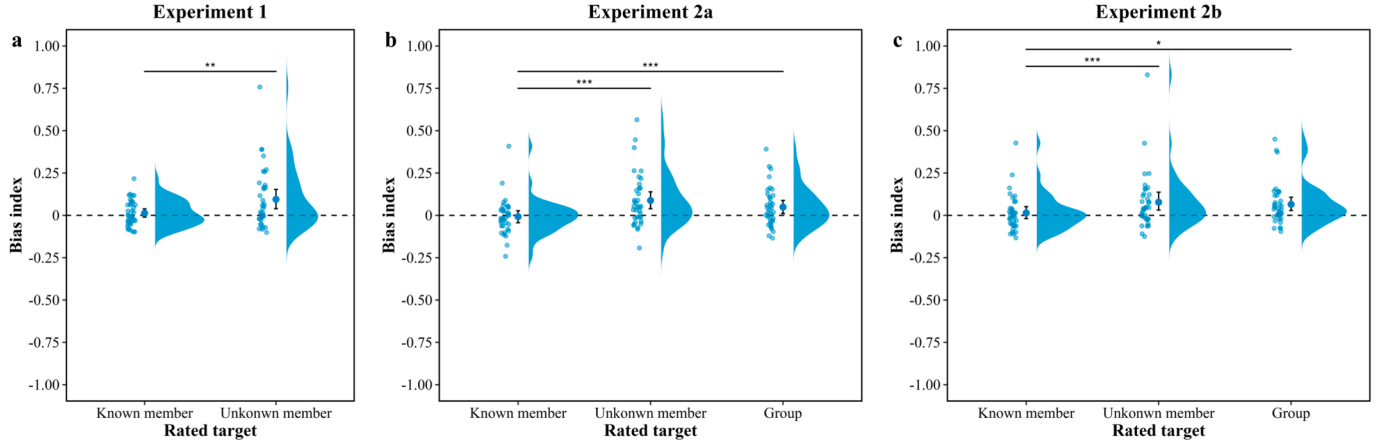

Fig. 2. Bias indices in different experiments. (a) Bias indices of known and unknown members in Experiment 1; (b) and (c) bias indices of known members, unknown members, and groups in Experiments 2a and 2b, respectively. The small dots represent the mean values of each participant in each condition, while the large dots indicate the means of the corresponding conditions. The black vertical lines indicate the bootstrapped confidence intervals (CIs) of the means. ***$p < 0.001$; **$p < 0.010$; *$p < 0.050$.

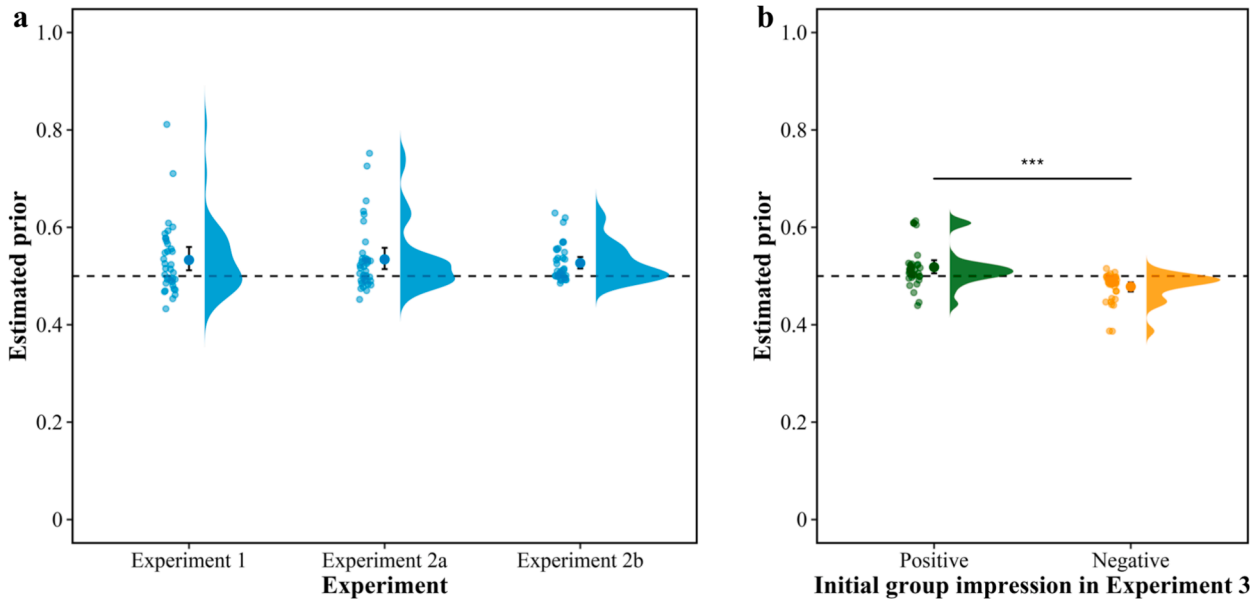

Fig. 3. Estimated priors in different experiments. (a) Estimated priors in Experiment 1, Experiment 2a and Experiment 2b; (b) estimated priors in Experiment 3. The small dots represent the mean values of each participant in each condition, while the large dots indicate the means of the corresponding conditions. The black vertical lines indicate the bootstrapped confidence intervals (CIs) of the means. ***$p < 0.001$.

3. Experiment 2a

The results of Experiment 1 revealed that participants formed equally strong negative and positive moral impressions of individuals engaging in helping and hindering behaviors within our dynamic learning context. They also provided initial evidence of a positivity bias in morality generalization. However, this bias was observed primarily in the generalization of moral impressions from a known individual to an unknown individual. Whether this approach truly captures generalization at the group level requires an inquiry into moral impressions concerning groups. To directly measure the generalization of moral impressions from individuals to groups and replicate the findings from Experiment 1, we conducted Experiment 2a using the same paradigm as Experiment 1 but with an additional task in which participants directly reported their subjective ratings of their group moral impressions.

3.1. Methods

A new group of 36 participants (17 men and 19 women; mean age = 20.40 years, SD = 2.17) was recruited for Experiment 2a. The procedure in Experiment 2a was identical to that in Experiment 1 except that after every third trial, participants were asked to provide ratings of their moral impression of the group to which the three actors belonged.

3.2. Results and discussion

Experiment 2a revealed that participants effectively formed individual impressions for both known helpers and hinderers. Specifically, the reported moral impressions of the helpers ($M = 64.1$, $SE = 1.5$; $t(35) = 9.61$, $p < 0.001$, $d = 1.60$, 95 % CI = [1.26, 1.94]) and the hinderers ($M = 35.2$, $SE = 1.6$; $t(35) = 14.83$, $p < 0.001$, $d = 1.54$, 95 % CI = [1.21, 1.88]) were significantly different from the middle value of 50.

Again, no bias was observed in the formation of individual moral impressions, as similar extents of positive and negative moral characters were assigned to known good and bad members ($M = -0.007$, $SE = 0.018$; $t(35) = -0.42$, $p = 0.678$, $d = -0.07$, 95 % CI = [-0.41, 0.27]). However, the bias indices for the unknown group member ($M = 0.088$, $SE = 0.026$; $t(35) = 3.36$, $p = 0.002$, $d = 0.56$, 95 % CI = [0.22, 0.90]) and the entire group ($M = 0.049$, $SE = 0.020$; $t(35) = 2.48$, $p = 0.018$, $d = 0.41$, 95 % CI = [0.07, 0.75]) were both greater than 0. Furthermore, the bias index was greater for the unknown member ($t(35) = 4.88$, $p < 0.001$, $d = 1.04$, 95 % CI = [0.50, 1.57]) and the group ($t(35) = 4.25$, $p < 0.001$, $d = 0.61$, 95 % CI = [0.25, 0.98]) than for the known group member (Fig. 2b). This finding suggests that positive moral impressions are generalized more to groups than are negative moral impressions, indicating a group-level positivity bias in morality generalization.

As in Experiment 1, in Experiment 2a, a clear preference for positivity was revealed in the estimated prior ($M = 0.534$; $SE = 0.011$; $t(35) = 2.99$, $p = 0.005$, $d = 0.50$, 95 % CI = [0.16, 0.84]; Fig. 3a). Moreover, the simulated group impressions generated from Bayesian models showed a similar positivity bias as the subjective ratings, as did the subsequent experiments (see the Supplementary Material).

Participants learned hinderers' moral characters ($M = 0.30$, $SE = 0.02$) more quickly than they learned helpers' moral characters ($M = 0.14$, $SE = 0.01$; $t(35) = 8.12$, $p < 0.001$, $d = 1.35$, 95 % CI = [1.01, 1.69]; Fig. S5b), indicating that participants had ample opportunity to learn about negative moral character. Moreover, participants did not exhibit discriminative uncertainty in forming differently valenced impressions (positive: $M = 29.5$, $SE = 2.3$; negative: $M = 30.2$, $SE = 2.4$; $t(35) = 0.90$, $p = 0.700$, $d = 0.07$, 95 % CI = [-0.27, 0.40]; Fig. S6b).

4. Experiment 2b

In the previous experiments, both the positive and negative blocks featured the same group of actors. Despite informing participants that the agents' behaviors in the two blocks were reset, this setting could introduce potential interference in the learned moral valence between the blocks. To assess the generalizability of positivity bias in the generalization of morality to groups, we conducted Experiment 2b using the same paradigm as Experiment 2a, except that different groups of actors (indicated with different colors) were used for the positive and negative blocks.

4.1. Methods

A new group of 36 participants (15 men and 21 women; mean age = 21.75 years, SD = 1.84) was recruited for Experiment 2b. The procedure in Experiment 2b was identical to that in Experiment 2a, except for the

use of different groups of actors (indicated with different colors) for the positive and negative blocks.

4.2. Results and discussion

Participants successfully learned moral characters for both known helpers and hinderers. Specifically, the reported moral impressions of the helpers ($M = 67.7$, $SE = 1.5$; $t(35) = 12.15$, $p < 0.001$, $d = 2.03$, 95 % CI = [1.69, 2.36]) and the hinderers ($M = 33.7$, $SE = 1.3$; $t(35) = 12.23$, $p < 0.001$, $d = 2.04$, 95 % CI = [1.70, 2.38]) were significantly different from the middle value of 50.

Experiment 2b produced results similar to those of Experiment 2a in terms of impression formation (Fig. 2c). Specifically, no discernible bias emerged when forming individual moral impressions ($M = 0.014$, $SE = 0.018$; $t(35) = 0.80$, $p = 0.431$, $d = 0.13$, 95 % CI = [-0.21, 0.47]). However, the bias indices for the unknown group members ($M = 0.078$, $SE = 0.028$; $t(35) = 2.76$, $p = 0.009$, $d = 0.46$, 95 % CI = [0.12, 0.80]) and the entire group ($M = 0.065$, $SE = 0.021$; $t(35) = 3.16$, $p = 0.003$, $d = 0.53$, 95 % CI = [0.19, 0.86]) were both greater than 0. Furthermore, the bias index was greater for the unknown member ($t(35) = 2.49$, $p = 0.054$, $d = 0.54$, 95 % CI = [-0.01, 1.08]; in Experiment 2b, with Tukey correction: $p = 0.046$, $d = 0.41$, 95 % CI = [0.07, 0.75]) and for the group ($t(35) = 4.51$, $p < 0.001$, $d = 0.43$, 95 % CI = [0.19, 0.66]) than for the known group member. Hence, participants exhibited a stronger tendency to generalize positive moral impressions to groups than negative impressions, revealing a pervasive positivity bias in how they generalized morality.

Similarly, in Experiment 2b, participants demonstrated a clear prior belief in the positivity of groups, with the estimated prior learning in the positive direction ($M = 0.527$; $SE = 0.006$; $t(35) = 4.23$, $p < 0.001$, $d = 0.70$, 95 % CI = [0.37, 1.04]; Fig. 3a).

As in Experiment 2a, participants learned negative moral characters ($M = 0.30$, $SE = 0.02$) more quickly than positive moral characters ($M = 0.14$, $SE = 0.01$; $t(35) = 8.12$, $p < 0.001$, $d = 1.35$, 95 % CI = [1.01, 1.69]; Fig. S5c), indicating that participants had ample opportunity to learn about negative moral character. Furthermore, participants' uncertainty in their impressions was not influenced by the characters' positive ($M = 31.7$, $SE = 3.0$) or negative ($M = 31.8$, $SE = 2.3$) moral character ($t(35) = 0.06$, $p = 0.957$, $d < 0.01$, 95 % CI = [-0.33, 0.35]; Fig. S6c).

5. Experiment 3

The results of Experiments 1 and 2 suggest that positive moral impressions are more readily generalized to groups than are negative impressions. This positivity bias can be attributed to a robust prior belief in the positivity of groups. This explanation is supported by the absence of bias in individual moral impressions as evidence supporting morality generalization, combined with the Bayesian modeling of prior beliefs about groups showing positivity. In Experiment 3, we further investigated the role of prior beliefs in morality generalization by manipulating participants' initial moral impressions of evaluated groups. If the positivity bias in morality generalization is rooted in prior beliefs, manipulating this belief through preexposure to group moral information before learning about individual moral characters should alter the positivity bias. Specifically, after acquiring initial information indicating a positive group moral character, the positivity bias should persist. Conversely, initial negative moral information about the group might lead to a change in, potentially even a reversal of, asymmetric morality generalization, indicating a negativity bias.

5.1. Methods

A total of 72 additional participants (31 men and 41 women, mean age = 20.88 years, SD = 2.21) participated in Experiment 3, with 36 participants assigned to each of the two between-subjects conditions. The procedure and design closely mirrored those of Experiment 2b, with an additional phase occurring at the beginning of each block to form an initial group impression (see Fig. 1b and Video S2). In this phase, as in previous studies (63–65), participants were presented with a word composed of two Chinese characters representing a moral trait. This word was displayed above the three actors after the introduction of the social groups. Participants were instructed that the word described the moral character of the group of three actors and that they should pay attention to the moral character associated with the group, as they would be questioned about the presented words and social groups later (however, no questions were asked after the testing). Two types of moral characters were presented: positive (e.g., 善良, "shanliang," kind; 友好, "youhao," friendly; 友善, "youshan," amiable; 友爱, "youai," affectionate; 热心, "rexin," enthusiastic) and negative (蛮横, "manheng," savage; 霸道, "badao," overbearing; 自私, "zisi," selfish; 冷漠, "lengmo," indifferent; 恶劣, "elie," nasty). Each type comprised five words. To generate this list of positive and negative words describing moral characters, an additional group of 30 participants was recruited to watch animations of helping and hindering behaviors. Participants were asked to list at least five words describing the actors' traits related to moral character. The five most frequently reported words were selected for each moral valence. Participants rated their moral impressions of the group before and after acquiring initial information about the group moral character.

5.2. Results and discussion

5.2.1. Initial group impression

The patterns of acquiring group moral impressions presented before each block were similar for the two groups, and presenting a group first or second did not affect the initially formed group impressions (see

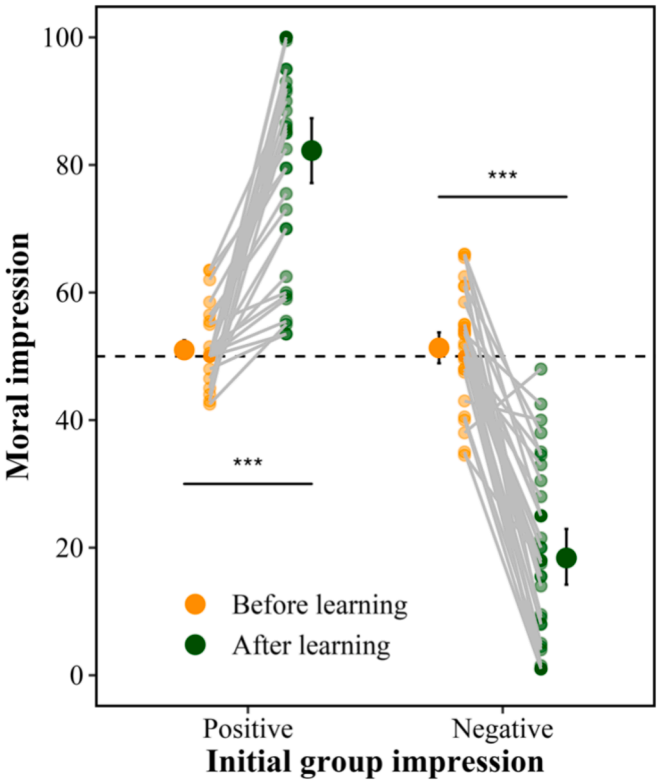

Fig. 4. Results of the task of acquiring an initial group moral impression in Experiment 3. The small dots indicate the mean value for each participant in each condition and are connected for the paired conditions. The large dots denote the means, and the black vertical lines indicate the bootstrapped confidence intervals (CIs) of the means. ***$p < 0.001$.

Supplementary Material for details). Hence, we collapsed the data without considering the factor of the presented group for further analyses.

The initial group impression was significantly influenced by the learning phase (Fig. 4), as evident in the interaction effect of the repeated-measures ANOVA with the learning phase and the valence of the initial group impression as factors ($F(1, 70) = 291.59$, $p < 0.001$, $\eta_p^2 = 0.81$, 95 % CI = [0.74, 0.85]). Simple effects analysis revealed that in the condition where information indicating a positive moral character was presented, the initial group impression was more positive ($M = 82.2$, $SE = 2.7$) after learning than before learning ($M = 51.0$, $SE = 0.8$; $t(35) = 11.08$, $p < 0.001$, $d = 1.85$, 95 % CI = [1.51, 2.19]). Conversely, in the condition where information indicating a negative moral character was presented, the initial group impression was more negative ($M = 18.4$, $SE = 2.2$) after learning than before learning ($M = 51.3$, $SE = 1.3$; $t(35) = 13.26$, $p < 0.001$, $d = 2.21$, 95 % CI = [1.87, 2.55]). Additionally, before learning, the group impressions were not significantly different from the mid-value (i.e., 50) of the scale ($ts \leq 1.22$, $ps \geq 0.232$, $ds \leq 0.20$), indicating a neutral moral impression of the group. The absolute change between acquiring initial positive and negative impressions was nearly identical according to the CIs. Hence, participants effectively formed an initial group moral impression in the first phase. In addition, changes in group moral impressions after initially acquiring positive and negative impressions were nearly identical.

5.2.2. Bias index

Concerning the bias index in the second learning phase, a 2 (initial group moral impression as a between-subjects variable: positive vs. negative) × 3 (rated target as a within-subject variable: unknown member vs. known member vs. group) ANOVA revealed that the two-way interaction effect reached significance ($F(2, 140) = 4.43$, $p = 0.025$, $\eta_p^2 = 0.06$, 95 % CI = [0.00, 0.15]). This finding indicates that the initially acquired group moral impression modulates subsequent moral impression formation. To further explore this effect, a simple effect analysis was conducted.

In cases where the initially acquired group moral impression was positive, the bias index significantly differed across the three types of rated targets (Fig. 5a; $F(2, 70) = 8.42$, $p = 0.003$, $\eta_p^2 = 0.19$, 95 % CI = [0.05, 0.36]). The bias index was higher for the unknown member ($M = 0.091$, $SE = 0.025$; $t(35) = 3.14$, $p = 0.010$, $d = 0.68$, 95 % CI = [0.14, 1.23]) and the overall group ($M = 0.048$, $SE = 0.020$; $t(35) = 2.84$, $p = 0.023$, $d = 0.28$, 95 % CI = [0.03, 0.53]) than for the known member ($M = 0.018$, $SE = 0.021$). Furthermore, the bias indices for the unknown member ($t(35) = 3.68$, $p < 0.001$, $d = 0.61$, 95 % CI = [0.27, 0.95]) and the group ($t(35) = 2.38$, $p = 0.023$, $d = 0.40$, 95 % CI = [0.06, 0.74]) were greater than 0. However, no significant bias emerged when moral impressions of known members were formed ($t(35) = 0.86$, $p = 0.395$, $d = 0.14$, 95 % CI = [-0.19, 0.48]). Therefore, when initial impressions of groups are positive, indicating that people hold a prior belief in the

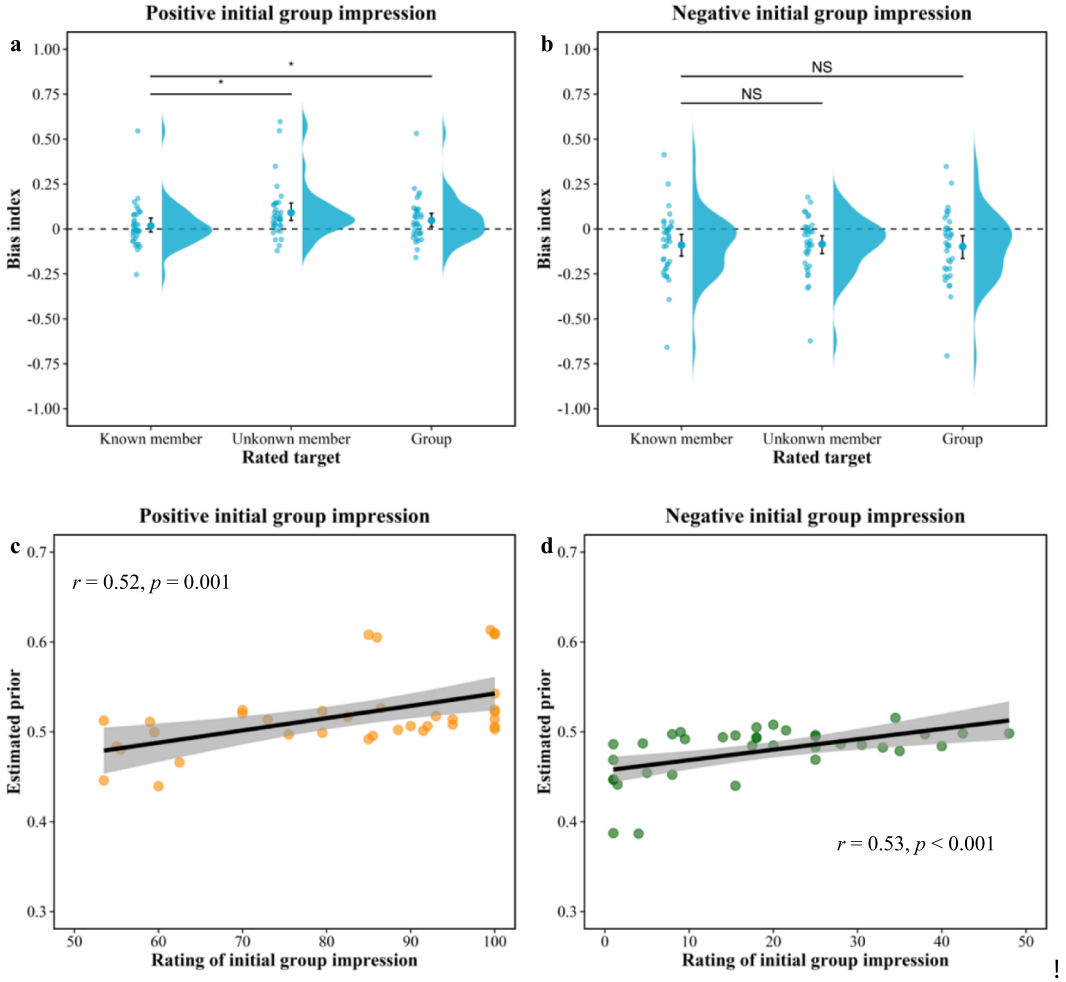

Fig. 5. Bias indices and correlation between estimated prior and initial group impression ratings computed in Experiment 3. Bias indices of known members, unknown members, and groups are displayed for initially acquired positive (a) and negative (b) group moral impressions. Pearson's correlation between the rating of initially acquired group moral impressions from the first phase and the estimated prior in the second phase is shown for initially acquired positive (c) and negative (d) group moral impressions. The small dots represent the mean values of each participant in each condition, while the large dots indicate the means of the corresponding conditions. The black vertical lines indicate the bootstrapped confidence intervals (CIs) of the means. *$p < 0.050$; NS, not significant.

positivity of groups, the positivity bias in generalizing moral impressions from individuals to groups is still present. However, when the initially acquired group moral impression was negative, different results were obtained (Fig. 5b). For all the rated targets (known member: $M = -0.089$, $SE = 0.031$; $t(35) = -2.89$, $p = 0.007$, $d = -0.48$, 95 % CI = [-0.82, −0.14]; unknown member: $M = -0.084$, $SE = 0.026$; $t(35) = -3.24$, $p = 0.003$, $d = -0.54$, 95 % CI = [-0.88, −0.20]; group: $M = -0.098$, $SE = 0.032$; $t(35) = -3.04$, $p = 0.004$, $d = -0.51$, 95 % CI = [-0.85, −0.17]), the bias indices were significantly lower than 0, indicating a negativity bias. Additionally, the extent of the bias did not significantly differ across the three rated targets ($F(2, 70) = 0.50$, $p = 0.568$, $\eta_p^2 = 0.01$, 95 % CI = [0.00, 0.09]). Hence, when the initial group moral impression is negative, indicating a change in the prior belief about groups from positivity to negativity, the observed asymmetry in morality generalization is reversed, indicating a negativity bias.

5.2.3. Estimated prior

The estimated prior when the initially acquired group moral impression was positive was more positive ($M = 0.518$; $SE = 0.007$) than the estimated prior when the initially acquired group moral impression was negative ($M = 0.478$; $SE = 0.005$; $t(70) = 4.66$, $p < 0.001$, $d = 1.10$, 95 % CI = [0.63, 1.57]; Fig. 3b). Furthermore, the estimated prior for initially acquired positive group moral impressions was greater than 0.5 ($t(35) = 2.60$, $p = 0.014$, $d = 0.43$, 95 % CI = [0.09, 0.77]), indicating a positive prior. Conversely, the estimated prior for initially acquired negative group moral impressions was less than 0.5 ($t(35) = 4.41$, $p < 0.001$, $d = 0.73$, 95 % CI = [0.40, 1.07]), suggesting a negative prior.

To explore the potential influence of prior group impressions on subsequent morality generalization, Pearson's correlation analysis was performed to examine the relationship between the subjective ratings of the initially acquired group impressions from the first phase and those estimated prior to the second phase. The results revealed that a more positive initial group moral impression corresponded to greater positive prior constraining morality generalization ($r = 0.52$, $p = 0.001$, 95 % CI = [0.23, 0.66]; Fig. 5c), and a more negative previous group moral impression correlated with greater negative prior constraining morality generalization ($r = 0.53$, $p < 0.001$, 95 % CI = [0.24, 0.73]; Fig. 5d). This finding suggests that the valence of the prior group moral impression determines the extent of bias in morality generalization from individuals to groups.

5.2.4. Learning rate

The 2 (initial group moral impression as a between-subjects variable: positive vs. negative) × 2 (moral valence as a within-subject variable: positive vs. negative) ANOVA revealed a significant main effect of initial group moral impression ($F(1, 70) = 7.99$, $p = 0.006$, $\eta_p^2 = 0.10$, 95 % CI = [0.02, 0.22]), showing that the learning rate after initially acquiring negative group impressions ($M = 0.36$, $SE = 0.02$) was faster than that after initially acquiring positive group impressions ($M = 0.29$, $SE = 0.02$). The significant main effect of moral valence ($F(1, 70) = 8.41$, $p = 0.005$, $\eta_p^2 = 0.11$, 95 % CI = [0.02, 0.23]) showed that the learning the negative moral character of the hinderers ($M = 0.35$, $SE = 0.02$) occurred more quickly than in the case of the helpers ($M = 0.29$, $SE = 0.01$). Moreover, the interaction effect was significant ($F(1, 70) = 36.00$, $p < 0.001$, $\eta_p^2 = 0.34$, 95 % CI = [0.20, 0.47]). Simple effects analysis revealed that when the initially acquired group moral impression was positive, the hinderer's morality was learned more quickly ($M = 0.38$, $SE = 0.02$) than the helper's morality ($M = 0.19$, $SE = 0.02$; $t(35) = 6.54$, $p < 0.001$, $d = 1.09$, 95 % CI = [0.75, 1.43]; Fig. S5d). However, when the initially acquired group moral impression was negative, the hinderer's morality was learned more slowly ($M = 0.33$, $SE = 0.03$) than the helper's morality ($M = 0.39$, $SE = 0.02$; $t(35) = -2.12$, $p = 0.042$, $d = -0.35$, 95 % CI = [-0.69, −0.01]). In summary, when the initially formed group moral impression does not match the newly learned group member's morality, the learning rate is enhanced.

5.2.5. Uncertainty

Concerning the uncertainty of moral impressions for learned individuals (i.e., known group members), the 2 (initial group moral impression as a between-subjects variable: positive vs. negative) × 2 (moral valence as a within-subject variable: positive vs. negative) ANOVA revealed a significant main effect of moral valence ($F(1, 70) = 8.15$, $p = 0.006$, $\eta_p^2 = 0.10$, 95 % CI = [0.02, 0.23]), showing that the moral impressions of negative characters of helpers ($M = 30.9$, $SE = 2.0$) were more uncertain than those of hinderers ($M = 27.4$, $SE = 2.0$; Fig. S6d). Neither the main effect of initial group moral impression ($F(1, 70) = 0.46$, $p = 0.499$, $\eta_p^2 < 0.01$, 95 % CI = [0.00, 0.07]) nor the interaction effect ($F(1, 70) = 0.86$, $p = 0.358$, $\eta_p^2 = 0.01$, 95 % CI = [0.00, 0.09]) reached significance.

6. General discussion

These experiments suggest that positive moral impressions are generalized from individuals to groups to a greater extent than negative moral impressions are, indicating a positivity bias in generalizing moral impressions from individuals to groups (i.e., group-level positivity bias). The findings from Bayesian modeling and the manipulation of prior beliefs regarding group moral impressions indicate that this group-level positivity bias can be attributed to a prior belief in the positivity of groups.

The positivity bias in morality generalization observed was not due to the limited number of chances to learn about negative character or uncertainty about individual moral impressions. In fact, we found that morality learning occurred more quickly for hinderers than for helpers, and the uncertainty in positive impressions of known members was not significantly different from that in negative impressions. However, research has indicated that negative impressions are often more uncertain and susceptible to change (Siegel et al., 2018; 2020). While previous studies have focused on moral impression formation for isolated individuals, our research studies individuals in a group context. Similar morality between two observed members may validate their respective morality, leading participants to have comparable levels of certainty in their evaluations. This question is beyond the scope of the current research, and future studies should explore this intriguing topic further.

One might argue that judgments about an unknown group member could reflect baseline expectations about general behavior in the task, rather than generalization from the observed members. However, several findings challenge this interpretation. First, if baseline expectations were solely responsible, then we would expect similar biases for both known and unknown members. However, we observed a specific positivity bias only for the unknown member, whereas no bias was found in evaluations of known members. Second, the proportion of participants who predicted that the unknown member would perform the same behavior as the known members was significantly greater than 50 % in both the positive and negative conditions across all experiments (see Supplementary Material), providing supporting evidence for generalization from the observed agents. Moreover, moral impression ratings for the unknown member changed over time in both the positive and negative blocks. By collapsing the data across the three experiments (Experiments 1, 2a, and 2b), we found that in the positive block, moral evaluations became more positive from the first half of trials ($M = 59.5$, $SE = 1.0$) to the last half ($M = 62.2$, $SE = 1.0$; $t(107) = 3.27$, $p = 0.001$, $d = 0.31$, 95 % CI = [0.12, 0.51]). In contrast, in the negative block, the evaluations became more negative (first half: $M = 49.0$, $SE = 1.2$; last half: $M = 46.7$, $SE = 1.3$; $t(107) = 2.63$, $p = 0.010$, $d = 0.25$, 95 % CI = [−0.24, −0.42]). These changes suggest that participants incorporated the learned impressions of individual group members into their evaluations of the group. Finally, when we experimentally manipulated prior beliefs about the group (e.g., initially presenting the group as negative), the positivity bias in group impression formation was reversed into a negativity bias. This result provides direct evidence that generalizations were shaped by observed information and were not driven by a stable

baseline positivity expectation.

These findings do not imply that negative individual morality cannot be generalized to groups. However, this process appears to be weaker than that for positive morality. Indeed, compared with a neutral level of morality, a significant negative group moral impression was evident in all the experiments, except for Experiment 1 (see Supplementary Material, Fig. S1). While people form similar strong moral impressions of individual helpers and hinderers, the resulting morality generalization to groups is more pronounced for helpers than for hinderers. Consistent with the results of previous studies, this finding suggests that group moral impressions are formed differently from individual moral impressions, as the former involves representing multiple individual moral characters (Crawford et al., 2002; Hackel et al., 2022; Shin & Niv, 2021). The present study contributes to the literature by introducing the role of moral valence in group impression formation and shedding light on positivity bias and its underlying mechanisms. Our findings seem inconsistent with previous results establishing a negativity bias (Ratliff & Nosek, 2011). However, it is important to highlight that this negativity bias was primarily reported for moral impression formation at the individual level with an unequal mix of positive (common) and negative (rare) events with textual descriptions of a few behaviors (Anderson, 1965; Skowronski & Carlston, 1989; Unkelbach et al., 2020). When evaluating individuals in isolation, negative information tends to be more diagnostic, and thus, it carries more weight, which is a hallmark of the well-documented negativity bias. This bias manifests in various ways, such as the stronger influence of negative traits on forming individual impressions of personality and the enhanced recognition and recall of negative information (Unkelbach et al., 2020). In contrast, when forming impressions about groups or generalizing morality to unknown group members, the belief in group positivity tends to dominate unless it is strongly contradicted by evidence. Supporting this distinction, although the positive and negative moral impressions for individuals were comparably strong (i.e., the bias index did not significantly differ from zero), our data showed a faster learning rate for the moral impressions of hinderers than for those of helpers, which is consistent with negativity bias at the individual level. Nevertheless, a positivity bias emerged specifically in the formation of group moral impressions and the generalization of morality to unknown group members. Moreover, in our experiments, moral impressions were dynamically updated, and similar moral impressions of group members were formed, akin to observing behaviors in daily life involving groups.

Our findings suggest that people do not simply treat similar individual morality as evidence for inferring group moral impressions; instead, they hold beliefs in the positivity of groups prior to making any observations of a group. This finding supports the claim that conceptual representations of categories depend on the prior value assigned to members within that category, in addition to their statistical frequency (Knobe & Cushman, 2023). Our study further suggests that this value may reflect a stronger prior belief in group-level positivity, which constrains inferences about social categories. The prior belief in group positivity may serve as a default parameter guiding the formation of group impressions. Accordingly, even when participants observe group members with similarly valenced characters, the generalized group impression is stronger for a group of helpful actors than for a group of harmful actors. Therefore, morality generalization to groups does not merely involve weighting individual moral impressions but rather considers both available priors and accumulated individual impressions through Bayesian inference. While numerous previous studies have highlighted the role of both prior beliefs and evidence in inferring group characteristics (Gershman & Cikara, 2023; Krueger & Clement, 1996; McGarty et al., 2002; Yon & Frith, 2021), this study contributes to the literature by explicitly identifying which prior is held in the process of moral generalization. This prior may be rooted in intuitive sociology, a natural and instinctive way of understanding group dynamics, and may provide a heuristic that maintains the moral impression of the goodness of human beings. Traditionally, stereotypes have been used to emphasize prior beliefs about specific social groups (Chan et al., 2012; Delplanque et al., 2019; Jussim et al., 2015; King et al., 2021). However, the prior in our study was an initial assumption of group moral impressions, which exert effects on the process of generalizing individual-level impression to group-level impression. A posterior group impression formed based on this prior, that is, a stereotype, can be generated after collecting information. We focused on the first prior rather than the priors generated after exposure to group members.

The prior belief in the positivity of groups can be updated after acquiring new moral information about the target group, as demonstrated in Experiment 3. Notably, when participants were initially presented with a negative group impression (i.e., a negative stereotype), their posterior group impression formed after observing group members' behaviors exhibited negativity bias. This finding further underscores the pivotal role of prior beliefs in how moral valence influences morality generalization from individuals to groups. Crucially, an initially acquired negative group impression also influenced individual moral impression formation, with higher morality attributed to hinderers than to helpers. However, this effect was absent when the initial group impression was positive. Thus, negative initial group impressions have more profound impacts on individual moral impression formation than positive initial group impressions, leading to negativity biases at both the individual and group levels. These biases can mutually reinforce each other, solidifying negative stereotypes. Interestingly, forming an initial positive group moral impression did not strengthen the default prior belief much, showing that, compared with Experiment 2b, the estimated priors did not significantly differ from each other ($t(70) = 0.89$, $p = 0.376$, $d = 0.21$, 95 % CI = [-0.26, 0.67]). It is possible that an initial positive group moral impression does not challenge a positive prior. Hence, the default prior may undergo substantial revision only when it is strongly contradicted by incoming information. This aligns with recent work emphasizing the role of prior beliefs in modulating impression updating (Park et al., 2021; Kim et al., 2021). For example, Park et al. (2021) demonstrated that the right temporoparietal junction tracks social prediction errors in relation to different strengths of prior expectations about others. Similarly, Kim et al. (2021) found that violations of strong versus weak priors elicit different neural and behavioral responses. Moreover, this default prior for a group likely influences the process through which individual-level impressions of observed members are integrated to form group-level impressions. Hence, when an agent displayed no explicit behavior and was evaluated in isolation without any group context, participants tended to report a neutral moral impression, as demonstrated by our pilot study. This result suggestion is further evidenced by participants' group moral impressions, which did not differ from the mid-value prior to being presented with positive or negative moral information about the groups. However, further research is needed to address these speculations.

Additionally, emphasizing the role of the prior belief in the positivity of groups in forming group-level moral impressions does not exclude motivational explanations, such as people avoiding the perception of stereotyping groups based on limited (negative) observations. Our conclusion regarding prior beliefs was supported by Bayesian modeling and the manipulation of participants' prior beliefs. It is possible that both motivational factors and cognitive priors contribute to the formation of moral impressions about groups. This research has several limitations. First, we utilized computer animation featuring nonhuman agents in our research. While this approach enabled strict control over potential confounding factors, enabling a correspondingly strict examination of moral impression formation, it is imperative to assess the external applicability of our findings using recorded human interactions in future studies. Second, we defined positive (negative) characters in terms of helping (hindering) behaviors. However, there are various other expressions of these characteristics, such as making donations or bullying. Third, due to spatial constraints within the visual scene, our study could include only three agents in each group. As group size often plays a pivotal role in group dynamics (Bond, 2005), future

investigations should consider incorporating different manifestations of positive and negative characters, as well as larger groups of agents, to further validate our findings. Fourth, although the current study sought to minimize baseline confounds through design features (e.g., contrasting known vs. unknown members, manipulating prior beliefs about groups), future research could strengthen the morality generalization account by incorporating a non-group-member condition, where participants occasionally predict the behavior of novel agents who do not belong to the learned group. Incorporating such a condition would help more directly rule out baseline explanations. Finally, in the current computational model, we assume that the prior and each piece of information play the same role in forming moral impressions. This is an idealized assumption, and in daily life, the importance of the prior may vary depending on the actual situation and experience of the group, such as experiencing high levels of intergroup conflict, and the weight of the evidence may change according to how much insight it can offer. How humans flexibly adjust the importance of priors and evidence during morality generalization needs to be further explored in future research.

In conclusion, our findings provide substantial evidence that people possess a prior belief in the positivity of groups. Consequently, this inclination exerts effects on the process of generalizing individual-level impression to group-level impression, indicating that positive moral impressions are more strongly generalized from individuals to groups than are negative moral impressions. Therefore, social groups not only impact the identification of shared moral character among their members but also contribute to prior knowledge that underpins the formation of group moral impressions. This prior likely emanates from the intuitive comprehension of social dynamics and functions, leading to the formation of a default belief applied in social evaluations.

Author Note.

This research was supported by the National Natural Science Foundation of China (Grant No. 32371090). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The study titled “Knowledge about social groups” was approved by the Research Ethics Board of the Department of Psychology at Ningbo University and was performed in accordance with the relevant guidelines and regulations. All data, analysis codes and videos showing different conditions have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/xuc83/?view_only=7aebad94347346fe92a30daac7e6b29b. No study designs were preregistered.

CRediT authorship contribution statement

Hui Huang: Visualization, Investigation, Conceptualization, Writing – original draft, Methodology, Data curation, Writing – review & editing, Formal analysis. **Jiecheng Huangliang:** Methodology, Formal analysis, Conceptualization, Writing – original draft, Data curation, Writing – review & editing, Visualization, Investigation. **Qinglu Xiao:** Methodology, Formal analysis, Writing – original draft, Investigation. **Ting Zhao:** Methodology, Formal analysis, Writing – original draft, Investigation. **Qingqing Ye:** Writing – original draft, Investigation, Methodology, Formal analysis. **Haokui Xu:** Writing – original draft, Investigation, Data curation, Writing – review & editing, Formal analysis, Conceptualization, Methodology. **Jun Yin:** Funding acquisition, Conceptualization, Writing – original draft, Supervision, Data curation, Writing – review & editing, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cogsys.2025.101396.

Data availability

I have shared the link to the data in the article.

References

Ahn, W. Y., Haines, N., & Zhang, L. (2017). Revealing neurocomputational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *Computational Psychiatry (Cambridge, Mass), 1*, 24–57.

Alves, H., Koch, A., & Unkelbach, C. (2017a). The “common good” phenomenon: Why similarities are positive and differences are negative. *Journal of Experimental Psychology: General, 146*(4), 512–528.

Alves, H., Koch, A., & Unkelbach, C. (2017b). Why good is more alike than bad: Processing implications. *Trends in Cognitive Sciences, 21*(2), 69–79.

Alves, H., Unkelbach, C., Burghardt, J., Koch, A. S., Krüger, T., & Becker, V. D. (2015). A density explanation of valence asymmetries in recognition memory. *Memory & Cognition, 43*, 896–909.

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology, 70*(4), 394–400.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour, 1*(4), 1–10.

Bao, H. W. S. (2022). bruceR: Broadly useful convenient and efficient R functions. *R Package Version*, (8), 9.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323–370.

Bond, R. (2005). Group size and conformity. *Group Processes & Intergroup Relations, 8*(4), 331–354.

Brambilla, M., Sacchi, S., Rusconi, P., & Goodwin, G. P. (2021). The primacy of morality in impression development: Theory, research, and future directions. *In Advances in experimental social psychology, 64*, 187–262.

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science, 3*(1), 14–25.

Chan, W., McCrae, R. R., De Fruyt, F., Jussim, L., Löckenhoff, C. E., De Bolle, M., Costa, P. T., Jr, Sutin, A. R., Realo, A., & Allik, J. (2012). Stereotypes of age differences in personality traits: Universal and accurate? *Journal of Personality and Social Psychology, 103*(6), 1050–1066.

Crawford, M. T., Sherman, S. J., & Hamilton, D. L. (2002). Perceived entitativity, stereotype formation, and the interchangeability of group members. *Journal of Personality and Social Psychology, 83*(5), 1076–1094.

Crockett, M. J., Everett, J. A. C., Gill, M., & Siegel, J. Z. (2021). The relational logic of moral inference. *In Advances in Experimental Social Psychology, 64*, 1–64.

Delplanque, J., Heleven, E., & Van Overwalle, F. (2019). Neural representations of groups and Stereotypes using fMRI repetition suppression. *Scientific Reports, 9*, 3190.

Duan, J., Jiang, Y., He, Y., Zhang, F., Shen, M., & Yin, J. (2021). Action generalization across group members: Action efficiency matters. *Cognitive Science, 45*(4), Article e12957.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160.

Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological Science, 25*(3), 656–664.

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of personality and Social Psychology, 38*(6), 889–906.

Gershman, S. J., & Cikara, M. (2023). Structure learning principles of stereotype change. *Psychonomic Bulletin & Review, 30*(4), 1273–1293.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology, 106*(1), 148–168.

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*, Article 104267.

Hamilton, D. L., Chen, J. M., Ko, D. M., Winczewski, L., Banerji, I., & Thurston, J. A. (2015). Sowing the seeds of stereotypes: Spontaneous inferences about groups. *Journal of Personality and Social Psychology, 109*(4), 569–588.

Hamilton, D. L., & Thurston, J. A. (2022). *13 Perceiving Group Attributes Spontaneously: Broadening the Domain* (pp. 228–255). A Social Psychological Approach: The Handbook of Impression Formation.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*(7169), 557–559.

Henrich, J., & Muthukrishna, M. (2021). The origins and psychology of human cooperation. *Annual Review of Psychology, 72*(1), 207–240.

Henrich, N., & Henrich, J. P. (2007). *Why humans cooperate: A cultural and evolutionary explanation*. Oxford University Press.

Hu, Y., Yang, Y., Huang, P., Ai, D., Sun, H., Zhou, D., Huangliang, J., & Yin, J. (2023). More evidence, greater generalization? the relation between the prevalence of observed action and the strength of generalization depends on action properties.

Journal of Experimental Psychology: Human Perception and Performance, 49(3), 306–326.
Huangliang, J., Hu, Y., Zheng, X., Xu, Z., Zhou, W., & Yin, J. (2024). Valence-dependent implicit action generalization among group members. *Cognitive Science, 48*(4), Article e13440.
Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences, 114*(43), E9145–E9152.
Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (in)accuracy in perceptions of groups and individuals. *Current Directions in Psychological Science, 24*(6), 490–497.
Kashima, Y., Woolcock, J., & Kashima, E. S. (2000). Group impressions as dynamic configurations: The tensor product model of group impression formation and change. *Psychological Review, 107*(4), 914–942.
Kim, M. J., Mende-Siedlecki, P., Anzellotti, S., & Young, L. (2021). Theory of mind following the violation of strong and weak prior beliefs. *Cerebral Cortex, 31*(2), 884–898.
King, T. L., Scovelle, A. J., Meehl, A., Milner, A. J., & Priest, N. (2021). Gender stereotypes and biases in early childhood: A systematic review. *Australasian Journal of Early Childhood, 46*(2), 112–125.
Knobe, J., & Cushman, F. (2023). The common effect of value on prioritized memory and category representation. *Trends in Cognitive Sciences, 27*(10), 892–900.
Krueger, J., & Clement, R. W. (1996). Inferring category characteristics from sample characteristics: Inductive reasoning and social projection. *Journal of Experimental Psychology: General, 125*(1), 52–68.
Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review, 103*(2), 284–308.
Lammers, J., Gast, A., Unkelbach, C., & Galinsky, A. D. (2018). Moral character impression formation depends on the valence homogeneity of the context. *Social Psychological and Personality Science, 9*(5), 576–585.
Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology, 93*(2), 234–249.
Liberman, Z., Woodward, A. L., & Kinzler, K. D. (2017). The origins of social categorization. *Trends in Cognitive Sciences, 21*(7), 556–568.
Ludwig, N. N., Hecht, E. E., King, T. Z., Revill, K. P., Moore, M., Fink, S. E., & Robins, D. L. (2020). A novel social attribution paradigm: The Dynamic Interacting Shape Clips (DISC). *Brain and Cognition, 138*, Article 105507.
Martijn, C., Spears, R., Van Der Pligt, J., & Jakobs, E. (1992). Negativity and positivity effects in person perception and inference: Ability versus morality. *European Journal of Social Psychology, 22*(5), 453–463.
McGarty, C., Yzerbyt, V. Y., & Spears, R. (2002). *Stereotypes as explanations: The formation of meaningful beliefs about social groups*. Cambridge University Press.
Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *Journal of Neuroscience, 33*(50), 19406–19415.
Park, B., Fareri, D., Delgado, M., & Young, L. (2021). The role of right temporoparietal junction in processing social prediction error across relationship contexts. *Social Cognitive and Affective Neuroscience, 16*(8), 772–781.
Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*, 195–203.
Powell, L. J., & Spelke, E. S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences, 110*(41), E3965–E3972.
Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology, 61*(3), 380–391.
Randell, J. A. (2021). Evolutionary Social Psychology. *Encyclopedia of Evolutionary Psychological Science*, 2782–2791.
Ranganath, K. A., & Nosek, B. A. (2008). Implicit attitude generalization occurs immediately; Explicit attitude generalization takes time. *Psychological Science, 19*(3), 249–254.
Ratliff, K. A., & Nosek, B. A. (2011). Negativity and Outgroup Biases in Attitude Formation and transfer. *Personality and Social Psychology Bulletin, 37*(12), 1692–1703.
Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review, 86*(1), 61–79.
Reeder, G. D., & Coovert, M. D. (1986). Revising an Impression of Morality. *Social Cognition, 4*(1), 1–17.
Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296–320.
Schroeder, D. A., & Graziano, W. G. (2018). Prosocial behavior as a human essence. In M. van Zomeren, & J. F. Dovidio (Eds.), *The Oxford handbook of the human essence* (pp. 109–121). Oxford University Press.
Shin, Y. S., & Niv, Y. (2021). Biased evaluations emerge from inferring hidden causes. *Nature Human Behaviour, 5*(9), 1180–1189.
Shutts, K., & Kalish, C. W. (2021). Intuitive sociology. *Advances in child development and behavior, 61*, 335–374.
Sidanius, J., & Pratto, F. (2001). *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge University Press.
Siegel, J. Z., Curwell-Parry, O., Pearce, S., Saunders, K. E. A., & Crockett, M. J. (2020). A Computational phenotype of disrupted moral inference in Borderline Personality Disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 5*(12), 1134–1141.
Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour, 2*(10), 750–756.
Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology, 52*(4), 689–699.
Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105*(1), 131–142.
Stangor, C., & Lange, J. E. (1994). Mental representations of social groups: Advances in understanding stereotypes and stereotyping. *Advances in Experimental Social Psychology, 26*, 357–416.
Unkelbach, C., Alves, H., & Koch, A. (2020). Negativity bias, positivity bias, and valence asymmetries: Explaining the differential processing of positive and negative information. *Advances in experimental social psychology, 62*, 115–187.
Unkelbach, C., Fiedler, K., Bayer, M., Stegmüller, M., & Danner, D. (2008). Why positive information is processed faster: The density hypothesis. *Journal of Personality and Social Psychology, 95*(1), 36–49.
Wickham, H., Chang, W., & Wickham, M. H. (2016). Package 'ggplot2'. *create Elegant Data Visualisations using the Grammar of Graphics. Version, 2*(1), 1–189.
Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592–598.
Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin, 24*(12), 1251–1263.
Wojciszke, B., Brycz, H., & Borkenau, P. (1993). Effects of information content and evaluative extremity on positivity and negativity biases. *Journal of Personality and Social Psychology, 64*(3), 327–335.
Ybarra, O., Chan, E., & Park, D. (2001). Young and old adults' concerns about morality and competence. *Motivation and Emotion, 25*, 85–100.
Yoder, K. J., & Decety, J. (2014). Spatiotemporal neural dynamics of moral judgment: A high-density ERP study. *Neuropsychologia, 60*, 39–45.
Yin, J., Duan, J., Huangliang, J., Hu, Y., & Zhang, F. (2022). Members of highly entitative groups are implicitly expected to behave consistently based on their deep-level goals instead of their shallow-level movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(1), 13–28.
Yin, J., Zhou, D., Ai, D., Sun, H., Duan, J., Sun, Z., & Guo, X. (2023). Event-related potential and behavioural evidence of goal-based expectations for consistent actions among group members. *British Journal of Psychology, 114*(3), 662–677.
Yon, D., & Frith, C. D. (2021). Precision and the Bayesian brain. *Current Biology, 31*(17), R1026–R1032.
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.