Journal of Experimental Social Psychology 126 (2026) 104968

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Selective traits and general rewards: Differentiating learning effects on social impression and choice☆

Kira Harris [a], Andrew Luttrell [b], Peter Mende-Siedlecki [c], Leor M. Hackel [a,*]

[a] University of Southern California, 3650 McClintock Ave, Los Angeles, CA 90089, USA
[b] Ball State University, North Quad (NQ), Room 104, Muncie, IN 47306, USA
[c] University of Delaware, 105 The Green, Wolf Hall, Room 108, Newark, DE 19716, USA

ARTICLE INFO

Keywords:
Reward learning
Trait
Social learning
Attitude
Social decision-making

ABSTRACT

People learn about their social world by inferring traits others possess (e.g., "generous"), allowing them to flexibly predict others' behavior in new settings. Yet, people also experience rewarding outcomes in social interactions (e.g., getting a gift). Rewards feel good and may cast a positive glow on others. We tested whether people apply these two types of learning differently to new circumstances. Specifically, we hypothesized that rewards lead to a more general positivity toward others, across judgments and choices, than trait inferences, which exert a more selective influence. In four experiments ($N = 52$, $N = 100$, $N = 91$, $N = 182$), online participants recruited from the university participant pool and Prolific learned about others in a sharing game. These individuals independently varied in reward value to the participant (the absolute amount of money they provided) and their generosity (the proportion of available money they shared). Participants made decisions across games, hypothetical scenarios, and trait ratings that varied in relevance to the sharing game. Participants used trait feedback more selectively than reward feedback: they preferred generous partners more in settings more relevant to generosity, but they preferred previously rewarding partners to a similar degree across scenarios and trait ratings. These results suggest a small, more consistent influence of rewarding outcomes on person representations, compared to trait learning's more selective influence. Beyond reinforcing behaviors, rewards lead people to ascribe positive attributes to and feel drawn toward rewarding individuals more consistently than traits, which convey specific meaning that people apply based on relevance.

1. Introduction

On any given day, we might decide whom to approach at a party, ask for advice on a difficult math problem, or ask for help moving a couch into a new apartment. One way to make these decisions is to consider whose traits best fit the corresponding scenario, relying on specific knowledge about what particular people are like. At a party, we might approach an extrovert we expect will make good conversation without considering their math abilities; for math advice, we might turn to someone we think is smart but ignore their physical strength; and for help moving into an apartment, we might call someone known for their weight-lifting prowess, regardless of how outgoing they are. Yet, beyond these trait inferences, people also learn from past positive outcomes: when people provide us with a compliment, a gift, material resources, or a good time, we experience these outcomes as psychologically rewarding, which can translate to a diffuse liking of the target.

Although both modes of social learning have been established in prior research, we propose that they differ in an important way: how broadly people generalize these impressions to new circumstances. Whereas trait inferences are likely to drive choices that are conceptually relevant to a specific trait, reward-based impressions may create more positive "gut feelings" about a person, drawing us toward social partners even when past rewards are less directly relevant to current choices. For instance, the trait of extroversion is relevant to predicting whether someone will be talkative at a party, but not necessarily to predicting their math ability; nonetheless, if an extrovert pays someone a compliment that is experienced as rewarding, she may have won a small edge for party invites *and* math tutoring gigs alike. Here, we examined whether rewarding interactions lead people to associate others with general positivity, in contrast to a larger but more selective influence of

☆ This paper has been recommended for acceptance by Chadly Stern.

* Corresponding author.
E-mail addresses: kiraharr@usc.edu (K. Harris), alluttrell@bsu.edu (A. Luttrell), pmende@udel.edu (P. Mende-Siedlecki), lhackel@usc.edu (L.M. Hackel).

https://doi.org/10.1016/j.jesp.2026.104968
Received 3 February 2026; Received in revised form 24 June 2026; Accepted 25 June 2026
Available online 29 June 2026

trait inferences in guiding social choices and judgments.

1.1. Learning from traits

Social cognition depends on multiple processes of learning and memory (Amodio, 2019). One such process is semantic memory, reflecting conceptual information we know about someone, such as their traits. People readily draw trait inferences from others' behaviors (Todorov & Uleman, 2002, 2003; Winter & Uleman, 1984).These traits reflect the gist of someone's behavior, capturing the action's abstract meaning beyond its details (Hackel & Mende-Siedlecki, 2023); for instance, sharing french fries and donating to a charity can both reflect a person's generosity. Accordingly, people use trait inferences to predict how others will behave across novel settings for which the abstract characteristic is relevant. If an acquaintance reveals their motto is "never let the party die," we may infer they will be extroverted rather than boring across a night out, a long car ride, or a camping trip.

Importantly, traits form a fine-grained social knowledge structure: people understand how traits relate to one another and when they are relevant to a given setting (Elder et al., 2023; Frolichs et al., 2022; Lin & Thornton, 2023; Mullin et al., 2025; Stolier et al., 2020). For example, we might have a friend who we know to be kind, and consequently, we might infer that this person would be a comforting, empathetic person to turn to for emotional support; however, this impression might not tell us how well our friend would do on a calculus exam (Mullin et al., 2025). Moreover, traits tend to cluster around two dimensions (warmth and competence; Judd et al., 2005; though see Connor et al., 2026; Koch et al., 2016; Lin et al., 2021; Liu et al., 2025; Lu & Lin, 2025), and trait descriptions can impact other trait inferences along one dimension but not the other, showing that people apply these concepts selectively when they are relevant (Zanna & Hamilton, 1972). Accordingly, trait inferences reflect semantic knowledge that can support new judgments specific to contexts we perceive as conceptually relevant.

Supporting this view, past work has found that people form and generalize trait inferences in a way that is sensitive to context (Hackel et al., 2022; Mullin et al., 2025). For instance, after learning about targets' verbal and math abilities, participants preferred verbally (vs mathematically) competent targets for activities perceived as related to verbal ability and vice versa (e.g., advising on a social conflict versus planning a move; Hackel et al., 2022). Altogether, when people form semantic trait inferences about others, they also have a conceptual understanding of how those traits relate to other concepts. People then tend to apply their learning to make decisions accordingly, weighing trait information more heavily when it is more relevant and discounting it when it is irrelevant. Much as people often apply other kinds of information to judgments based on their relevance, or usability (Higgins, 1996), people apply trait knowledge to new contexts based on its social meaning and relevance to the new context.

1.2. Learning from rewarding outcomes

In addition to forming trait impressions of others that draw on semantic knowledge, other memory systems also guide our ability to learn about others (Amodio, 2019). One such system supports instrumental learning from rewards. Rewards are stimuli that evoke psychological processes of liking (i.e., feelings of pleasure), wanting (i.e., motivational drive), and learning (e.g., reinforcing behavior) (Berridge et al., 2009). Although rewards can be non-social, such as receiving money or eating tasty foods, rewards can also reflect social feedback, such as receiving a smile (Tamir & Hughes, 2018), being accepted by others (Babür et al., 2024), or more abstractly, experiencing fairness, good reputation, or a sense of fitting in (Cushman, 2024; Hsu et al., 2008; Izuma & Adolphs, 2013; Sanfey et al., 2003; Zaki et al., 2011).

Learning from rewarding outcomes in social interaction depends on distinct neural and cognitive bases from those that support trait inference. Supporting this idea, Hackel et al. (2015) dissociated learning from traits and rewards in an economic game. Participants played a sharing game with givers who could share a proportion of available money; givers also had different pools available and therefore provided different absolute amounts of money regardless of proportional generosity. Whereas the absolute amount received reflected a reward, the proportion shared revealed the target's trait-level generosity. For example, a target who shared 20 cents out of $1 (20%) would be less generous than a target who shared 20 of 50 cents (40%), even though both yielded the same reward. Participants returned to both generous and rewarding partners, indicating both kinds of feedback reinforced choice.

However, these kinds of learning involved dissociable brain regions. Reward-based feedback was primarily associated with activity in ventral striatum—a core marker of reward-based learning in non-social settings (Daniel & Pollmann, 2014; Pagnoni et al., 2002). In contrast, trait-based feedback recruited not only ventral striatum but also brain regions previously associated with impression updating (e.g., ventrolateral prefrontal cortex; Mende-Siedlecki et al., 2013, b; Mende-Siedlecki & Todorov, 2016). These activations suggest that, even if being treated generously carries affective weight as a social reward (Tamir & Hughes, 2018), generosity feedback also sparked conceptual impression updates above and beyond any affective impact, whereas economic reward feedback did not. Notably, reward feedback had a persistent effect on later choices, such as choosing a partner for a puzzle-solving task, even in the absence of further monetary reward. Altogether, although generosity had a stronger impact on social choice, economic reward had a robust influence that persisted in the absence of further reward opportunities.

Why and how does past reward influence future choices? In contrast to traits associated with specific semantic content, reward feedback may instill a more general sense of positivity toward others. In particular, there are two pathways through which this pattern may emerge. First, reward feedback may lead people to repeat actions in a habitual manner. Although reward learning can rely on causal models of the environment ("model-based learning"), it can also lead people to repeat previously rewarded behaviors without considering why actions led to reward ("model-free learning"; Daw et al., 2011; Doll et al., 2015; Hackel et al., 2019). In the long run, people may repeat actions habitually even when doing so is no longer strategic, and this pattern may extend to social interaction (Hackel et al., 2019; Wood, 2017). Importantly, although habits are cued by features of an environment, they are not sensitive to goals (Wood, 2017). Accordingly, people may feel drawn to the same partners regardless of their current goals or the *conceptual* features of an interaction (e.g., whether they need career advice versus statistical advice, or whether they are playing an economic points-sharing game versus seeking a partner for puzzle-solving).

Second, beyond reinforcing behavior, rewards feel good. People often use subjective feelings to inform their judgments and beliefs, treating their feelings as information (Melnikoff & Strohminger, 2024; Schwarz, 2012). People may attribute rewarding feelings to an individual responsible for those feelings, providing one pathway through which rewards can reinforce our attraction to a stimulus (Clore & Byrne, 1974). Critically, unlike the specific semantic knowledge formed by trait inferences, "liking" can exert a more general influence, providing an affective signal which we may interpret as informative for any judgment. For instance, regardless of whether we are judging a person's likelihood of being a good mentor or a good dance partner, positive feelings from past rewards may inform these judgments to a similar degree. In this manner, reward might not only reinforce behavior but also bias broader beliefs and judgments of rewarding individuals. In other words, people may think a rewarding target is generally "good" with no sense of domain specificity. In turn, people may assume that someone they like is also smarter, kinder, more attractive, and a better baseball player, which could inform inferences that they are *always* the better interaction partner, regardless of context. Once people have updated these beliefs, future decisions might reflect inferences based on

perceiving the target as having more positive traits.

Indeed, amassing evidence indicates that rewards prompt more positive beliefs across a wide range of social and non-social domains. In the non-social domain, people misconstrue gambles as being better in the long-term after receiving short-term rewards, even when sufficient information is available to determine otherwise (Fischer et al., 2017). In social interactions, people rate trivia players as more competent if they provided larger rewards, despite these individuals demonstrating equivalent competence to others (Hackel et al., 2022); think they are better liked by social partners who provided more rewarding acceptance outcomes, despite these individuals providing identical cues to liking as others (Babür et al., 2024; Cho & Hackel, 2022); reciprocate more with wealthy individuals who provided them with larger economic rewards, even when these individuals demonstrate equivalent generosity to their less wealthy counterparts (Hackel & Zaki, 2018); and estimate they performed better in a task if they received more frequent rewards for correct answers, despite having full information about their accuracy (Luo et al., 2025). Moreover, in work dissociating pure reward association (model-free learning) from goal-directed learning using a causal model of the environment (model-based learning), model-free learning has a unique impact on social evaluations (Hackel et al., 2019), suggesting people may sometimes like partners associated with reward without considering *why* those partners led to reward.

Altogether, rather than forming connections to semantic knowledge maps as traits do, rewarding outcomes may directly cause people to approach others and, by eliciting positive feelings, indiscriminately evoke more positive social judgments of them. However, it remains unknown whether reward learning produces this broad positivity across contexts and types of judgments. The prior work cited above has typically examined the effects of reward within the same domain in which reward learning occurred—for instance, examining how reward learning biases competence judgments after a task in which participants learned about trivia performance (Hackel et al., 2022), or examining how reward learning biases self-concept after a task in which participants learned about their own ability (Luo et al., 2025). However, to our knowledge, past work has not examined the generalized effects of reward on contexts unlike the one in which the reward was experienced. We posit that rewards cast a positive light on others that is not limited to the social meaning of the interactions in which the rewards were originally experienced. That is, people might simply want to interact with rewarding others in *any* situation and think of people associated with rewards as being a little bit better in *every* way. If a person provides a reward by sharing money, for example, the recipient may have biased beliefs not only about the giver's generosity but also about their smarts, sense of humor, and ability to perform Shakespeare, and may want to interact with them across contexts simply because they like and feel good about them.

1.3. Overview of present research

We hypothesized that reward would have a general positive effect across social choices and person perception, as compared to trait inferences. That is, we hypothesized that trait inferences would have a relatively more selective influence on choice, such that participants would seek partners seen as generous more in settings in which generosity is relevant (e.g., a trust game) than in settings in which generosity is irrelevant. In contrast, we hypothesized that reward learning should introduce a more general influence on decisions across interaction contexts or judgments of others, regardless of the conceptual relatedness of the new decision or judgment to the original learning context. For instance, after learning that a person provides large rewards in a sharing game, we hypothesized that participants would also seek that individual in an unrelated trivia game or investment game. As a secondary question, we explored whether the effects of reward might depend on changes in other contextual features besides the conceptual relatedness of the social interaction (e.g., making a decision in an environment that looks different).

In each experiment, participants played a sharing game with a group of targets, presented as participants in a prior study who, across trials, varied in how much of their resources they shared with the participant. This game allowed us to independently manipulate how rewarding each target is to a participant (the absolute amount they shared) and the target's generosity toward the participant (the proportion of available resources they shared). We then measured participants' decisions to interact with the targets in a subsequent game and other judgments about the targets.

In Study 1, participants decided whom to interact with in new game contexts and rated their preferences to interact with different targets in hypothetical scenarios. In Study 2, participants made decisions in new game contexts that looked visually similar to or different from the learning context, rated their preferences to interact with the targets in hypothetical scenarios, and made semantic trait judgments about the targets. Study 3 manipulated a new dimension of the decision contexts (specifically, the reward-relevance of the new context) and introduced more varied semantic trait judgments. Finally, in Study 4, participants rated interaction preferences and made predictions about how targets would perform across a wide array of hypothetical real-life scenarios. Altogether, these experiments allowed us to characterize judgment and decision making across a wide array of economic choices, trait judgments, and day-to-day scenarios.

If learning from rewarding outcomes has a more general influence on decision-making than trait inferences, then reward-based decision-making should be less context-sensitive than generosity-based decision-making. The key test of our hypotheses in most studies, therefore, is either 1) the difference in context-sensitivity between generosity-based choice versus reward-based choice (i.e., interaction coefficients between generosity or reward and the relatedness of a new choice to the original learning context), or 2) the difference in variability for judgments rooted in generosity versus reward, reflecting selective judgments versus a flat effect. We report all relevant studies, measures, manipulations, and participant exclusions in the main text or supplemental materials. Participants for all studies provided informed consent in accordance with the University of Southern California Institutional Review Board. Studies were pre-registered through as-predicted (https://researchbox.org/5570?PEER_REVIEW_passcode=QZETGG), and de-identified data and analysis scripts are accessible through OSF (https://osf.io/k95cy/overview?view_only=d9f18f98fc37450cb4173b5960d01e03).

2. Study 1

To examine the generality of reward learning in social choice, Study 1 tested the extent to which participants would rely on trait learning and reward learning in new settings that ranged in their conceptual similarity to the initial learning context. Participants completed a sharing game used in prior work to dissociate learning about generosity and reward (Hackel et al., 2020). Afterward, they made additional choices involving the same individuals in two new games. In one game—the *trust game*—the trait most relevant to a target's value was their trustworthiness, which is conceptually related to generosity (the trait most relevant to the learning game). In the other game—the *trivia game*—the trait most relevant to a target's value was their competence, which is conceptually unrelated to generosity.

We hypothesized that participants would be more likely to choose generous over stingy targets in the trust game than in the trivia game, suggesting specificity in trait learning. In contrast, we expected that participants would choose previously rewarding targets over non-rewarding targets to a similar extent in the related and unrelated games. This finding would suggest a general positivity and attraction toward individuals who have provided rewards in the past, regardless of whether behavior in the new setting has a similar or distinct social meaning to the past scenario in which rewards were encountered.

2.1. Method

2.1.1. Participants

We recruited 70 participants from the University of Southern California participant pool who completed the study online for course credit. This sample was collected in November of 2023. The sample size was pre-registered and chosen based on previous similar tasks using a within-subjects design and multiple learning trials (Hackel et al., 2015; Hackel et al., 2022). After excluding participants who failed the attention check, those who clicked the same button for every response, and those who responded to fewer than 80% of trials during learning or test, we analyzed data from 52 participants ($M_{age} = 20.92$, $SD_{age} = 3.11$; 58% Women, 42% Men). Participants could select multiple racial/ethnic identities. The most common responses were White or Caucasian only (28.8%), East Asian only (25.0%), Hispanic or Latino/a only (21.2%), South Asian only (9.6%), and Black or African American only (5.8%); the remaining participants selected multiple identities (9.6%).

A simulation-based post-hoc sensitivity analysis in R (1,000 simulations, $\alpha = 0.05$), using the fitted mixed-effects logistic model, estimated that a linear contrast of the two targeted two-way interaction terms could detect a difference in log odds greater than or equal to 0.29 (ratio of odds ratios greater than or equal to 1.25) with 82.9% power. While we did not explicitly take measures to ensure the responses were from human respondents, the learning task requires fast paced key presses over multiple trials presented in randomized order, which involve learning from feedback. This design would not have been amenable to simple bots or to humans copying and pasting answers from a large language model (LLM). In addition, it is unlikely that university participant-pool participants would have had access to a complex enough AI agent to complete these tasks for their limited participant-pool requirements during the timeframe in which the study was conducted. We report the timing of data collection for all subsequent studies.

2.1.2. Design

The study used a 2 (target generosity; generous, stingy) x 2 (target reward association; rewarding, non-rewarding) x 2 (context relevance; high relevance, low relevance) within-subjects design. In an initial sharing game, participants learned about four targets who tended to be high or low in their generosity (average proportion of points shared by the target in the game) and association with rewarding outcomes (average number of points shared by the target in the game). Then, participants made decisions about pairs of targets in a trust game (a context with high generosity-relevance) and in a trivia game (a context with low generosity-relevance). The primary dependent variables of interest were participants' binary decisions for which target to select on each trial, as well as participants' ratings for how much they would want to turn to each of the targets in a variety of contexts on a 0–100 scale.

2.1.3. Procedure

2.1.3.1. Learning phase. Participants were informed they would be playing a sharing game with four individuals who had participated in a previous study. The instructions explained that each of these previous participants was represented by an animal avatar on a colored background; these animal avatars were used to make each target easily visually distinguishable and memorable while avoiding attributions of group identity, such as gender and race, to the interaction partners. After reading instructions explaining the game, participants played 72 rounds of a sharing game in which they pressed the “E” or “I” button on their keyboard to make a series of pairwise decisions between two of the targets, represented by their animal avatars. For each participant, the four animal avatars were randomly assigned to each of the four roles defined by generosity (proportion of points the target tended to share) and reward (number of points that target tended to share), such that the same animal was not always generous or rewarding between subjects. The pairwise decisions were such that participants viewed each possible unordered combination of targets 8 times (with each target appearing on the left four times and on the right the other four times). The 72 trials made up of these pairwise decisions were presented in a random order for each participant.

After each decision, participants saw the total point pool that the target they chose was offered and the amount of that point pool that the target chose to share with a future participant. The number of points shared indicated a concrete reward, whereas the proportion of the pool shared revealed the giver's trait generosity. Participants were informed that most people tend to share between 10% and 50% of their pool and that, usually, they could expect to receive 10–50 points on a given round. In reality, the targets were computer-generated in order to manipulate generosity and reward independently. Generous targets tended to share 40% of their pool while stingy targets tended to share 20% of the pool ($SD = 10\%$). Rewarding targets tended to share 40 points while non-rewarding targets tended to share 20 points ($SD = 10$).

On each trial of the sharing game, participants first saw a screen in which two avatars were presented for two seconds and had to make a binary choice. They then saw the outcome of how many points their chosen target shared out of the point pool for two seconds. There was an inter-trial interval of 0.5 s during which a fixation cross was presented.

2.1.3.2. Test phase. After a break, participants read instructions about two more games they would play with the same targets: the trust game and the trivia game. In the trust game, participants decided how to entrust an endowment to the other players. They were told they would receive 100 points for each round, which would be tripled to 300 points and given to whichever target they selected to be the trustee on a given round. That target could choose whether to return 150 points back to the participant, reflecting a prosocial decision leading each to have more points than they started with, or to return nothing, reflecting a self-serving decision leading the target to have 300 points and the participant to have 0. In the trivia game, the participant wagered on which of the two targets had answered a trivia question correctly the fastest on a given round. They were told they would receive 150 points if they chose correctly and 0 points otherwise; the economic value of choosing correctly was therefore matched across games. However, the social meaning of the games differed, such that one was related to prosociality and the other was not. Participants were given comprehension checks after reading the instructions for each game, and they reread the instructions if they answered at least one question incorrectly.

During the test phase, participants played the trust game and the trivia game. On each trial, a screen informed them which game they would play in the next round. Then two of the four animal avatars appeared, and participants selected “E” or “I” to choose one, as they had during the sharing game. In these two games, however, participants did not receive feedback on the outcome of their decision in order to avoid further learning, providing a measure of choice based on prior learning alone; participants were told that they would not learn about their performance until after all their choices were made. Participants played 24 rounds of the trust game and 24 rounds of the trivia game in randomly interleaved order throughout the test phase (see Fig. 1). As in the learning phase, the order of these 48 total trials was random, and each target was presented an equal number of times against each other target and on each side of the screen.

Finally, participants answered hypothetical preference questions about how much they would want to turn to each of the four targets in a series of different scenarios on a scale from 0 (Not at all) to 100 (Very much). These scenarios were selected from a previous pilot study ($N = 33$) where participants rated how relevant generosity would be to one's decision to turn to each target in a variety of situations on a 0–100 scale. We selected two scenarios that were generally rated high in generosity: “You are organizing a fundraiser for your volunteer organization and

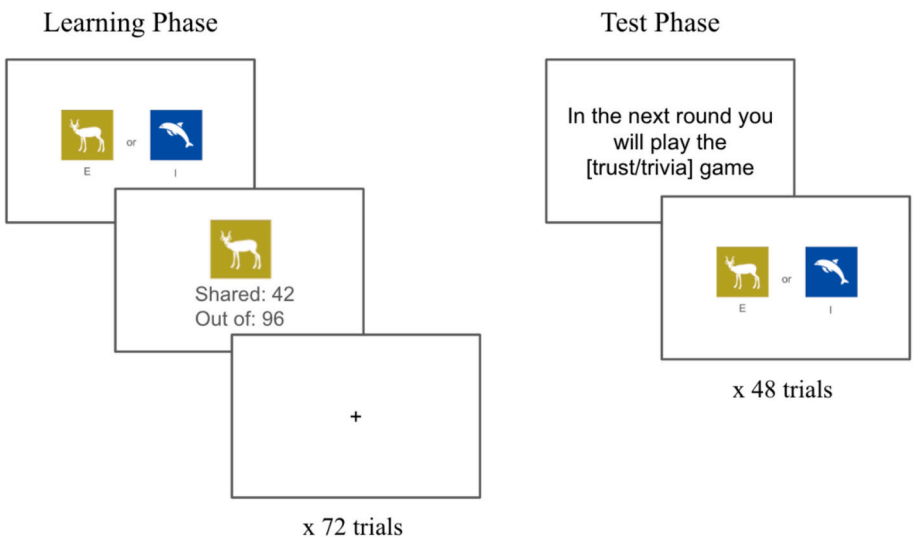

Fig. 1. Figure illustrating the learning and test phases in Study 1. In the learning phase, participants made pairwise decisions between two of the four animal avatars on 72 trials with feedback. In the test phase, participants saw which game (the high generosity-relevance trust game or the low generosity-relevance trivia game) they would play in the next round and made pairwise decisions without feedback for 24 trials in each game. These trials were randomly intermixed.

want donations" ($M = 78.55$, $SD = 22.14$) and "You want a loan to pay rent for the next month" ($M = 72.19$, $SD = 33.24$); two rated in the middle: "You want someone to teach you how to surf" ($M = 56.06$, $SD = 29.58$) and "You want to go on an unplanned road trip over the weekend" ($M = 53.71$, $SD = 28.92$); and two rated low: "You want medical advice" ($M = 45.06$, $SD = 31.33$) and "You want to hire a computer programmer at your company" ($M = 40.90$, $SD = 33.58$). We also included another scenario meant to be unrelated to any personality trait ("You want someone to choose your lottery numbers for you"; $M = 31.16$, $SD = 27.05$). The order in which the scenarios were presented was randomized, and within each scenario category, the target order for ratings of how much participants would want to turn to a particular target in that scenario was randomized. Moreover, participants also rated the extent to which they had positive and negative reactions to each of the targets on a 0 (Not at all) – 100 (Very much) scale, as well as their perceptions of each target's generosity on a 0 (Not at all generous) – 100 (Very generous) sliding scale. Participants reported their positive reactions and then their negative reactions for each target, but the order for which targets they made these ratings about was randomized. (For analyses of ambivalence toward the targets, see Supplemental Materials p. 14.) For the generosity ratings, we also randomized the order in which targets were presented.

2.1.4. Data analysis

2.1.4.1. Choices in novel settings. To examine how past reward and generosity feedback influenced choices in novel games, we conducted a mixed effects logistic regression, assessing how likely participants were to choose targets in each game as a function of prior generosity and reward. In the learning phase, each target was assigned an average generosity value (0.4 or 0.2, the average proportion of points they shared) and reward value (40 or 20, the average total points they shared). For each trial in the test games, we computed the differences between the two generosity values of the two targets shown onscreen and the difference in their reward value, thus generating generosity and reward difference scores. We arbitrarily set whichever target appeared on the right side of the screen as the anchor. Therefore, a positive generosity difference meant that the target on the right had played the learning game more generously than the target on the left. We standardized each type of difference score across test trials for each participant.

The statistical model predicted whether participants chose the target on the right (vs. left) in a given test trial based on 1) the standardized generosity difference, 2) the standardized reward outcome difference, 3) the generosity-relevance of the context (trust game = 1, trivia game = −1), and 4) the interactions between the context and each of the difference scores. We included all random effects terms in the model and iteratively removed the smallest random effect until the model converged (Barr, 2013). The final model included a random intercept term and random slopes corresponding to all fixed effects.

Our goal was to test whether context moderates the application of generosity more than it moderates the application of reward. We therefore conducted a pre-registered linear contrast of coefficients using the doBy package in R, asking whether the Context x Generosity coefficient was significantly greater than the Context x Reward coefficient. This approach can be used to test whether a given coefficient is larger than another (Hackel et al., 2022).

As a non-preregistered exploratory follow-up analysis, we also used a Bayesian approach with the brms package in R (Bürkner, 2017) to model the same predictors and outcome as listed in the frequentist approach (see Supplemental Materials p. 71). The aim in doing so was to assess evidence in favor of a null effect for the interaction of Reward x Context.

2.1.4.2. Preference judgments in hypothetical real-life situations. We conducted a mixed effects linear regression, predicting participants' ratings in each scenario by the target's generosity, the target's reward association, and the generosity relevance of each scenario, calculated by z-scoring pilot ratings. We also included two-way interactions between the generosity-relevance of the scenarios and the generosity/reward of the targets. We included all random effects terms with respect to Participant and Scenario in the model and iteratively removed the smallest effect until the model converged (Barr, 2013). Our preregistration only included random effects with respect to Participant, and we report these pre-registered analyses in the Supplemental Materials (pp. 29–30), but we report the model including random effects for Scenario in order to account for stimulus effects (Judd et al., 2012). The final model reported here included a random intercept of Participant ID and random slopes for target generosity, target reward, and the interaction between target generosity and generosity-relevance with respect to Participant ID. It also included a random intercept of scenario and a random slope for target generosity with respect to scenario. We again used a Bayesian approach to model the same predictors and outcomes as a non-preregistered follow-up (see Supplemental Materials p. 71).

2.2. Results

2.2.1. Choices in novel settings

How did participants use their earlier learning of generosity and reward to make choices in the novel trust and trivia games? To address this question, we regressed whether participants chose the target on the right side of the screen on 1) the standardized generosity difference between the right and left target, 2) the standardized reward outcome difference, 3) the generosity-relevance of the context (trust game = 1, trivia game = −1), and 4) the interactions between the context and each of the difference scores.

Replicating prior work, we observed a significant main effect of generosity ($b = 0.77$, $SE = 0.16$, $OR = 2.15$, 95% CI [1.59, 2.92], $z = 4.93$, $p < .001$) such that participants were significantly more likely to choose the target higher in generosity, and a main effect of reward ($b = 0.25$, $SE = 0.13$, $OR = 1.28$, 95% CI [1.00, 1.64], $z = 1.98$, $p = .048$), such that participants were more likely to choose the previously more rewarding target. More importantly, a Target Generosity x Context Relevance interaction ($b = 0.32$, $SE = 0.11$, $OR = 1.37$, 95% CI [1.11, 1.69], $z = 2.98$, $p = .003$) indicated that participants were more likely to choose the generous target in the trust game, which was highly relevant to generosity, than in the trivia game, which was less relevant to generosity. Participants thus used trait knowledge selectively, drawing on the conceptual relevance of past behavior to new settings. In contrast, the Reward x Context Relevance interaction was not significant or in the same direction ($b = -0.06$, $SE = 0.07$, $OR = 0.94$, 95% CI [0.83, 1.08], $z = -0.89$, $p = .37$), suggesting participants relied on past reward feedback to a similar extent across the Trust Game and Trivia Game.

These two effects were significantly different from one another. Specifically, a linear contrast of coefficients found that the Generosity x Context Relevance interaction term coefficient was significantly greater than the corresponding coefficient for the Reward x Context Relevance interaction ($b = 0.38$, $SE = 0.12$, $t(1) = 9.06$, $p = .003$). Thus, participants showed significantly more context sensitivity for generosity than reward (see Fig. 2).

Although we found a null effect for the Reward x Context Relevance interaction, we conducted a secondary exploratory Bayesian analysis to evaluate whether the data supported a null effect over the presence of an interaction. A Bayes Factor of $BF_{10} = 0.105$ assessing evidence for the presence versus absence of a Reward x Context Relevance interaction indicated moderate evidence in favor of the null hypothesis ($BF_{10} < 1/3$) (Stefan et al., 2019). This analysis provides modest support to the idea that, in addition to being relatively more general than generosity, learning from reward leads to a general positivity that may be independent from the conceptual context.

Finally, although we found participants used generosity knowledge more in the trust game than the trivia game, we examined whether participants nonetheless still choose based on generosity in the trivia game. A follow-up exploratory analysis using dummy coding, with the trivia game as the reference group, demonstrated that participants were more likely to choose the generous target than the stingy target ($b = 0.42$, SE = 0.16, $OR = 1.52$, 95% CI [1.11, 2.10], $z = 2.58$, $p = .01$). Altogether, participants chose generous targets in both games, but did so to a greater extent in the trust game ($b = 0.75$, $SE = 0.21$, $OR = 2.12$, 95% CI [1.41, 3.18], $z = 3.61$, $p < .001$) than they did in the trivia game; in contrast, participants chose rewarding targets without demonstrating differentiation across contexts.

2.2.2. Preference judgments in hypothetical real-life situations

Did this pattern—a general influence of reward but a context-sensitive influence of generosity—also emerge in preferences for hypothetical real-life social settings? To investigate this question, we regressed participants' ratings in each scenario on the target's generosity, the target's reward association, the generosity-relevance of each scenario, and interactions between the generosity-relevance of the scenarios and the generosity/reward of the targets.

Consistent with the game results, we observed a significant main effect of target generosity, which indicated that participants generally preferred to turn to generous targets ($b = 7.18$, $SE = 1.74$, $t(34.46) = 4.13$, $p < .001$). Although participant ratings showed a qualitative

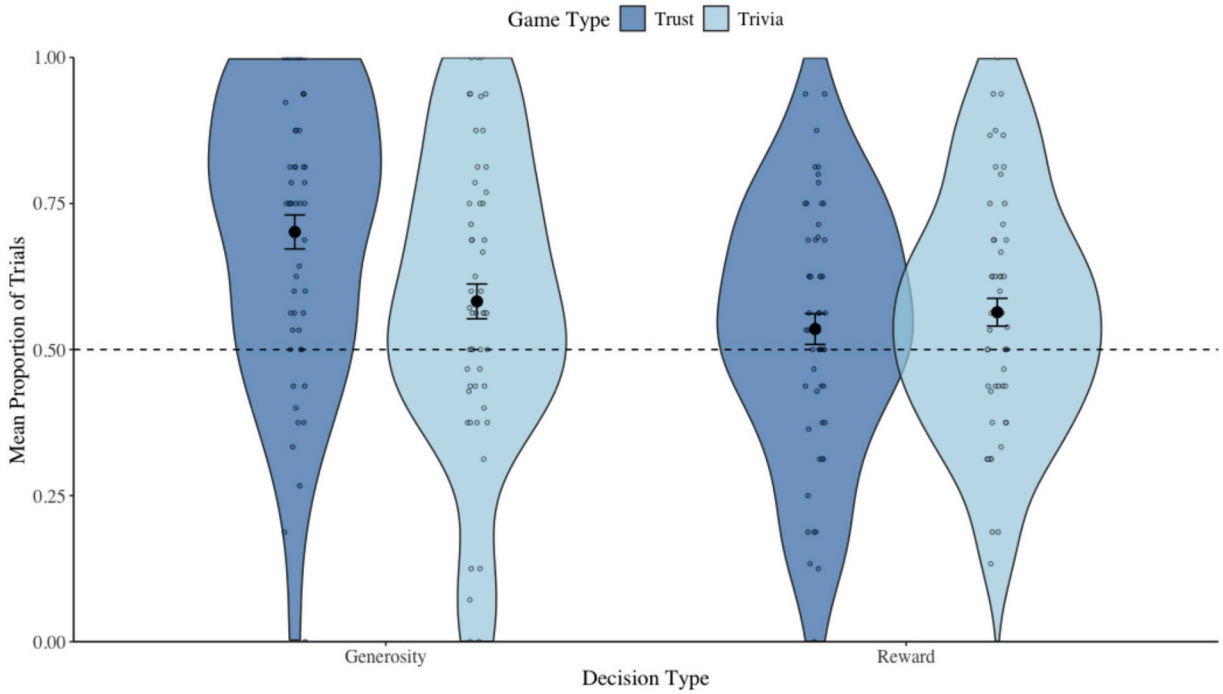

Fig. 2. Generosity-based decisions are more sensitive to context than reward-based decisions. The graph illustrates the proportion of trials in which there was a difference between the two targets' generosity or reward associations on which participants chose the generous or rewarding target. The error bars represent standard errors corrected for the within-subjects design, and the black dotted line is at 0.5, marking where the bars would fall if people were making decisions by chance. Individual data points for the proportion of trials on which an individual participant chose a particular target are represented by the gray dots.

pattern that suggested a weak preference for more rewarding targets, this result was not statistically significant ($b = 1.38$, $SE = 1.07$, $t(51.08) = 1.29$, $p = .204$). We did not observe statistically significant evidence of a Generosity x Generosity-Relevance interaction ($b = 2.23$, $SE = 0.98$, $t(6.04) = 2.28$, $p = .062$), although the data qualitatively were in the direction of participants preferring generous targets more in more generosity-relevant contexts (we test this effect with a larger sample in Study 2). There was no evidence of a Reward x Generosity-Relevance interaction ($b = -0.09$, $SE = 0.53$, $t(1231.15) = -0.17$, $p = .868$), and again, the pattern of this effect was in the opposite direction to the generosity effect. A Bayes Factor of $BF_{10} = 0.02$ indicated there was very strong evidence in favor of the null hypothesis over the presence of an interaction of Reward x Generosity Relevance ($BF_{10} < 1/30$) (Stefan et al., 2019). This analysis provides stronger evidence that reward has a general positive effect across preference, regardless of preference type.

Notably, an exploratory linear contrast (in the spirit of the analysis pre-registered for the test phase) between interaction term coefficients for Generosity x Generosity-Relevance and Reward x Generosity-Relevance demonstrated that they were statistically significantly different ($b = 2.32$, $SE = 1.11$, $t(1) = 4.33$, $p = .037$). Thus, although the Generosity x Generosity-Relevance interaction was not significant, it was significantly more positive than the Reward x Reward-Relevance interaction (see Fig. 3).

This result raises the possibility that the influence of targets' generosity was more context-sensitive than the influence of their reward value when evaluating real-life situations. At the same time, given that neither the Generosity x Generosity-Relevance interaction nor the main effect of reward was significant, this contrast of coefficients should be interpreted with caution. We therefore replicate this preference question procedure in Study 2 using a larger sample size.

2.2.3. Evaluations of targets

How positive were participants' self-reported attitudes toward each of the targets? We conducted an exploratory two-way repeated measures ANOVA predicting the difference score in participants' positive reactions minus negative reactions to each target as a function of target reward and target generosity. Participants had more positive attitudes toward more generous targets than more stingy targets ($F(1,50) = 30.59$, $p < .001$, $\eta^2_G = 0.182$). However, we did not observe a significant effect of reward ($F(1,50) = 0.05$, $p < .82$, $\eta^2_G < 0.001$), nor was there evidence for a Target Generosity x Target Reward interaction ($F(1,50) = 0.41$, $p < .52$, $\eta^2_G = 0.001$). Given the smaller sample size of Study 1, and given past work finding that both Reward and Generosity impact evaluations (Hackel et al., 2020), we returned to this question in Studies 2 and 3 with larger sample sizes.

2.3. Discussion

In Study 1, participants learned about individuals who varied in the generosity they displayed and the rewards they provided. Importantly, these signals were experimentally dissociated; for example, generous individuals could provide small rewards, and stingy individuals could provide large rewards. Later, participants chose generous targets in a game related to generosity to a significantly greater extent than they did in a game unrelated to generosity (though they preferred generous targets to stingy targets in both games). This finding indicates that, although people have an overall preference for interacting with generous individuals (Brambilla et al., 2021), trait knowledge about generosity was deployed selectively, increasing preferences for choosing generous partners in settings that were conceptually relevant to the social meaning of generosity. In contrast, participants were drawn

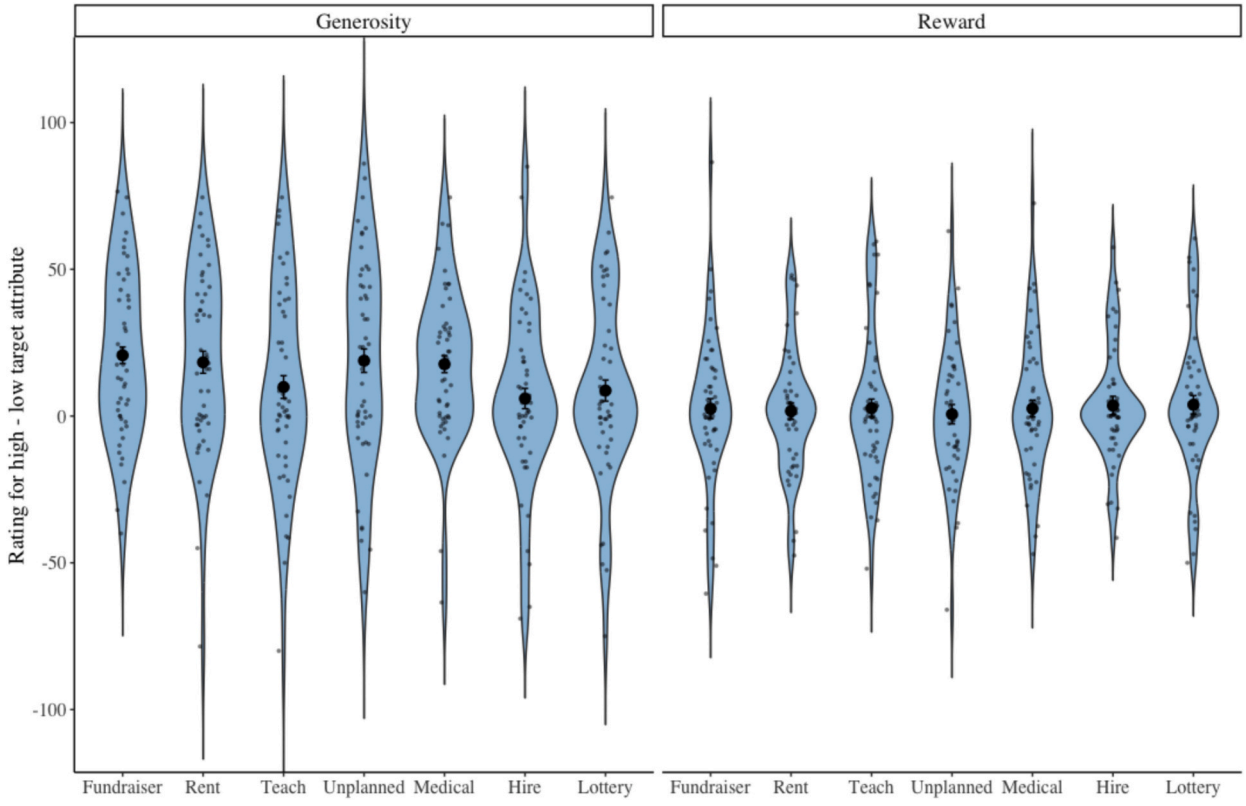

Fig. 3. Generosity-based preferences are more sensitive to the type of preference than are reward-based preferences. The graph illustrates the difference scores between average preference ratings for the generous minus stingy targets and average preference ratings for the rewarding minus non-rewarding targets in Study 1. The error bars represent standard errors corrected for the within-subjects design. Individual data points for a participant's difference in average ratings for targets high in generosity/reward minus targets low in generosity/reward are represented by the gray dots. The contexts are arranged such that those that had higher generosity-relevance scores are on the left.

toward rewarding targets to a similar extent regardless of the setting. These findings suggest that learning from rewarding outcomes is less context-sensitive, potentially reflecting a general positivity that draws people to interact with rewarding interaction partners again, even in conceptually distinct circumstances.

Notably, because rewards varied independently of generosity, point pools in the task were positively correlated with reward values (and negatively correlated with generosity values). This is a typical and necessary feature of this study design to decorrelate generosity and reward (e.g., Hackel et al., 2015). To address the possibility that observed reward effects reflected variation in point pools rather than reward itself, we repeated each analysis while controlling for point pools. Across the four primary studies and one supplemental study, reward remained a significant predictor in 8 of 10 key analyses, whereas point pools themselves did not positively predict responses in any analysis, indicating that the reward effects reported here were generally robust to the inclusion of point pools as a covariate. Further details and discussion are provided in the Supplemental Materials (p. 17).

3. Study 2

Having demonstrated initial evidence that reward learning promotes general but indiscriminate positivity independent of trait learning, we sought to test the robustness and generalizability of this effect. First, Study 2 aimed to replicate the findings of Study 1, in which decisions rooted in past reward, rather than generosity, promoted more general positivity across contexts. Second, we recruited a larger sample size to enhance our power—in particular, to support a more robust test of whether this finding extends to hypothetical real-life scenarios, given that this preference for rewarding targets and the interaction between generosity and context were not statistically significant in Study 1.

Third, as an exploratory question, Study 2 also examined whether the effects of reward would depend on other dimensions of similarity. Study 1 manipulated the *conceptual* similarity of the two scenarios in terms of whether the same trait (generosity) was relevant to each; however, the scenarios were both presented in a way that visually matched the initial learning context. In Study 2, we further tested whether decision-making based on rewarding outcomes might be sensitive to the similarity of *perceptual* features to the original learning context. Although humans tend to generalize based on inferences about latent causes (Gershman et al., 2015), past work suggests that perceptual similarity can guide the generalization of reward feedback, including in social scenarios (FeldmanHall et al., 2018; Kahnt et al., 2012; Verosky & Todorov, 2010). Given that participants applied reward learning regardless of the conceptual relevance of a scenario, we tested an alternative perceptual mechanism, as it is possible that each type of learning may selectively generalize in different ways. Whereas trait learning selectively generalizes to scenarios relevant to the social meaning of behavior (i.e., traits), reward learning might selectively generalize to perceptually similar scenarios. We therefore manipulated the perceptual similarity of the new games to the original sharing game, in addition to manipulating the generosity-relevance of the different contexts. Although this manipulation does not provide a strong basis for generalization–given that perceptual similarity was not diagnostic for decision-making–it tests an alternative pathway observed in prior work that relies less on conceptual similarity. At the same time, we note that this particular test of perceptual relevance should be interpreted with caution, as this was only an exploratory manipulation of some conceptually-irrelevant perceptual elements of the situation. Other manipulations viewed by participants as more relevant to potential latent causes of differences in reward may yield different results.

Finally, we explored whether the effect of reward arose due to people uniquely inferring a different personality trait from the rewarding outcomes they experienced, similar to the way people inferred generosity from the proportion of points targets shared. Specifically, we tested whether rewarding outcomes uniquely predicted ratings of how "lucky" targets were perceived to be.

3.1. Method

3.1.1. Participants

We recruited 120 participants on Prolific who were compensated for their participation, and 118 participants were recorded as having completed the study pre-exclusions, presumably due to a platform error or other error in recording. The sample was collected online in July of 2024. The sample size was pre-registered. This sample size was heuristically increased relative to Study 1 since we were manipulating another variable. Again, we excluded participants who answered the attention check question incorrectly, those who clicked the same button on every trial, and those who responded to fewer than 80% of trials during learning or test. There were 100 participants after these exclusion criteria (M_{age} = 39.76, SD_{age} = 12.93; 43% Women, 57% Men). Participants could select multiple racial/ethnic identities. Responses were White or Caucasian only (64.0%), Black or African American only (17.0%), Hispanic or Latino/a only (7.0%), East Asian only (2.0%), American Indian or Alaskan Native only (1.0%), South Asian only (1.0%), multiple racial/ethnic identities (5.0%), and chose not to disclose (3.0%).

A simulation-based post-hoc sensitivity analysis in R (1,000 simulations, $\alpha = 0.05$), using the fitted mixed-effects logistic model, estimated that a linear contrast of the two targeted two-way interaction terms could detect a difference in log odds greater than or equal to 0.17 (ratio of odds ratios greater than or equal to 1.19) with 85.6% power. Regarding measures to ensure our data represents human responses, Prolific uses internal platform screening that attempts to screen out simple bots and agentic AI. Additionally, as in Study 1, human respondents would not be able to copy and paste responses from an LLM due to the nature of the task, which requires fast-paced key presses in randomized order. We report the timeline for data collection for all other studies and apply the same logic to other data collected on Prolific.

3.1.2. Design

This study used a 2 (target generosity; generous, stingy) x 2 (target reward association; rewarding, non-rewarding) x 2 (conceptual similarity; conceptually similar, conceptually dissimilar) x 2 (perceptual similarity; perceptually similar, perceptually dissimilar) within-subjects design. Target generosity and target reward were manipulated within-subjects by varying the proportion shared by each target (generosity) and the total amount shared by each target (reward) in the initial sharing game, which was the same as in Study 1.

Conceptual similarity was manipulated by presenting the participants with different instructions for four different types of games in the test phase. Two games, the trust game and the casino game (described below), were conceptually similar to the sharing game, as revealed in pilot testing. In contrast, two games, the trivia game and investment game (described below), were conceptually dissimilar to the sharing game. Specifically, we conducted a pilot study with 20 participants to rate a hypothetical high-performing target and a hypothetical low-performing target on each game. The average difference between generosity ratings for a high-performing target versus a low-performing target were greater for the conceptually relevant games (trust game: M = 69.55, SD = 34.14; casino game: M = 48.15, SD = 30.08) as compared to the conceptually irrelevant games (trivia: M = 21.75, SD = 29.85; investment: M = 11.4, SD = 25.80), (b = 21.14, SE = 3.5, $t(19)$ = 6.08, $p < .001$; see Supplemental Materials p. 10 for further details).

Perceptual similarity was manipulated by changing the font type, font color, background color, and associated image cue for each of the trials, such that the visual cues were the same as the initial sharing game for perceptually similar games and different for perceptually dissimilar games (see Fig. 4). Which of the two conceptually similar games was perceptually similar, and which of the two conceptually dissimilar games was perceptually similar, was randomized across participants.

3.1.3. Procedure

3.1.3.1. Learning phase. Participants completed 72 rounds of the sharing game. This game was identical to the game participants played in Study 1 except that the instructions included a visual cue to represent the game, and this cue was also presented after every decision at the bottom of the feedback page (see Fig. 4). The background color of these sharing game trials was also light green rather than the default white background color. These changes were made to allow later manipulations of perceptual similarity.

3.1.3.2. Test phase. After the sharing game, participants read instructions and completed 24 trials of each test phase game in turn.

3.1.3.2.1. High conceptual relevance games. In the casino game (high conceptual relevance), participants were told that the player they chose would decide whether to flip a coin or roll a die to determine the likelihood that the participant or the other player would win points; the participant would win 100 points upon a heads in the coin flip game or a 6 in the dice game. The coin flip was therefore a fair decision, such that either the target or the participant had an even chance of winning, whereas the die roll favored the target over the participant. In the trust game (high conceptual relevance), the participant again sent 100 points to the player they chose. This point value was tripled, and then that player could supposedly decide whether to return 150 points back to the participant.

3.1.3.2.2. Low conceptual relevance games. In the investment game (low conceptual relevance), participants selected another player who chose how many points out of 100 to invest in a stock on a given round; the participant's money is supposedly invested the same way as that Decider's, winning points if the investor invested in a good stock or losing points if the investor invested in a bad stock. In the trivia game (low conceptual relevance), the participant was again told they would get 150 points if they chose the player who correctly answered a trivia question fastest on a given round.

The order of each game block was randomized, but all 24 trials of a given game were presented together in a block following the instructions rather than interleaving the games, in order to enhance the clarity of the task. Included in these instructions was an image of a money bag leading to two outcomes (for perceptually similar games) or a pie chart representation of the outcomes (for perceptually dissimilar games). After each set of instructions, participants had to answer comprehension questions correctly to advance. Then, they played the corresponding game by pressing the "E" or "I" key to indicate which of the two targets presented on the screen they wanted to choose on that trial. Participants did not receive feedback after each trial and, instead, simply saw a fixation cross and then the same visual image cue they saw in the instructions, along with the title of the game.

Next, participants completed the same preference questions as in Study 1 in random order, rating the extent to which they would want to interact with each target in a variety of real-life scenarios, and, again, indicated the extent to which they had positive reactions and negative reactions to each target. Finally, participants indicated how generous and how lucky they thought each target was on 0–100 scales, completed a demographics questionnaire, and were debriefed and thanked for their participation.

3.1.4. Data analysis

3.1.4.1. Choices in novel settings. We again conducted a mixed effects logistic regression to test how participants chose rewarding and generous targets in novel settings. The model predicted the likelihood of choosing the target presented on the right side of the screen based on 1) the standardized difference in generosity between the targets, 2) the standardized difference in reward between the targets, 3) the conceptual similarity of the context (1 = similar, −1 = dissimilar), 4) the perceptual similarity of the context (1 = similar, −1 = dissimilar), and 5) the two- and three-way interactions between the similarity variables and each of the differences in target attributes. We included random intercepts as well as random slopes for each of the fixed effects, including interactions with respect to participant ID. We again conducted a non-preregistered follow-up analysis using a Bayesian model with the same model specification as the frequentist model (see Supplemental Materials p. 71).

3.1.4.2. Preference judgments in hypothetical real-life situations. We next examined social preferences for hypothetical real-life situations, asking whether participants demonstrated more context-sensitivity in their generosity-based preferences than reward-based decisions in everyday scenarios with a greater sample size. We again used mixed effects linear regression to predict participants' ratings as a function of target generosity, target reward association, and the generosity-relevance of each scenario, along with the two-way interactions between generosity-relevance and generosity and between generosity-relevance and reward. We included a random intercept term and random slopes for target generosity, target reward, and the interaction between generosity

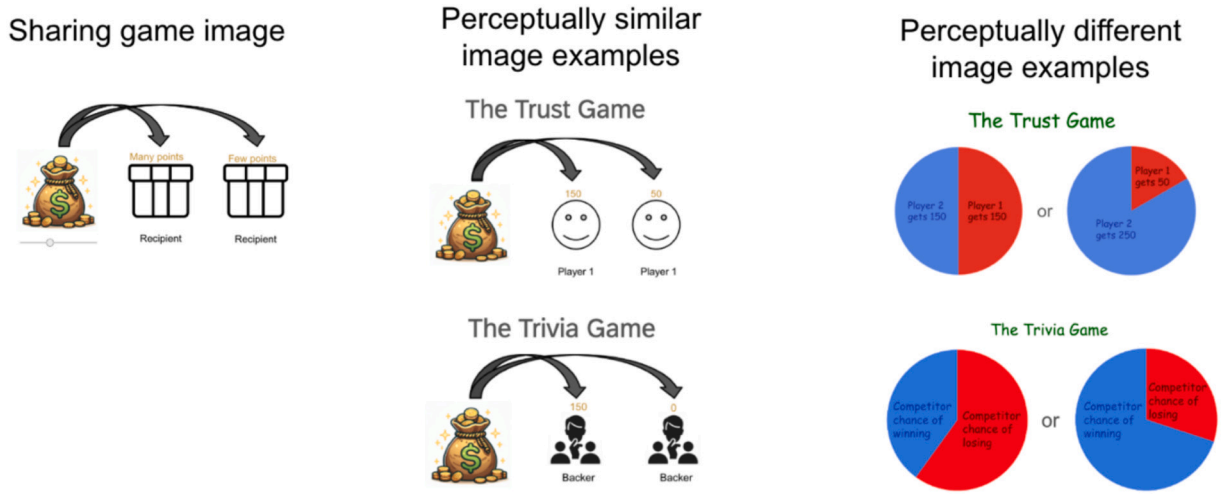

Fig. 4. The image on the left was shown to participants to represent the sharing game. Perceptually similar images included the same structure of a money bag leading to two different potential outcomes, while perceptually dissimilar images were in the form of a red and blue pie chart instead. This figure shows the trust game and the trivia game; however, games were randomly assigned to be perceptually similar or dissimilar, and the other games used a similar format. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and generosity relevance with respect to Participant ID. As in Study 1, we added a random intercept of Scenario and a random slope for target generosity with respect to Scenario to account for stimulus effects (for the pre-registered version of the model without these random effects, see Supplemental Materials p. 39). To quantify generosity-relevance, we again used average ratings of each scenario's generosity relevance from an independent sample of raters who made ratings on a 0–100 scale (see Study 1). Finally, we repeated this analysis in brms in a non-preregistered follow-up Bayesian analysis (see Supplemental Materials p. 71).

3.1.4.3. Semantic trait inferences. In an exploratory analysis, we fit a linear mixed effects regression predicting ratings of generosity (a trait we expected to track generous behavior) and luck (a trait that might uniquely track rewarding behavior) as an outcome measure, including rating type and its interaction with target generosity and target reward as predictors, along with a random intercept for each participant. We again followed up this analysis with a Bayesian analysis using the same model specification (see Supplemental Materials p. 71).

3.2. Results

3.2.1. Choices in novel settings

We first tested how participants made decisions in the test game by regressing whether participants chose the target on the right side of the screen on 1) the difference in targets' generosity 2) the difference in targets' rewardingness, 3) the conceptual similarity of the context, 4) the perceptual similarity of the context, and 5) the interactions between the similarity and target variables. We observed a significant main effect of generosity, which indicated that, during the test phase, participants were more likely to choose targets who had been higher in generosity ($b = 1.13$, $SE = 0.13$, $OR = 3.09$, 95% CI [2.39, 4.00], $z = 8.58$, $p < .001$). Similarly, a significant main effect of reward indicated participants were more likely to choose targets who had provided larger rewards ($b = 0.78$, $SE = 0.11$, $OR = 2.17$, 95% CI [1.74, 2.70], $z = 6.87$, $p < .001$).

As hypothesized, a Conceptual Similarity x Generosity interaction indicated that participants applied their generosity knowledge differently based on how relevant the context was to the concept of generosity ($b = 0.18$, $SE = 0.06$, $OR = 1.20$, 95% CI [1.06, 1.35], $z = 2.97$, $p = .003$). Simple effects analyses conducted as non-preregistered follow-ups demonstrated that when games were conceptually similar, there was a strong effect of generosity ($b = 1.23$, $SE = 0.14$, $OR = 3.42$, 95% CI [2.60, 4.49], $z = 8.83$, $p < .001$), whereas when games were conceptually dissimilar, the effect of generosity was still positive but weaker ($b = 0.88$, $SE = 0.13$, $OR = 2.40$, 95% CI [1.87, 3.09], $z = 6.88$, $p < .001$). Thus, although participants preferred generous targets across both game types, they did so to a greater extent when the games were conceptually relevant to generosity. (See Fig. S1 in supplemental materials for effects by each of the four games, p. 22).

In contrast, the Conceptual Similarity x Reward interaction was not significant ($b = 0.02$, $SE = 0.04$, $OR = 1.02$, 95% CI [0.94, 1.11], $z = 0.47$, $p = .64$), indicating there was no evidence that reliance on reward depended on the social meaning of the game. Importantly, a pre-registered linear contrast found that the Generosity x Conceptual Similarity interaction term was significantly greater than the Reward x Conceptual Similarity interaction term ($b = 0.16$, $SE = 0.07$, $t(1) = 4.89$, $p = .03$). Thus, participants used generosity in a more context-sensitive manner than reward value. Moreover, a Bayes Factor of $BF_{10} = 0.049$ provided strong evidence in favor of the null hypothesis of no Reward x Conceptual Similarity interaction ($BF_{10} < 1/10$) (Stefan et al., 2019), indicating that the influence of reward on decisions did not depend on the social meaning of the game.

As a secondary question, we asked whether reward-based choice would depend on the perceptual similarity of the new game to the original learning game, identifying alternative bases for generalization. The Perceptual Similarity x Reward interaction was not significant ($b = 0.05$, $SE = 0.03$, $OR = 1.05$, 95% CI [0.99, 1.12], $z = 1.49$, $p = .14$), indicating reward reliance did not significantly vary across perceptually similar or dissimilar contexts. Reward also did not show significantly greater moderation by perceptual-relevance than did generosity in a pre-registered linear contrast of coefficients ($b = 0.08$, $SE = 0.05$, $t(1) = 2.11$, $p = .15$). Therefore, we did not observe evidence that reward was sensitive to the perceptual similarity of the context. A Bayes Factor of $BF_{10} = 0.101$ provides moderate evidence in favor of the null hypothesis ($BF_{10} < 1/3$) (Stefan et al., 2019), suggesting that reward was not sensitive to the perceptual similarity of context.

Unexpectedly, a Perceptual Similarity x Conceptual Similarity x Reward interaction suggested participants were *less* likely to choose previously rewarding targets in conceptually similar but perceptually dissimilar games ($b = 0.08$, $SE = 0.04$, $OR = 1.08$, 95% CI [1.00, 1.17], $z = 1.99$, $p = .047$). Simple effects analysis revealed a Perceptual Relevance x Reward interaction in conceptually similar contexts ($b = 0.12$, $SE = 0.04$, $OR = 1.12$, 95% CI [1.04, 1.22], $z = 2.81$, $p = .005$) such that participants preferred previously rewarding targets in both conceptually similar, perceptually similar ($b = 0.85$, $SE = 0.12$, $OR = 2.33$, 95% CI [1.84, 2.95], $z = 7.06$, $p < .001$) and conceptually similar, perceptually dissimilar contexts ($b = 0.62$, $SE = 0.12$, $OR = 1.85$, 95% CI [1.46, 2.35]. $z = 5.13$, $p < .001$) but to a smaller extent. However, there was not a Perceptual Relevance x Reward interaction in conceptually dissimilar contexts ($b = -0.02$, $SE = 0.04$, $OR = 0.98$, 95% CI [0.91, 1.06], $z = -0.58$, $p = .57$). While not hypothesized, this result suggests that reward learning may have been sensitive to perceptual relevance only in conceptually similar contexts (see Fig. 5).

3.2.2. Preference judgments in hypothetical real-life situations

To examine how target generosity and reward contributed to participants' preferences to interact in hypothetical real-life situations, we regressed preference ratings on target generosity, target reward, the generosity-relevance of the scenarios, and the interactions between generosity-relevance and target attributes. As in Study 1, we observed a significant main effect of target generosity, indicating that participants preferred to turn to generous targets overall ($b = 7.87$, $SE = 1.07$, $t(51.88) = 7.36$, $p < .001$). Unlike Study 1, a significant main effect of target reward now similarly indicated that participants prefer rewarding targets overall ($b = 5.93$, $SE = 1.07$, $t(98.99) = 5.53$, $p < .001$).

More importantly, there was a significant Generosity-Relevance x Target Generosity interaction, such that participants' generosity-based ratings depended on the context ($b = 1.76$, $SE = 0.56$, $t(6.24) = 3.16$, $p = .019$). That is, for scenarios highly relevant to generosity (centered one SD above the mean), participants very strongly preferred generous targets over stingy ones ($b = 9.64$, $SE = 1.19$, $t(28.27) = 8.09$, $p < .001$); for scenarios less relevant to generosity (centered one SD below the mean), preferences for generous targets over stingy ones were weaker ($b = 6.11$, $SE = 1.17$, $t(26.86) = 5.20$, $p < .001$). Interestingly, even when Generosity-Relevance was centered at its lowest value, we still observed a significant positive effect of generosity ($b = 5.45$, $SE = 1.27$, $t(18.72) = 4.31$, $p < .001$). Altogether, although participants always preferred generous partners to stingy ones, this preference grew stronger in scenarios more relevant to generosity.

In contrast, the interaction between reward and generosity-relevance was not significant and was in the opposite direction ($b = -0.18$, $SE = 0.36$, $t(2387.99) = -0.51$, $p = .61$), indicating that there was not evidence that participants' preferences for rewarding targets over non-rewarding targets varied by the scenario's generosity-relevance. An exploratory linear contrast between the coefficients for Generosity-Relevance x Target Generosity and Generosity-Relevance x Target Reward provided evidence that the effect of generosity on ratings depended on the generosity-relevance of the context more than did the effect of reward on ratings ($b = 1.95$, $SE = 0.66$, $t(1) = 8.60$, $p = .003$). A Bayes Factor of $BF_{10} = 0.017$ also provides very strong evidence for the null hypothesis ($BF_{10} < 1/30$) (Stefan et al., 2019), providing support for

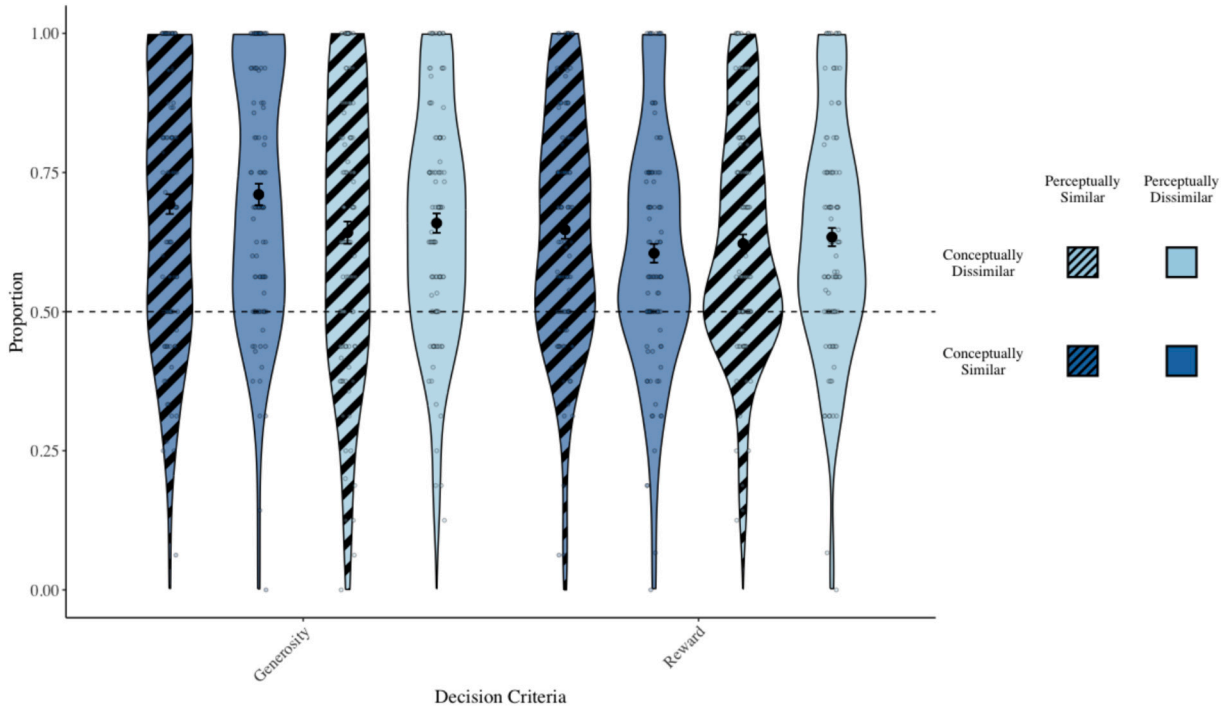

Fig. 5. Generosity-based decisions are sensitive to the generosity-relevance of a context. The graph illustrates the proportion of trials in which there was a difference between the two targets' generosity or reward associations on which participants chose the generous or rewarding target. The error bars represent standard errors corrected for the within-subjects design, and the black dotted line is at 0.5, marking where the bars would fall if people were making decisions by chance. Individual participants' data points are represented as jittered dots.

the idea that there is not evidence of a Generosity-Relevance x Target Reward effect. This analysis provided further evidence that people have a general preference for rewarding individuals, regardless of the conceptual content of a decision (see Fig. 6).

3.2.3. *Semantic trait inferences*

How did generosity and reward feedback impact participants' explicit semantic knowledge about each target? To address this question, we examined ratings of generosity and of luck; in doing so, we aimed to exploratorily test whether participants uniquely inferred

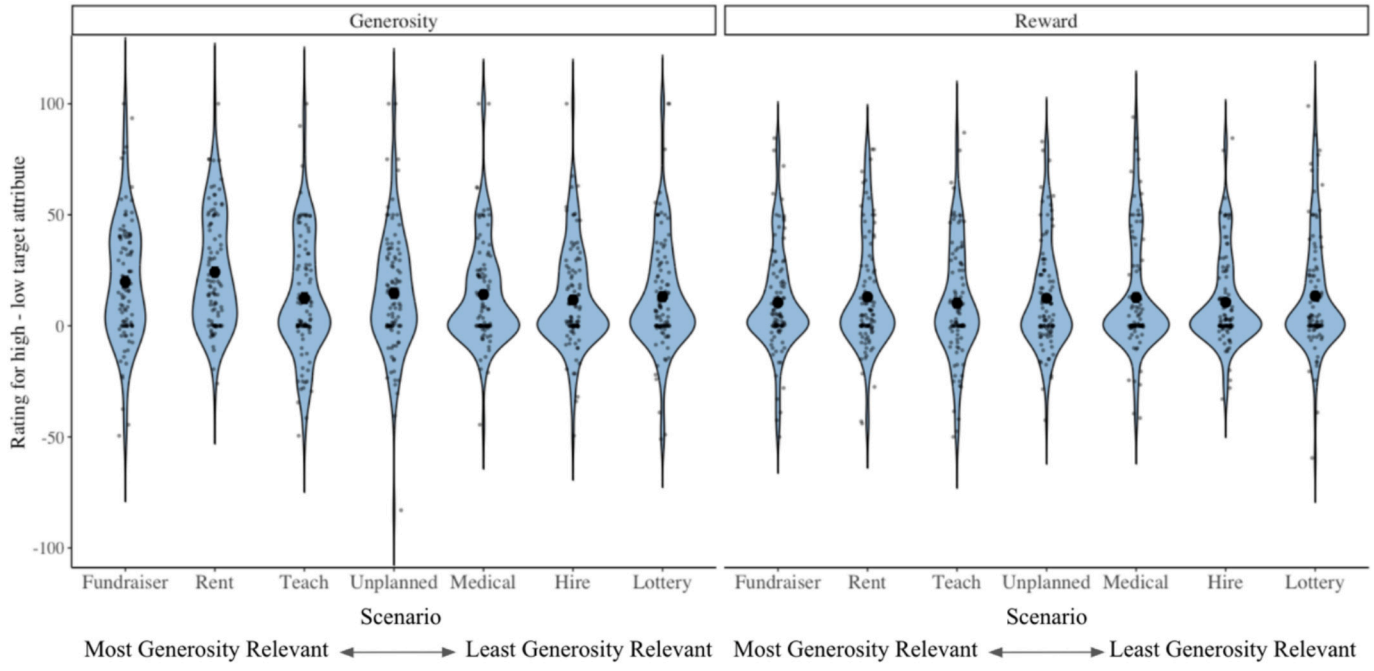

Fig. 6. Generosity-based preferences depend more on the content of the preference question than do reward-based preferences. This graph illustrates the difference scores between average preference ratings for the generous minus stingy targets and average preference ratings for the rewarding minus non-rewarding targets in Study 2. The error bars represent standard errors corrected for the within-subjects design. The contexts are arranged such that those that had higher pilot generosity-relevance scores are on the left of each panel. Individual participants' data points are represented as jittered dots.

rewarding participants were "lucky", therefore suggesting an alternative selective trait inference influenced by reward. Both generosity ($b = 7.85$, $SE = 0.77$, $t(695) = 10.25$, $p < .001$) and reward ($b = 6.01$, $SE = 0.77$, $t(695) = 7.85$, $p < .001$) predicted ratings overall. However, there was only a Generosity × Rating Type interaction ($b = 5.07$, $SE = 0.77$, $t(695) = 6.62$, $p < .001$), indicating that generosity's influence depended on rating type. Specifically, participants rated generous (versus stingy) targets higher when rating their generosity ($b = 12.91$, $SE = 1.08$, $t(695) = 11.93$, $p < .001$) but showed a weaker differentiation of these targets when rating their luck ($b = 2.78$, $SE = 1.08$, $t(695) = 2.57$, $p = .01$). In contrast, we observed no significant Reward x Rating Type interaction ($b = 0.23$, $SE = 0.77$, $t(695) = 0.29$, $p = .77$), indicating that the influence did not differ by rating type. In other words, reward did not uniquely influence ratings of luck; instead, reward inflated estimates of both positive attributes. Moreover, an exploratory linear contrast of coefficients indicated that the Generosity x Rating Type interaction was significantly greater than the Reward x Rating Type interaction ($b = 4.84$, $SE = 1.08$, $t(1) = 20.01$, $p < .001$). A Bayes Factor of $BF_{10} = 0.032$ also provides very strong evidence for the null hypothesis that there is no Reward x Rating Type interaction ($BF_{10} < 1/30$) (Stefan et al., 2019). This result suggests that generosity feedback selectively affected impressions of generosity, whereas reward had a more consistent positive effect on both impressions rather than uniquely influencing impressions of luck. We further test this interpretation in Study 3 by including ratings of additional traits (see Fig. 7).

3.2.4. Evaluations of targets

As in Study 1, we conducted an exploratory analysis of target evaluations and found that more generous targets were evaluated more positively than stingy targets ($F(1,99) = 89.93$, $p < .001$, $\eta^2_G = 0.207$). Unlike Study 1, there was also a significant effect of target reward such that participants evaluated previously-rewarding targets more positively than less rewarding targets ($F(1,99) = 33.69$, $p < .001$, $\eta^2_G = 0.076$). There was still no evidence of an interaction between these target attributes ($F(1,99) = 0.01$, $p = .93$, $\eta^2_G < 0.001$).

3.3. Discussion

In Study 2, we again found evidence that generosity-based learning was more selectively applied than reward-based learning. Whereas participants' preference for generous partners extended to the casino and trust games, but less to the investment or trivia games, their preference for previously rewarding partners extended similarly across all of these games. These patterns further extended to hypothetical preferences for interaction partners in real-world scenarios (e.g., organizing a fundraiser for a volunteer organization versus seeking medical advice). In Study 2, with a larger sample size, the preference for rewarding individuals in these scenarios overall was also statistically significant. Altogether, these findings suggest that reward-based learning was applied regardless of the social meaning of a scenario, whereas generosity-based learning was applied more when conceptually relevant to a new decision than when it was less conceptually relevant.

Moreover, we found initial evidence that trait impressions follow a similar pattern: participants applied learning from generous behavior more to a high relevance trait impression (generosity) than a low relevance trait (luck) but applied learning from rewarding outcomes to a similar extent across traits. We extend this idea in Study 3.

As a secondary question, Study 2 also manipulated perceptual similarity to explore whether learning from reward is sensitive to a different kind of similarity than trait learning is. Reward-based choice did not straightforwardly vary across perceptually similar or dissimilar contexts overall, although there was some evidence that perceptual similarity may interact with conceptual similarity. Given that this interaction was not hypothesized and provides unclear evidence for context dependence of reward, we further explored context dependence of reward learning in Study 3.

4. Study 3

In Study 3, we had three aims. First, we sought to replicate the finding that reward is applied less flexibly than generosity in social decisions and social preferences. Second, we sought to expand on the

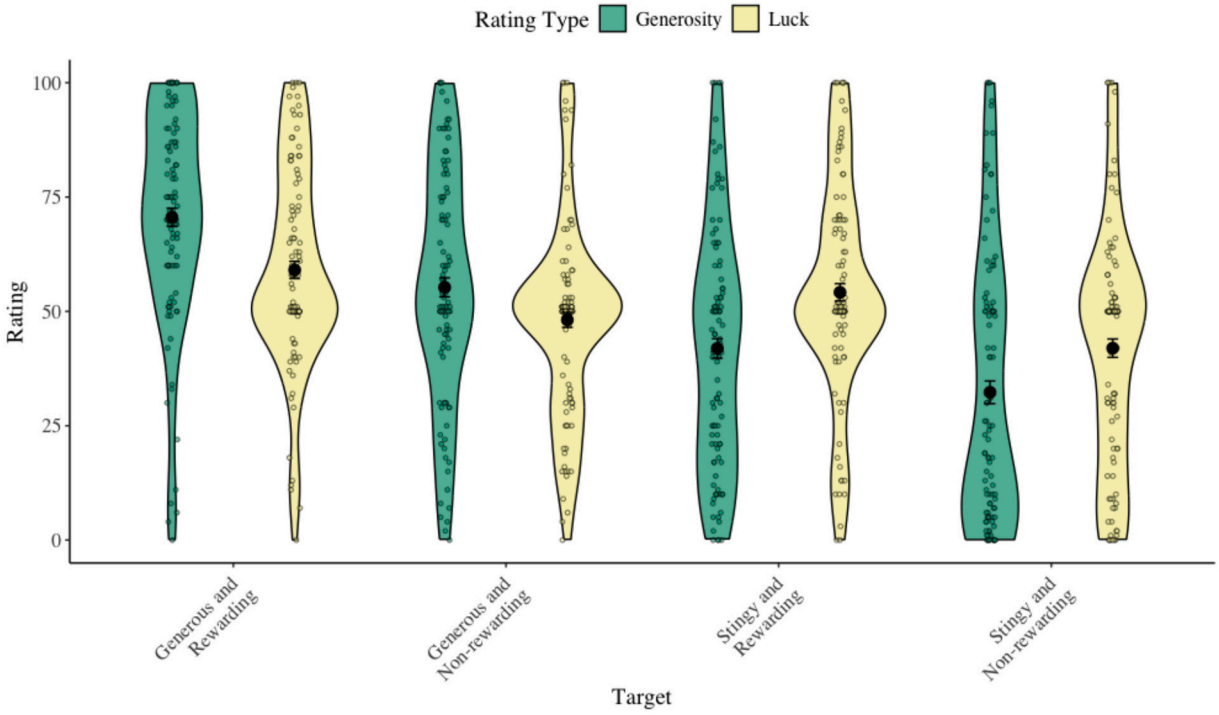

Fig. 7. The graph depicts participants' ratings of target generosity and target luck for each of the four targets. Error bars represent the standard error corrected for the within-subjects design. Individual participants' data points are represented as jittered dots.

trait rating findings in Study 2, which suggested that reward builds general positivity in impressions. Specifically, Study 2 found that rewarding partners were rated more positively than non-rewarding targets to a similar extent across two traits (generosity and luck), whereas generous partners were rated more positively for generosity as opposed to luck. In Study 3, we tested a larger set of traits, asking whether these effects hold across a broader range of judgments.

Finally, as a secondary question, we again considered whether reward learning might be constrained by some alternative aspect of the context besides its social meaning. Studies 1 and 2 did not find evidence that the influence of a target's reward value is sensitive to whether a new decision context resembles either the salient trait priorities (Studies 1 and 2) or visual appearance (Study 2) of the learning environment. However, perhaps the application of reward learning depends on the new context's relevance to reward-based instrumental learning.

To address this possibility, in Study 3, we manipulated whether or not a game scenario was one in which participants could expect to win economic rewards for themselves (high reward-relevance) or merely to predict another person's outcomes (low reward-relevance). This manipulation could impact choice in a few ways.

First, reward learning can build habitual response tendencies (Wood, 2017), but these tendencies might be expressed to a greater extent in subsequent instrumental choices (i.e., selecting who to interact with) as opposed to more distanced predictions. Second, some evidence indicates people use feelings as information to a greater extent when making decisions for themselves as opposed to others (Raghunathan & Pham, 1999; Schwarz, 2012).

Third, this manipulation could provide an explicit basis for considering past reward learning more or less relevant to current decisions. Notably, in Study 3, even in the "reward-relevant" condition, reward outcomes during the learning phase were not actually predictive of rewards in the subsequent games; in the learning phase, some targets had larger pools of money and were more rewarding, whereas in the test phase games, all targets had the same amounts of money available at stake. That is, as in Studies 1 and 2, the actual reward contingencies from learning were non-predictive of outcomes in all Study 3 test games; however, games involving direct choice were *generally* relevant to gaining reward, whereas games involving prediction were not. Therefore, just as some settings were seen as "generosity-relevant," we aimed to test whether making some settings seem generally "reward-relevant" would moderate reward learning.

4.1. Method

4.1.1. Participants

We recruited 120 participants from the University of Southern California participant pool who participated online in exchange for course credit. The sample was collected in September and October of 2024. We pre-registered the number of participants we would recruit to be the same as in Study 2. After excluding participants who responded incorrectly to the attention check question, participants who responded to fewer than 80% of trials in the learning and test phases, and participants who clicked the same key on every trial, we analyzed data from 91 participants (M_{age} = 19.92, SD_{age} = 1.18; 71% Women, 27% Men, 1% Did not report gender). Participants could select multiple racial/ethnic identities. Responses were East Asian only (24.2%), White or Caucasian only (20.9%), Hispanic or Latino/a only (13.2%), South Asian only (11.0%), Other only (5.5%), Black or African American only (4.4%), American Indian or Alaskan Native only (1.1%), Pacific Islander or Native Hawaiian only (1.1%), multiple racial/ethnic identities (14.3%), chose not to disclose (4.4%), and no response (1.1%).

A simulation-based post-hoc sensitivity analysis in R (1,000 simulations, α = 0.05), using the fitted mixed-effects logistic model, estimated that a linear contrast of the two targeted two-way interaction terms could detect a difference in log odds greater than or equal to 0.15 (ratio of odds ratios greater than or equal to 1.16) with 80.4% power.

4.1.2. Design

This study used a 2 (target generosity; generous, stingy) x 2 (target reward association; rewarding, non-rewarding) x 2 (generosity-relevance; high generosity-relevance, low generosity-relevance) x 2 (reward-relevance; high reward-relevance, low reward-relevance) within-subjects design. As in Study 1 and Study 2, target generosity and target reward were manipulated by the proportion and total amount, respectively, that each target shared in the sharing game. Generosity-relevance was manipulated by the game instructions in the test phase. Once again, there were two high generosity-relevance games—the effort game and the casino game—and two low generosity-relevance games—the trivia game and the lottery game.

We confirmed the relevance of each game in a separate group of pilot participants (N = 25), who rated how relevant generosity would be to the other player's behavior in each game on a scale from 0 to 100 ("Not at all relevant" to "Extremely relevant"). The high generosity-relevance games received mean generosity-relevance ratings that were higher than the low generosity-relevance games ($t(24) = 4.28$, $p < .001$; M_{Effort} = 64.64, SD_{Effort} = 30.98, M_{Casino} = 58.84, SD_{Casino} = 33.23, $M_{Lottery}$ = 35.12, $SD_{Lottery}$ = 30.88, M_{Trivia} = 25.44, SD_{Trivia} = 28.75).

After the games, Study 3 participants also rated each target on several trait adjectives. The primary dependent variables of interest were the targets participants chose on each trial in the test phase games and the ratings participants gave on each of the traits.

4.1.3. Procedure

4.1.3.1. Learning phase. As in the previous studies, participants consented to participate in the study after reading a consent form and then read the instructions for the sharing game. This time, the sharing game used the same default white background as the rest of the study, given that there was no manipulation of perceptual relevance in this study.

4.1.3.2. Test phase. After the sharing game, participants read the instructions and completed a test phase without feedback for each of the four games: the effort game, the casino game, the trivia game, and the lottery game. Some of these were the same as those used in Study 2, but others were newly generated. One generosity-relevant game was the effort game (different from Study 2), in which participants were told that the targets had the opportunity to press the space bar 40 times in 10 s on several trials, with the knowledge that doing so would earn points for a future participant. The other generosity-relevant game was the casino game (same as Study 2), in which the target chose to flip a coin or roll a die to determine whether they or a future participant received points on that round. The games that were conceptually unrelated to generosity were the trivia game (same as Study 2), where the targets were competitors trying to correctly answer a trivia question the fastest, and the lottery game (different from Study 2), in which the target was a recipient who randomly received a certain number of points, and a future recipient would be a duplicator who would receive the same number of points. The games were presented in blocks as in Study 2 rather than interleaving trials, and the order in which these game blocks were presented was randomized.

Critically, there were two versions of each game: 1) a high reward-relevance version in which the participant could win or lose points toward a raffle for a $10 bonus that they themselves would receive and 2) a low reward-relevance version in which the participant made predictions about others' outcomes but could not personally gain or lose points.

4.1.3.2.1. The effort game. In the high reward-relevance condition of the effort game, the participant was told that they were the beneficiary of the target's prosocial effort and would, therefore, receive points for choosing a target who had pressed the space bar a sufficient number of times that round. In the low reward-relevance condition, the participant was asked to select the player that they thought had pressed the

space bar 40 times on that round and therefore won points for a third-party beneficiary, but participants were told that they personally would not receive any points from guessing correctly.

4.1.3.2.2. The casino game. In the high reward-relevance condition of the casino game, participants were told that they themselves would have a 50% chance of gaining points (as opposed to the target gaining points) if the target chose to flip a coin and a 16.67% chance of gaining points (as opposed to the target gaining points) if the target chose to roll a die. In the low reward-relevance condition of the casino game, participants were told that a different participant was Player B in this game and that the target was Player A, and participants were asked to identify a Player A who chose to flip the coin (with no opportunities to receive points from their decision).

4.1.3.2.3. The trivia game. In the high reward-relevance condition of the trivia game, participants were told they would receive 10 points each time they correctly chose the target who was the fastest to provide a correct answer in a trivia game. In the low reward-relevance condition of the trivia game, participants were simply asked to predict which target they believed was the fastest to provide a correct answer in a trivia game, with no ability to gain points.

4.1.3.2.4. The lottery game. Finally, in the high reward-relevance condition of the lottery game participants were told that they would receive the same number of points as the target they chose (the target received a random number of points from a lottery), while in the low reward-relevance condition, the participant was simply told that they would win the round if they chose the target who received more points in the lottery (but that they would not receive points themselves).

Which of the two generosity-relevant games was reward-relevant and which of the two low generosity-relevance games was low in reward-relevance was randomized for each participant.

Participants completed all four games by reading the corresponding instructions, answering comprehension questions, and completing 24 trials per game, selecting one target from the pair presented and receiving no feedback after each trial.

4.1.3.2.5. Additional ratings. Next, participants rated their positive and negative reactions to each target and rated their impressions of each target on different positive traits on a scale from 0 to 100. To identify a set of traits to test in a hypothesis-neutral manner, we selected the positive trait words previously used by Oosterhof and Todorov (2008), aiming to obtain a variable sample of trait words without specifically looking for variability along a particular dimension (e.g., warmth or competence). In the Oosterhof and Todorov (2008) study, participants had viewed 66 faces with neutral descriptions and wrote down anything that came to mind about the faces. The researchers categorized these descriptions, and we used the positive traits they identified, along with dominance, which these researchers added due to its importance for person perception. These traits, therefore, span a representative array of the positive traits people spontaneously use to describe others. We also added "generous" and "lucky" since we asked about these traits in Study 2. Altogether, the traits were: attractive, caring, confident, dominant, emotionally stable, generous, lucky, responsible, sociable, and trustworthy. The order of the traits was randomized for each target, as was target order. Next, participants completed a couple of short questions for an unrelated study and a demographics form, after which they were debriefed and thanked for their participation.

4.1.4. Data analysis

4.1.4.1. Choices in novel settings. We again analyzed the game data with a pre-registered mixed effects logistic regression, predicting whether a participant chose the target on the right side of the screen by 1) the standardized difference in the generosity of the two targets, 2) the standardized difference in reward, 3) the generosity-relevance and reward-relevance of the context, and 4) the two- and three-way interactions between each target attribute and the two context relevance predictors. We again included all random effects in the model with respect to Participant ID. We also repeated a non-preregistered Bayesian version of this same analysis (see Supplemental Materials p. 71).

4.1.4.2. Semantic trait inferences. We conducted an exploratory mixed effects linear regression analysis predicting trait rating by target reward value and its interaction with trait generosity relevance, along with target generosity and its interaction with trait generosity relevance, in addition to the corresponding main effects. This analysis was not preregistered; however, we included it because it mirrored the analysis of choices reported above and provided a straightforward test of whether ratings varied with the social meaning of a trait. We included a random intercept for Participant ID, and random slopes for target reward, target generosity, and the interaction between target generosity and generosity relevance with respect to Participant ID. Generosity relevance values were obtained using the ConceptNet/relatedness API (Speer et al., 2017) to obtain similarity values between the word "generous" and each additional trait adjective. Again, we conducted a non-preregistered Bayesian version of this same analysis (see Supplemental Materials p. 71).

Although this analysis examined variance based on the generosity-relevance of traits, it is possible that the effects of reward depend on other dimensions. We therefore next aimed to examine overall variability in trait ratings based on reward and generosity, following the logic of the preregistration but with a slight deviation, using a permutation test instead of bootstrapping. We asked whether the effects of target generosity were simply more variable across traits. For each trait, we computed the difference scores in ratings given to high minus low generosity targets, collapsing across reward, producing a "generosity difference score." Analogously, we computed the difference scores in ratings given to high minus low reward targets, collapsing across generosity, producing a "reward difference score." Next, for each participant, we computed the standard deviation in their generosity difference scores and, separately, the standard deviation in their reward difference scores, across ratings given for different traits. These measures indicate the extent to which a participant's ratings based on generosity and reward varied from one trait to another. A low score would mean that participants gave higher ratings to generous over stingy targets, or rewarding over non-rewarding targets, to a similar degree across traits; a high score would mean that participants gave a higher rating to generous over stingy targets, or rewarding over non-rewarding targets, to different degrees for different traits.

Finally, we asked whether the standard deviation of generosity difference scores was greater than the standard deviation of reward difference scores. In order to test whether this difference was statistically significant, we used a permutation test, which offers a non-parametric approach. This approach creates a simulated null distribution, against which we compared the true value. More specifically, the mean difference in standard deviations was compared to a non-parametric null distribution generated by randomizing whether each participant's standard deviation scores were labeled "generosity" or "reward." This process of shuffling and calculating differences in standard deviations was repeated 10,000 times to simulate a null distribution in which there was no more variability in decisions made from generosity than in decisions made from rewarding outcomes. That is, by randomly reassigning labels on each iteration, this procedure removes any systematic differences between these dimensions and simulates a distribution in which any differences arise through chance alone. If our true observed value is greater than 95% of these simulated values, it suggests that the true value was so large as to be unlikely to emerge from chance alone. Significance was therefore determined by comparing the observed value of the mean difference in standard deviations to the simulated null distribution, with the *p*-value representing the proportion of simulations that had a mean difference equal to or more extreme than the observed value.

4.2. Results

4.2.1. Choices in novel settings

Did participants apply reward learning less selectively than they applied generosity learning? To address this question, we regressed whether participants chose the target on the right side of the screen based on the difference in target generosity and reward and the generosity-relevance and reward-relevance of the context as well as the interactions between the target and context variables.

We observed a significant main effect of target generosity, indicating that participants were more likely to choose the more generous targets overall ($b = 0.40$, $SE = 0.09$, $OR = 1.49$, 95% CI [1.24, 1.79], $z = 4.25$, $p < .001$). In addition, a significant main effect of target reward indicated participants preferred more rewarding targets overall ($b = 0.22$, $SE = 0.08$, $OR = 1.25$, 95% CI [1.07, 1.46], $z = 2.81$, $p = .005$).

More importantly, as in Studies 1–2, we observed a significant Target Generosity x Generosity-Relevance interaction, indicating that the extent to which participants preferred generous targets depended on the generosity-relevance of the context ($b = 0.16$, $SE = 0.06$, $OR = 1.17$, 95% CI [1.05, 1.31], $z = 2.85$, $p = .004$). Simple effects analyses using dummy coding revealed that participants preferred generous partners to a greater degree in games with closer conceptual relevance to generosity ($b = 0.52$, $SE = 0.10$, $OR = 1.68$, 95% CI [1.39, 2.04], $z = 5.35$, $p < .001$), as opposed to games with less conceptual relevance to generosity, in which they had a positive but weaker preference ($b = 0.22$, $SE = 0.10$, $OR = 1.25$, 95% CI [1.03, 1.51], $z = 2.26$, $p = .02$). Replicating studies 1–2, there was not a significant Target Reward x Generosity-Relevance interaction ($b = -0.04$, $SE = 0.03$, $OR = 0.96$, 95% CI [0.91, 1.03], $z = -1.12$, $p = .26$), and the Target Generosity x Generosity-Relevance effect was larger than the Target Reward x Generosity-Relevance effect ($b = 0.20$, $SE = 0.06$, $t(1) = 9.23$, $p = .002$). A Bayes Factor of $BF_{10} = 0.062$ provides strong evidence supporting the null hypothesis that there is no Reward x Generosity-Relevance interaction ($BF_{10} < 1/10$) (Stefan et al., 2019).

However, an analogous Target Reward x Reward-Relevance interaction was also significant ($b = 0.12$, $SE = 0.038$, $OR = 1.12$, 95% CI [1.04, 1.21], $z = 3.04$, $p = .002$); dummy coded simple effects analyses revealed that participants preferred rewarding targets to a greater extent when making instrumental choices that would earn them rewards ($b = 0.32$, $SE = 0.08$, $OR = 1.38$, 95% CI [1.18, 1.61], $z = 3.99$, $p < .001$), as opposed to when predicting the outcome of another person ($b = 0.10$, $SE = 0.08$, $OR = 1.11$, 95% CI [0.94, 1.30], $z = 1.22$, $p = .22$).

Additionally, the pre-registered linear contrast between the Generosity x Generosity-Relevance and Reward x Reward-Relevance interaction terms was not significant ($b = 0.04$, $SE = 0.07$, $t(1) = 0.44$, $p = .51$). Thus, this study provides evidence that the application of reward learning can depend on reward-relevance of a new context: participants applied prior reward learning to a greater extent when they could actually receive rewards, and less so when they could not (see Fig. 8).

4.2.2. Semantic trait inferences

To understand how generosity and reward learning influenced trait impressions, we regressed trait ratings on target generosity, target reward, the generosity-relevance of the traits, and the interactions between the target attributes and generosity-relevance. This analysis revealed a significant main effect of target reward ($b = 2.00$, $SE = 0.87$, $t(89.87) = 2.30$, $p = .024$), indicating that participants tended to give higher positive trait ratings to rewarding targets. Similarly, a main effect of target generosity demonstrated that targets tended to rate generous targets higher on positive traits as well ($b = 4.06$, $SE = 0.82$, $t(90.08) = 4.96$, $p < .001$). As we would expect based on our previous findings, the more a trait was conceptually related to generosity, the more ratings on that trait were affected by the target's generosity ($b = 1.96$, $SE = 0.39$, $t(90.11) = 5.08$, $p < .001$). Although participants rated generous targets more positively than stingy ones even when centering the data at the least generosity-relevant trait ($b = 2.56$, $SE = 0.82$, $t(100.76) = 3.12$, $p = .002$), this difference was greater for traits relevant to generosity.

In contrast, the effect of reward on ratings did not significantly

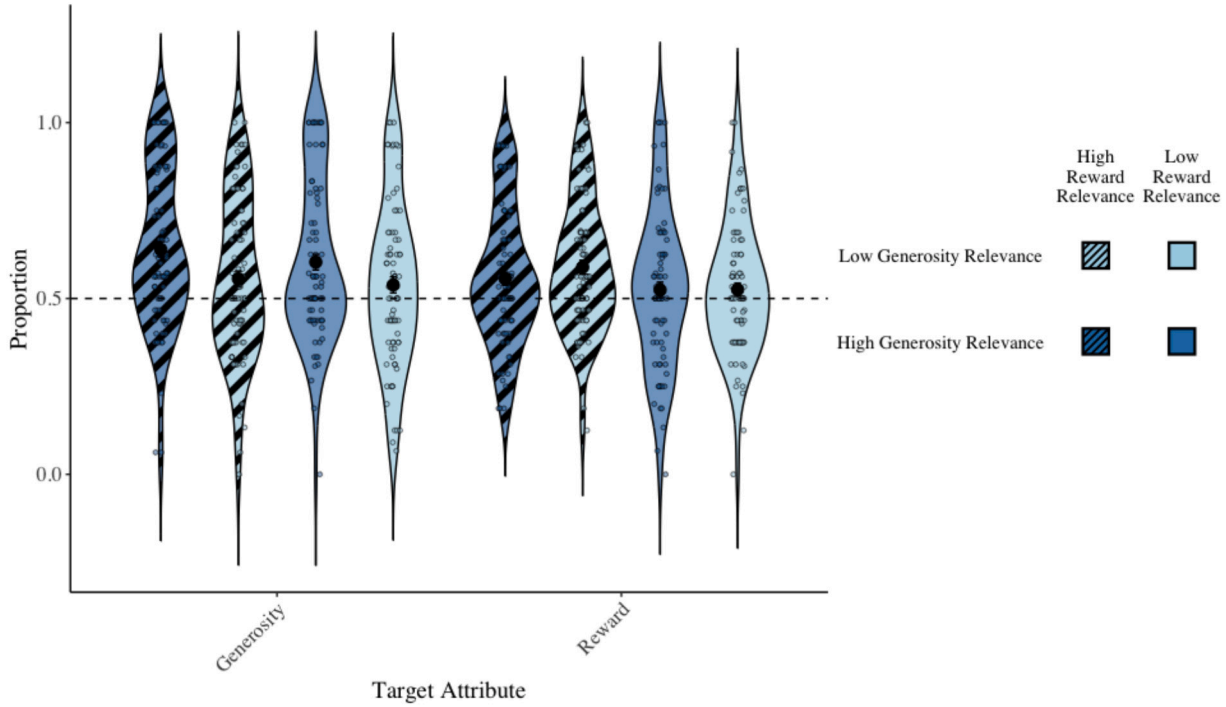

Fig. 8. Generosity-based decisions are sensitive to the generosity relevance of the decision context, and reward-based decisions are sensitive to whether participants had the opportunity to win an economic reward from their decision. The graph illustrates the proportion of trials in which there was a difference between the two targets' generosity or reward associations on which participants chose the generous or rewarding target. The error bars represent standard errors corrected for the within-subjects design, and the black dotted line is at 0.5, marking where the bars would fall if people were making decisions by chance. Individual data points representing the proportion of trials on which an individual participant chose a particular target are represented by the jittered gray dots.

depend on the generosity-relevance of the trait being rated ($b = -0.41$, $SE = 0.34$, $t(3268.38) = -1.20$, $p = .23$), and the interaction coefficient for Target Generosity x Generosity Relevance was significantly greater than the coefficient for Target Reward x Generosity Relevance ($b = 2.37$, $SE = 0.52$, $t(1) = 21.09$, $p < .001$). A Bayes Factor of $BF_{10} = 0.028$ provided very strong evidence in favor of the null hypothesis that there is not a Target Reward x Generosity-Relevance interaction. Thus, reward learning inflated ratings of positive traits, regardless of the extent to which a new trait related to generosity (see Fig. 9).

This analysis found that the influence of a target's reward value did not reliably vary by a trait's generosity-relevance; however, there could still be some meaningful dimension of traits that matters for whether reward biases evaluations on them. To test this possibility more broadly, we conducted a pre-registered analysis asking how much reward's effects varied *at all* across traits. If generosity is selectively applied, different trait dimensions (caring, dominant, etc.) should vary from one another in the extent to which participants prefer generous targets over stingy ones. In contrast, if reward produces a blanket sense of positivity in impressions of others, then people may rate high-reward targets more positively than low-reward targets across traits in a more uniform manner. For instance, a target who had been rewarding in the sharing game may be rated as higher not only in kindness but in attractiveness and confidence as well. On the other hand, if there is some hidden feature that matters for reward's influence, then we would expect to observe variance in its effect across traits to a similar degree as generosity.

To test this possibility, we conducted a permutation test to assess whether generosity or reward led to more variable differences in ratings. This analysis revealed that the effect of target generosity across trait ratings was more variable ($M_{SD\_Generosity} = 20.21$) than the effect of reward ($M_{SD\_Reward} = 16.30$). Specifically, the observed mean difference in generosity variability versus reward-based variability was 3.90, which was significantly greater than expected in the null distribution ($M = 0.03$, $SD = 1.37$; $p = .004$).

Moreover, the variability in the effect of reward was no greater than the variability in ratings that would be expected by chance; even in the absence of systematic variation in how much reward matters for a trait, there can be variability in ratings due to noise. To formalize this comparison, a null distribution was generated exploratorily by randomly reassigning the labels of all four targets, thus shuffling which targets were labeled high- or low-reward, re-computing reward difference scores, and re-calculating the average standard deviation of reward 10,000 times with this statistic. This approach reflects variability in ratings that cannot be related to any target characteristics in the task. The actual observed standard deviation of reward scores ($M_{SD\_Reward} = 16.30$) was not significantly greater than the mean of this null distribution and was actually directionally lower than the variability expected in the null distribution ($M_{\text{null}} = 17.08$, $SD_{\text{null}} = 0.68$; 95% *CI* [15.78, 18.46]; $p = .26$). In contrast, variability in generosity scores ($M_{SD\_Generosity} = 20.21$) was greater than that expected by the null distribution ($p < .001$).

Altogether, this finding again suggests that past generosity impacted ratings to a different extent for different traits, with generous targets being rated more highly for some traits than for others; in contrast, past reward boosted trait ratings overall, and we did not find evidence that traits meaningfully varied from one another in the size of this boost. These findings are consistent with the hypothesis that the effect of generosity systematically varies by the traits' conceptual relevance to generosity, whereas reward produces a general positive boost in impressions across traits.

4.2.3. Evaluations of targets

We conducted the same two-way repeated measures ANOVA as in studies 1 and 2 to assess effects on self-reported evaluations of the targets. The results replicated the findings from Study 2. Participants evaluated generous targets more positively than stingy targets ($F(1,88) = 46.37$, $p < .001$, $\eta^2_G = 0.136$), and participants also evaluated previously-rewarding targets more positively than less rewarding targets ($F(1,88) = 6.42$, $p = .013$, $\eta^2_G = 0.020$). There was still no evidence of an interaction between these factors ($F(1,88) = 0.65$, $p = .421$, $\eta^2_G = 0.002$).

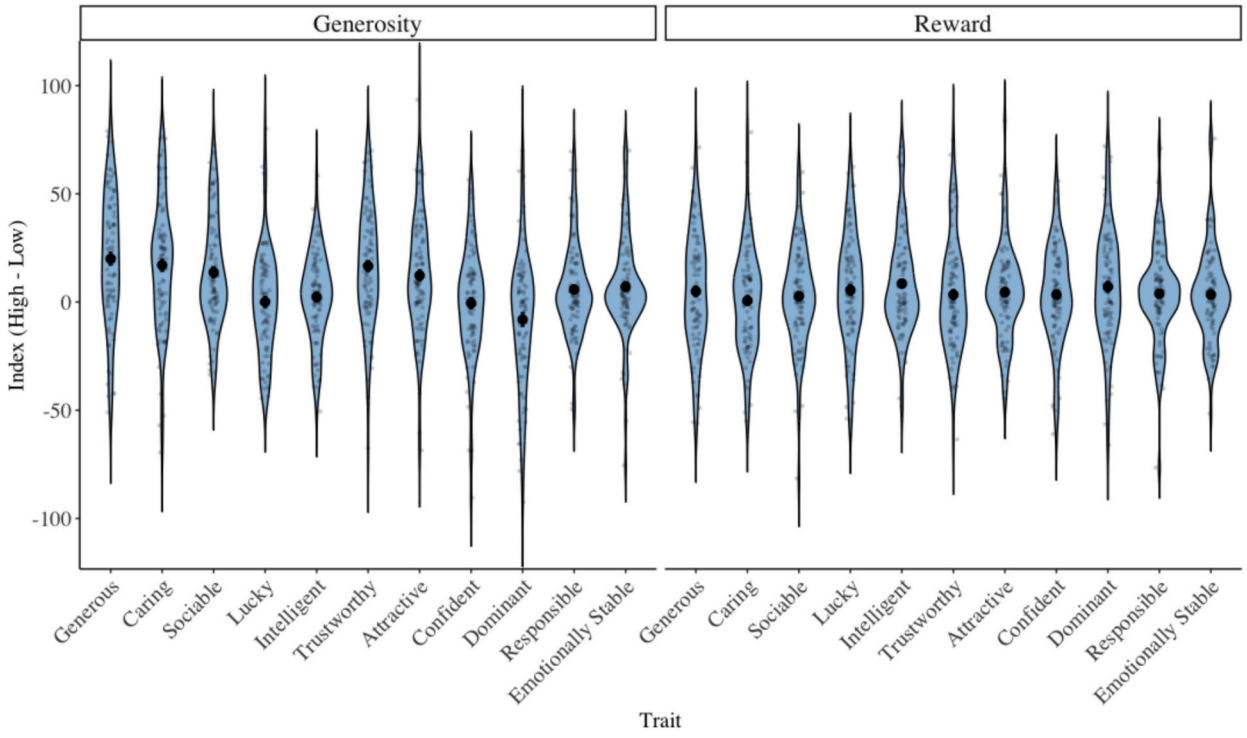

Fig. 9. Generosity-based trait ratings are more variable than reward-based trait ratings. The graph illustrates the difference scores between average trait ratings for the generous minus stingy targets and average trait ratings for the rewarding minus non-rewarding targets in Study 3. The error bars represent standard errors corrected for the within-subjects design. Individual participants' data points are represented as jittered dots.

4.3. Discussion

Study 3 expanded our understanding of how rewards bias choice and impressions, while also identifying distinct factors that can amplify trait-based versus reward-based learning. First, as in prior studies, participants again preferred generous partners to a greater extent when their behavior had a similar conceptual meaning to the sharing game (e.g., games involving prosociality) as opposed to dissimilar conceptual meaning (e.g., games involving luck). In contrast, participants preferred rewarding partners to a similar degree across these games.

A similar pattern was observed for trait ratings. When asked to rate these targets across a variety of traits, participants demonstrated more variability in their application of generosity knowledge (e.g., differentiating more between generous and stingy targets when rating trustworthiness as opposed to luck) than in their application of reward-based learning (e.g., rating a high-reward target more positively than a stingy target to a similar degree across trustworthiness and luck). This finding suggests that rewards generate a consistent boost across trait impressions, supporting our hypothesis that reward leads to a general but indiscriminate sense of positivity about others. Indeed, the reward effect showed no more variability across traits than would be expected by chance, consistent with the idea that conceptually different traits were impacted to a relatively similar degree by reward learning. People may thus draw on general positive affect from reward learning as information that a target is more likely to have positive traits. Moreover, this bias in trait ratings correlated with reward-based choice in the games; participants who were more biased in their trait ratings were also more sensitive to reward in their game decisions (see Supplemental Materials p. 15). Thus, these inflated trait inferences might contribute to people's increased propensity to interact with rewarding targets.

Study 3 also identified a factor that influenced the expression of reward learning in the games: participants applied prior reward learning more when they made instrumental choices that could win personal rewards as compared to when they were predicting the outcomes others would receive. Notably, this was true even though there was no strategic reason to choose previously rewarding targets in any of the games; in the novel games, there were no point pools that varied like they did in the learning phase, making prior reward learning irrelevant. Nonetheless, participants particularly preferred previously rewarding partners in games that *generally* involved instrumental choice and economic reward.

Why did this pattern emerge? One possibility is that these games were seen as more relevant to economic reward. However, responses in these games were also more self-relevant, involving choice for oneself as opposed to responses about others—a variable that may also impact construal level or psychological distance. Finally, these responses also involved instrumental action as opposed to prediction. This pattern may therefore reflect a habitual effect of reward that is sensitive to a general motivational drive to obtain rewards, and which may be activated under sufficiently motivating conditions (Niv et al., 2006). That is, habits may be expressed to a greater extent when making a motivated and immediate instrumental response, rather than a more distanced consideration. In Study 4, we separately manipulated economic-relevance and self-relevance, allowing us to test whether either factor moderates reward learning. In addition, we manipulated these factors while participants rated preferences for real-world scenarios, rather than immediate instrumental choice, to better understand when and why reward learning would show context sensitivity.

5. Study 4

Having demonstrated the widespread general positivity elicited by reward in most domains, as well as its sensitivity to some variation in context, we asked how these effects might generalize outside of an immediate instrumental decision-making context. In Study 4, we therefore addressed two aims. First, we aimed to replicate the effects of generosity and reward across a wider set of hypothetical real-life scenarios. While Study 3 identified this effect across an array of trait impression ratings, we aimed to detect whether reward learning would promote general social preferences across a wider range of settings.

Second, we manipulated the economic-relevance and self-relevance of the choices in order to better inform when and how generosity and reward learning each may lead to selective versus blanket preferences for others. These two explanations were confounded in Study 3 and both provide potential explanations for how reward processes might impact deliberations about the relevance of the semantic meaning of reward in new contexts. Study 3 also used contexts divorced from the instrumental decision process. We, therefore, wanted to distinguish whether either of these semantic-meaning-based explanations could explain the reward context-sensitivity discovered in Study 3 and, if so, which was driving the effect.

5.1. Method

5.1.1. Participants

We recruited 200 participants, 50 from the University of Southern California participant pool and 150 from Prolific, to participate in the study. This sample was collected online in December of 2024. The sample size was pre-registered. The university participant-pool participants received course credit for their participation, and the participants on Prolific received payment. After implementing pre-registered exclusion criteria (answering the attention check incorrectly, not answering at least 80% of learning phase trials, pressing the same button on every learning phase trial), 182 participants remained for inclusion in data analyses (M_{age} = 36.05, SD_{age} = 14.31; 57% Women, 42% Men, 2% Nonbinary). Participants could select multiple racial/ethnic identities. Responses were White or Caucasian only (52.75%), Black or African American only (13.74%), East Asian only (12.64%), Hispanic or Latino/a only (6.59%), Other only (0.55%), South Asian only (0.55%), multiple racial/ethnic identities (12.64%), and chose not to disclose (0.55%).

A simulation-based post-hoc sensitivity analysis in R (1,000 simulations, $\alpha = 0.05$), using the fitted linear mixed-effects model, estimated that a linear contrast of the two targeted two-way interaction terms could detect a difference in beta coefficients greater than or equal to 0.74 with 80% power.

5.1.2. Design

This study used a 2 (target generosity; generous, stingy) x 2 (target reward association; rewarding, non-rewarding) x 2 (generosity-relevance; high generosity-relevance, low generosity-relevance) x 2 (economic-relevance; high economic-relevance, low economic-relevance) x 2 (self-relevance; high self-relevance, low self-relevance) within-subjects design. Target generosity and reward association were manipulated through the sharing game, as in the previous three studies.

Generosity-relevance and economic-relevance of the contexts were manipulated by presenting different scenarios that varied independently in their relevance to these dimensions, based on pilot ratings (from the same 25 participants who rated the effort, casino, lottery and trivia games) in which participants rated how much each behavior reflected "generosity" and how much each behavior reflected "gaining money" on a 0 (Not at all) to 100 (Very much) scale. High generosity-relevance scenarios were each given a rating above 50 on generosity-relevance by these participants, and high economic-relevance scenarios were each given a rating above 50 on how much they reflect gaining money. Low generosity-relevance and low economic-relevance scenarios were rated below 50 on each score, respectively.

We selected four scenarios from each category (high generosity-relevance, high economic-relevance ($M_{Generosity}$ = 74.61, $SD_{Generosity}$ = 22.74; $M_{EconomicReward}$ = 72.67, $SD_{EconomicReward}$ = 28.00), e.g., "You want to ask someone to give you $100 because you have been saving money to buy a new car"; high generosity-relevance, low economic-relevance ($M_{Generosity}$ = 76.15, $SD_{Generosity}$ = 26.50; $M_{EconomicReward}$ = 29.08, $SD_{EconomicReward}$ = 34.34), e.g., "You want to ask someone to

carry your backpack while you are hiking when it is getting too heavy for you to carry"; low generosity-relevance, high economic-relevance ($M_{Generosity}$ = 26.08, $SD_{Generosity}$ = 29.96; $M_{EconomicReward}$ = 74.55, $SD_{Generosity}$ = 27.54), e.g. "You want to bet on a horse jockey that wins the race so that you win a lot of money"; low generosity-relevance, low economic-relevance ($M_{Generosity}$ = 21.04, $SD_{Generosity}$ = 27.37; $M_{EconomicReward}$ = 20.31, $SD_{EconomicReward}$ = 30.02), e.g., "You want someone to play well on your dodge ball team").

To manipulate self-relevance, we randomly assigned some scenarios to be framed as a choice participants had to make for themselves and others to be framed as uninvolved evaluations. We developed a high- and low-self-relevant version of each of the 16 scenarios, resulting in 32 total scenarios in this study. The high self-relevance versions asked participants to rate how likely they would be to choose a particular target in that context. For instance, the example scenarios above use the high self-relevance framing. The low self-relevance versions asked participants to rate how likely they thought the target would be to do "well" in that context. Which scenarios were assigned to the high or low self-relevance conditions was randomized for each participant, as explained in the procedure.

5.1.3. Procedure

After consenting to participate in the study, participants read the instructions for and completed the sharing game in the same way as in previous studies (the format was identical to that of the sharing game in Study 3). Next, participants rated each target in eight self-relevant scenarios. For each participant, these scenarios were the self-relevant versions of eight scenarios—two drawn randomly from each category in the 2 (generosity-relevance) x 2 (economic-relevance) set of scenarios. Participants rated all self-relevant scenarios first so that participants who saw the low self-relevance scenarios first would not apply the same reasoning to the high self-relevance situations, which we thought was a more likely possibility than the reverse effect (i.e., people interpreting low self-relevance prediction questions about behavior as asking about their own desire to interact with a target). The ratings were in response to the instruction: "How likely would you be to choose this person if…" followed by the scenarios, such as "You want to ask someone to give you $100 because you have been saving money to buy a new car." The order of the target ratings and the order of the scenarios within each target were randomized.

Then, participants responded to the low self-relevance versions of the remaining eight scenarios for each target. These ratings were in response to the instruction "How likely do you think this person would be to…" The equivalent low self-relevance example scenario to the high self-relevance example given would be "give someone $100 because they have been saving money to buy a new car." Only one version of these two examples would be presented to a single participant. Again, the order of the targets and scenarios about each target were both randomized. Ultimately, each participant contributed 64 responses: eight ratings for high self-relevance scenarios and eight ratings for low self-relevance scenarios for each of the four targets they learned about in the generosity game.

Next, participants answered a free response question about their impression of each participant for future exploratory analysis. Finally, participants completed a demographics questionnaire and were debriefed and thanked for their participation.

5.1.4. Data analysis

5.1.4.1. Preference judgments in hypothetical real-life situations. We conducted a mixed effects linear regression analysis predicting participants' ratings by the target's generosity, generosity-relevance of the scenario, and its self-relevance, and the two- and three-way interactions between these variables, as well as by the target's reward, economic-relevance of the scenario, and its self-relevance. We started with all of the relevant random effect terms with respect to both Participant ID and Scenario. After iteratively removing terms until the model converged, our final model had a random intercept for Participant ID and for Scenario, random slopes for target generosity, generosity relevance, self relevance, target reward, reward relevance, two-way interactions for Target Generosity x Generosity Relevance, Target Generosity x Self Relevance, Target Reward x Self Relevance, and Reward Relevance x Self Relevance, and the three-way interaction for Target Generosity x Generosity Relevance x Self Relevance with respect to Participant ID, and a random slope for Target Generosity with respect to Scenario. This model was similar to the pre-registration, but we did not pre-register random effects of scenario. The pre-registered version is reported in the supplemental materials (p. 56). As in the prior studies, we repeated an exploratory Bayesian version of this same analysis using brms (see Supplemental Materials p. 71).

5.1.4.2. Variability based on generosity and reward. To test the difference in variability between the effects of generosity and reward, we conducted an exploratory analysis in which we again computed difference scores of generous minus stingy targets in each self-relevant scenario, as well as difference scores of high-reward minus low-reward targets in each self-relevant scenario. For each participant, we then computed how variable these generosity scores and reward scores were across scenarios. Specifically, we compared each participant's standard deviation in generosity difference scores to their standard deviation in reward difference scores. Using a similar process to the one described for trait ratings in Study 3, we conducted a permutation test. We randomly shuffled the standard deviation labels and computed a difference score between them 10,000 times, in order to create a non-parametric null distribution.

5.2. Results

5.2.1. Preference judgments in hypothetical real-life situations

How did generosity and reward learning shape social preferences in hypothetical real-world scenarios, and to what extent did self-relevance and/or economic-relevance of a context's content influence reward-based decisions? To address this question, we regressed ratings on each target attribute and the interactions between these attributes and contextual-relevance features of the situations.

We observed a significant main effect of generosity, such that participants gave higher ratings to generous than to stingy targets ($b = 7.27$, $SE = 0.88$, $t(148.15) = 8.28$, $p < .001$). There was also a significant main effect of reward, such that participants gave higher ratings to rewarding than to non-rewarding targets ($b = 3.44$, $SE = 0.56$, $t(181.02) = 6.11$, $p < .001$).

There were significant two-way interactions for Target Generosity x Generosity-Relevance ($b = 2.19$, $SE = 0.46$, $t(25.28) = 4.77$, $p < .001$) and for Target Generosity x Self-Relevance ($b = 0.52$, $SE = 0.23$, $t(180.88) = 2.20$, $p = .03$). These interactions indicate that the extent to which people relied on their learning from targets' generosity depended on whether the decision-making context was relevant to generosity, as in prior studies, and whether it was a self-relevant decision. Participants were more likely to rate generous targets more positively in contexts related to generosity, and they were also more likely to rate generous targets more positively in self-relevant contexts. Notably, in an exploratory follow-up analysis setting low generosity relevance scenarios as the reference category, participants again preferred generous over stingy individuals, even in the least generosity-relevant scenarios ($b = 5.07$, $SE = 0.77$, $t(185.75) = 6.58$, $p < .001$). Thus, participants again always preferred generous (over stingy) partners, but this preference was greater when scenarios were more generosity-relevant.

Finally, we observed a Generosity x Generosity-Relevance x Self-Relevance interaction ($b = -0.69$, $SE = 0.19$, $t(179.80) = -3.70$, $p < .001$). This interaction indicated that there was a Generosity x Self-

Relevance interaction only within contexts that were less generosity-relevant ($b = 1.21$, $SE = 0.28$, $t(351.03) = 4.30$, $p < .001$. Specifically, in less generosity-relevant contexts, participants preferred generous targets in self-relevant contexts ($b = 6.28$, $SE = 0.96$, $t(94.02) = 6.51$, $p < .001$) and also preferred generous targets in prediction contexts but to a smaller extent ($b = 3.87$, $SE = 0.94$, $t(87.83) = 4.12$, $p < .001$). In generosity-relevant contexts, there was not evidence that participants favored generous targets based on whether the context was self-relevant or predictive ($b = -0.18$, $SE = 0.27$, $t(317.54) = -0.64$, $p = .52$).

In contrast, none of the interactions with target reward were statistically significant. Specifically, we did not observe a significant interaction of Reward x Economic-Relevance ($b = 0.26$, $SE = 0.16$, $t(9{,}602.82) = 1.68$, $p = .09$) or of Reward x Self-Relevance ($b = 0.17$, $SE = 0.21$, $t(181.26) = 0.77$, $p = .44$). A linear contrast revealed that the Generosity x Generosity-Relevance was significantly larger than the Reward x Economic-Relevance interaction ($b = 1.93$, $SE = 0.49$, $t(1) = 15.76$, $p < .001$). A Bayes Factor of $BF_{10} = 0.026$ also provides very strong evidence in favor of the null hypothesis that there is *not* a Reward x Economic-Relevance interaction. A Bayes Factor of $BF_{10} = 0.012$ similarly provides very strong evidence that there is not a Reward x Self-Relevance interaction ($BF_{10} < 1/30$) (Stefan et al., 2019). Unlike Study 3, the effect of reward-based learning thus did not depend on the economic-relevance or self-relevance of new scenarios. However, the Generosity x Self-Relevance interaction was not significantly larger than the Reward x Self-Relevance Interaction ($b = 0.35$, $SE = 0.32$, $t(1) = 1.22$, $p = .27$).

Consistent with our pre-registration, and in order to simplify the model, we did not examine the interaction between reward and generosity relevance or the interaction between generosity and economic-relevance. We made this decision because we did not expect decisions based on reward to depend on generosity relevance or decisions based on generosity to depend on economic relevance, as established in our prior studies. However, an additional analysis adding these interaction terms found that these interactions were not statistically significant (Reward x Generosity-Relevance $b = -0.02$, $SE = 0.19$, $t(181.2) = -0.10$, $p = .92$; Generosity x Economic-Relevance $b = -0.19$, $SE = 0.42$, $t(14.14) = -0.44$, $p = .67$). A Bayes Factor $BF_{10} = 0.007$ provides extreme evidence in support of the null hypothesis that there is not a Reward x Generosity-Relevance interaction ($BF_{10} < 1/100$), and a Bayes Factor $BF_{10} = 0.020$ provides very strong evidence that there is not a Generosity x Economic-Relevance interaction ($BF_{10} < 1/30$) (Stefan et al., 2019).

Altogether, there was no evidence for context-sensitivity of reward in Study 4, and furthermore, the effect of targets' generosity was more context-sensitive than was reward-based decision-making. These results are consistent with the primary hypothesis that reward-based decision-making is largely insensitive to the conceptual content of a decision, as compared to trait-based decision-making (see Fig. 10).

5.2.2. *Variability based on generosity and reward*

In Study 3, we found that generosity had a more flexible influence on semantic trait ratings when compared to reward; participants rated generous targets higher than stingy targets on some positive traits but not on others, whereas participants rated high-reward targets as slightly higher than low-reward targets to a similar degree across all positive traits. We exploratorily repeated this analysis with the Study 4 preference ratings to test whether participants also applied generosity knowledge more selectively than reward knowledge in these scenarios, limiting our analysis to the self-relevant condition. Again, the *p*-value represents the proportion of simulations on which the simulated value was greater than or equal to the observed difference in mean standard deviations.

Generosity difference scores were more variable ($M_{SD\_Generosity} = 14.77$) than reward difference scores ($M_{SD\_Reward} = 11.60$); our observed test statistic was 3.17, and the null distribution had a mean of 0.01 (*SD*

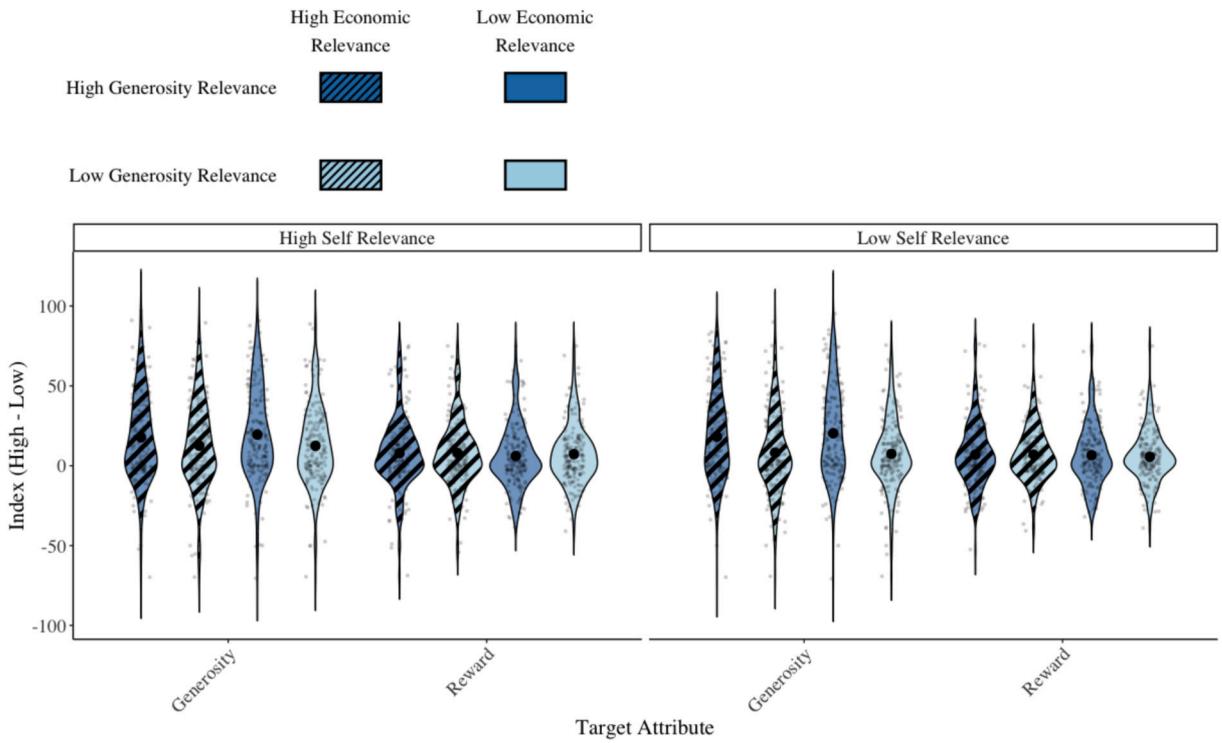

Fig. 10. Generosity-based decisions are sensitive to the generosity-relevance and self-relevance of a context, but reward-based decisions are more consistent across contexts, even those that vary in economic-relevance. The graph illustrates the difference scores between average ratings for the generous minus stingy targets and average ratings for the rewarding minus non-rewarding targets in Study 4. The error bars represent standard errors corrected for the within-subjects design. Individual participants' data points are represented as jittered dots.

= 0.74, 95% *CI* [−1.45, 1.47]). Results indicated that the standard deviation in the effect of target generosity on ratings across scenarios was higher than the standard deviation in the effect of reward ($M_{SD\_Difference}$ = 3.17, $p < .001$).

As in Study 3, we also compared these observed standard deviations to a null distribution in which we re-assigned the labels of all four targets, thus shuffling which targets were labeled as high or low, re-computed difference scores, and re-calculated the average standard deviation 10,000 times with this statistic ($M_{null} = 12.26$, $SD_{null} = 0.39$). Here, the *p*-values represent the proportion of simulations in which the standard deviation was greater than or equal to the observed standard deviation. Comparing the observed reward standard deviation to this distribution demonstrated that reward variability was not significantly greater than would be expected by chance ($p = .96$) and was directionally lower. However, the observed variability in generosity-based ratings was significantly greater than would be expected by chance ($p < .001$).

5.3. Discussion

By measuring participants' preferences regarding hypothetical social scenarios, Study 4 supported the hypothesis that reward history has a more general positive influence on social preferences across contexts, compared to a target's history of displaying the trait of generosity. After learning about the generosity and reward value of targets, participants preferred generous partners to a greater extent in scenarios that were conceptually relevant to generosity (e.g., asking someone for a loan) than in scenarios irrelevant to generosity (e.g., asking someone to join a dodgeball team). Although participants always preferred generous over stingy partners across every scenario, the size of this preference grew for the scenarios in which generosity was relevant. In contrast, participants preferred previously rewarding partners across these scenarios to a more similar extent, showing significantly lower variability in this preference compared to variability in generosity-based preferences.

In addition, we manipulated whether scenarios were economically-relevant (i.e., involved monetary gain) and self-relevant (i.e., whether participants reported personal preferences for interaction or mere predictions about behavior)—two dimensions of reward-relevance that were confounded in Study 3. Unlike Study 3, we observed no evidence of context-sensitivity for reward along either dimension. This finding suggests that economic reward is not a crucial ingredient driving the effects observed in Study 3, and that reward biases judgments of others to a similar degree across personal preferences and mere predictions. Instead, other features of Study 3, such as the mode of instrumental choice, might have shaped responses in that study—a question we return to in the General Discussion. Overall, however, these findings suggest that people apply learning from rewarding outcomes to a similar extent across social preferences for interaction partners in a wide array of scenarios.

6. General discussion

When people receive rewarding outcomes from others, do they seek out more interactions with these individuals and evaluate them more positively? In four studies, we found that receiving rewards from others engenders a general tendency to choose them as interaction partners and perceive them positively in new settings, even when these settings may not be conceptually related to the initial behaviors that provided reward. We contrasted this learning to trait inferences rooted in semantic memory, which informed decision-making more selectively to the extent that a new context was conceptually relevant to the original behavior that sparked a trait impression.

In each experiment, participants learned about targets that differed independently in how generous and rewarding they were to the participant. Importantly, participants were never explicitly told that a particular target was rewarding or generous but made these inferences themselves by learning from experience. In new settings, participants flexibly deployed their knowledge about generosity, as demonstrated by the fact that preferences for generous individuals increased the more a decision was conceptually related to generosity. For instance, they rated generous targets as highly trustworthy compared to stingy targets but had less of a preference for generous over stingy targets on less generosity-relevant traits. In the case of dominance, participants even preferred stingy to generous targets. In contrast, participants evaluated rewarding targets more positively than non-rewarding targets across the board—for instance, rating rewarding targets as trustworthy and dominant to a similar degree. Moreover, secondary Bayesian analyses provided evidence ranging from moderate to very strong in favor of the null hypothesis of no modulation of Reward by context, further suggesting that the effects of Reward did not vary by conceptual relevance of new situations. Whether participants were asked who they would want to hire as a computer programmer; who they thought was lucky, sociable, or attractive; who they would ask to pet sit for them; or who they expected to be good at dodgeball, they preferred individuals who had given them more points worth money in a game. These findings support the idea that rewards have a general positive influence on behavior and person perception, indiscriminately biasing us toward rewarding targets across multiple domains.

This work highlights the scope of reward learning in social interaction. While a wealth of work has illuminated impression formation through passive forms of learning—such as watching videos or reading text—interacting with others creates psychologically rewarding experiences that shape how we make sense of others. Accordingly, social learning through interaction may differ in meaningful ways from more passive impressions formed through observation.

In this manner, these findings also offer new insights into how distinct learning systems interact in social cognition. The human brain contains multiple systems for learning and memory that contribute to social behavior, including semantic knowledge about others (e.g., their character traits) and instrumental learning from rewards (Amodio, 2019). Recent work has begun to demonstrate how these systems interact, focusing on how semantic knowledge can bias instrumental learning. For instance, negative stereotypes about a group—a form of semantic knowledge (Amodio, 2019)—can bias reward learning about members of that group (Derreumaux et al., 2023; Schultner et al., 2024). However, the reverse may also be true, such that reward learning can bias the conceptual representations we form of others. The present findings, in which reward learning inflated a wide array of trait ratings, suggest a broad influence of this nature.

Importantly, these findings should not be taken to mean that participants *primarily* chose or evaluated partners based on reward feedback. In all studies, generosity had larger overall effects than reward, as participants primarily selected or rated partners positively based on how generous they had been; reward feedback provided a more modest boost to social preferences. However, the effects of trait knowledge appear large and selective, whereas the effects of reward learning appear smaller but more general. This modest benefit endowed by reward learning can nonetheless have meaningful impacts—for instance, in winner-take-all settings, leading one person to accrue outsized gains (Hackel & Zaki, 2018).

In addition, our use of the wording “trait selectivity” is not meant to imply that participants preferred generous partners *only* when generosity was highly relevant; participants preferred generous (over stingy) partners in almost all cases, but they had a *stronger* preference for generous partners in these scenarios. We suggest two explanations for this finding. First, even the low-generosity-relevance situations may not have been truly generosity-irrelevant; when choosing a trivia teammate, people still may find interacting with a generous person more pleasant than interacting with a stingy one, all things considered. This prediction aligns with research demonstrating the primacy of morality-related information in social cognition (Brambilla et al., 2021). Alternatively, experienced generosity may also serve as a kind of social reward

prompting positive feelings; indeed, prior work demonstrates that generosity learning in this task prompts activation in ventral striatum (Hackel et al., 2015)–a classical reward learning region–in addition to regions associated with impression formation. The total effect of generosity in this task may therefore include a selective component rooted in conceptual trait inference *and* a general positivity rooted in social reward processing. Future work can test these possibilities. Altogether, the present findings indicate that participants applied generosity knowledge in a *relatively* selective manner, guided by conceptual knowledge about traits and situations.

Past work has described a "halo effect" in impression formation, in which individuals with a positive attribute are assumed to also have other positive attributes (Nisbett and Wilson, 1977). The present work differs from that work in two key respects. First, in traditional studies of the halo effect, the attribute studied is an endogenous feature of the target person (e.g., attractiveness); in the present work, rewards were a feature of the participant's experience in a social interaction with the target, exogenously determined by the experimenters and absent after the learning phase. More importantly, past work indicates that the halo effect is not as general as often assumed; instead, the halo effect shows trait selectivity (Laham & Forgas, 2022). For instance, attractive individuals are believed by others to be socially competent but not to have higher integrity than their less attractive counterparts (Eagly et al., 1991); attractive individuals are even judged more negatively on some traits, such as vanity (Dermer & Thiel, 1975; Han & Laurent, 2023). Halo effects have therefore been proposed to depend on implicit personality theories that specify how different features of a person relate to one another (Laham & Forgas, 2022), rendering these effects empirically and theoretically more similar to the trait impressions studied here. In contrast, reward learning showed no evidence of context selectivity in the present work, offering a more general influence across trait ratings, choices, and preferences. Altogether, rather than using an endogenous feature of a person (e.g., attractiveness) to make selective inferences about related features (e.g., sociability), participants used an exogenously determined aspect of their interaction to draw globally positive inferences about the interaction partner.

6.1. Mechanisms of reward learning in social interaction

Why did reward feedback have a general positive influence on social approach and impressions? We suggest that two psychological mechanisms might promote the observed general positive influence of rewards. First, instrumental learning may have a direct effect on future behavior, such that when people receive reinforcing feedback, they develop habitual tendencies to repeat the actions. While habits are sensitive to the concrete context in which an action takes place, they are insensitive to the goals of an interaction (Wood, 2017), such as approaching someone based on their generosity versus their competence.

However, this pathway alone would not explain why rewards bias trait impressions, which do not involve approach behavior. A second explanation, as proposed by Clore and Byrne (1974), is that the influence of reward on behavior may be mediated, at least in part, by a conscious experience of liking, which people come to have for a rewarding stimulus in situations in which the rewarding feeling seems to be "about" the stimulus (Higgins, 1998). In new situations in which the feeling of liking is not attributed to other sources, people might interpret this positive feeling as a source of information guiding judgments of the target (Schwarz, 2012). Under this explanation, there may still be contextual factors that influence whether one's feelings as a *whole* are relevant to a particular decision (i.e., whether the decision is relevant to oneself, involves hedonic pleasure, and is made under low processing capacity; Raghunathan & Pham, 1999; Pham, 1998; Schwarz, 2012). However, the informational content of these feelings should engender a general positive boost to distinct contexts, regardless of the social meaning of behaviors involved.

Alternatively, rather than a bias at the time of judgment, it is also possible that a conceptual bias occurs at the time of reward feedback. In other words, rather than interpreting a positive feeling at the time of judgment as providing information for how we should respond to a particular stimulus, reward feedback may immediately alter our conceptual understanding of another person. The current project does not distinguish between these possibilities. The latter explanation (bias at feedback) suggests a more stable conceptual representation that is altered by reward, while the former explanation (bias at judgment) suggests an affective influence of reward that we reference each time we make a judgment. Future work might investigate these possibilities by manipulating whether people attribute their positive feelings about rewarding targets to a different source at learning or later judgment and then determining whether people still prefer these individuals.

These two overall mechanisms—habit and a positive effect of one's feelings (whether at the time of learning or judgment)—might also explain a discrepancy between two of the present studies. Specifically, Study 3 did reveal a source of context-sensitivity in reward learning: participants preferred rewarding partners more often when choosing partners in economic games as opposed to predicting how targets treated others. In contrast, in Study 4, participants rated rewarding targets similarly regardless of economic-relevance (whether a scenario involved monetary gain) and self-relevance (whether a rating expressed a preference for interaction or a prediction about behavior) (see also Supplemental Study 1 in Supplemental Materials, p. 2).

Why did Study 3 observe an effect of choice versus prediction, whereas Study 4 did not? One possibility is that the choice condition of Study 3 invoked a habit pathway due to the instrumental nature of the choice, as participants directly performed an immediate action for perceived benefit. Reward cues can enhance the vigor of instrumental behavior (Lovibond & Colagiuri, 2013), and people often express instrumental actions habitually in the immediate choice context in which a behavior was learned (Wood, 2017). In contrast, Study 4 collected more distanced preference ratings, rather than immediate actions, and these ratings might have depended on a more conceptual bias pathway alone. Such a pathway may have a more general effect across preferences and predictions, if it makes the targets appear to have more positive attributes. Future work can aim to replicate the Study 3 finding and more directly test these possible pathways. Nonetheless, the present findings robustly demonstrate that reward learning is applied more evenly than trait learning across scenarios that vary in social meaning (e.g., trust games and trivia games); across trait ratings that vary in their relevance to a behavior previously observed (e.g., generosity and competence); and across preferences that vary in their economic- or self-relevance (e.g., asking someone for a loan or predicting who would be a good actor for a play).

The distinct effects of reward-based and semantic evaluative learning may also speak to a broader debate in the literature as to how attitudes more generally could operate via associative and/or propositional processes. Briefly, some models propose that implicit measures capture primarily simple learned associations between an attitude object and valence (e.g., "vaccines = good"), whereas explicit measures capture propositional knowledge about the object with evaluative implications (e.g., "vaccines prevent severe illness") (Gawronski & Bodenhausen, 2011). Alternatively, some have argued that even implicit reactions are fundamentally propositional in nature (De Houwer et al., 2020). Much of this debate has occurred in the context of research on evaluative conditioning, which typically finds that a target stimulus will acquire the valence of whatever it is presented alongside (Moran et al., 2023). While some evidence highlights how propositional reasoning can affect implicit reactions that emerge from such simple co-occurrences (De Houwer et al., 2020), the present studies may suggest that there is still room for non-propositional accounts of evaluative learning. Indeed, although a recent meta-analysis found that explicit semantic information about pairs of stimuli plays a dominant role in evaluative learning, the mere co-ocurrence of the stimuli still accounts for some variance in acquired attitudes (Kurdi et al., 2023). The finding that our participants

tended to like rewarding targets in a relatively context-free manner that was not driven by specific trait inferences may suggest a limited role of semantic inferences drawn through logical consistency supporting these reward-based judgments. We invite future work considering the parallels between the literature on reward versus trait learning and theories regarding associative versus propositional processes driving basic attitudes.

Finally, these findings also have implications for non-social models of reward learning. Models of reinforcement learning often assume that people either use an internal model of the world to guide choices (model-based learning) or else simply repeat rewarded actions while using no model of the world (model-free learning). However, an intermediate possibility raised in past work is that rewards bias people's internal model of the world, changing conceptual representations and inferences (Babür et al., 2024; Fischer et al., 2017; Melnikoff & Strohminger, 2024). If so, individuals might repeat rewarded actions because they use a *biased* model of the world, not because they used *no* model of the world. The present findings are suggestive of such an effect, and future work can explore these rich possibilities in the non-social domain.

6.2. Limitations and future directions

In the present work, we dissociated learning from rewards and learning from generosity, building on past work. Future research can also assess whether these findings generalize to other forms of trait learning and instrumental learning. For example, we would hypothesize that a similar pattern can shape learning from traits and rewards in settings invoking other trait dimensions, such as competence. Indeed, Hackel et al. (2022) asked participants to learn about targets who received points from scratch-off lotteries and earned a proportion based on their trivia competence. Rewards biased impressions of competence and lottery success to a similar degree. Future work can examine this influence across a broader array of trait dimensions choices.

This work specifically manipulated instrumental learning, requiring participants to take action to interact with their interaction partner and then receive a reward. Future work can test whether these findings generalize to Pavlovian conditioning, in which people more passively learn that a person is a predictive cue for reward, without requiring an instrumental decision. Pavlovian learning may lead to similar effects, given that it can also involve a positively valenced outcomes; on the other hand, it is possible that receiving feedback on one's actions plays a role in the present effects–for instance via approach motivation, habit, or a gut feeling reflecting instrumental value. In addition, the only type of instrumental learning we investigated was positive reinforcement (i.e., forming positive associations with targets who disseminated higher rewards). Future work can test the effects of negative reinforcement, asking whether participants learn to negatively evaluate and avoid targets they associate with loss and, further, whether the addition or removal of punishing stimuli functions similarly to the gain or loss of rewarding stimuli. If people experience negative feelings about individuals who administer punishments and positive feelings about individuals who remove punishing stimuli, these forms of learning might also influence feelings more broadly.

These findings also leave room to further investigate the circumstances under which reward-learning in social interaction may be context-sensitive. As discussed above, the present studies raise the possibility that this may depend on the mode of the decision (i.e., instrumental action versus judgment), rather than the content of the decision (e.g., when a decision involves a different action, takes place in a different setting, or when the decision-maker is in a different state). At the same time, past work suggests people do often generalize reward based on perceptual similarity, in addition to conceptual similarity (FeldmanHall et al., 2018; Shepard, 1987); although this effect was not observed in Study 2, the present work cannot rule out the existence of such effects.

In addition, research on the MODE model highlights how automatically activated global attitudes toward a stimulus bias subsequent judgments and behavior unless people are motivated and able to consciously bring additional considerations to bear on that outcome (Fazio & Olson, 2014). Therefore, the general positivity associated with more rewarding targets may have less influence in circumstances that encourage people to correct for any bias they might have or to explicitly consider some other objective characteristic when making a decision.

Indeed, as described by Schwarz (2012), people rely on feelings more when they are less motivated to make systematic judgments (Rothman & Schwarz, 1998), have lower processing capacity (Siemer & Reisenzein, 1998), are experts in a particular domain (Ottati & Isbell, 1996), or can attribute their feelings to another source (Schwarz & Clore, 1983). Demonstrating that these conditions moderate the influence of rewards on decision-making would provide evidence that people are using their positive feelings to guide their decision-making when they experience rewards at the time of judgment rather than the time of encoding. Alternatively, rewards may have an immediate biasing effect on person perception that later persists across these changes in circumstance. Future work can also test the extent to which participants are aware of their own reward learning or experience reward associations as a general intuition. Similarly, future work can examine the extent to which trait-based generalization relies on controlled processing, which may be required to evaluate whether or not trait information applies to a particular context. For example, under conditions of cognitive load during judgment, people may rely more on affective components of learning or at least may fail to make choices that are as selective based on traits.

Finally, the present findings may have implications for social behaviors related to prejudice and stereotyping, which can be explored. To the extent that stereotypes are rooted in semantic knowledge whereas prejudice is rooted in affective learning (Amodio, 2014), it is possible that stereotypes may also apply more flexibly to new situations while prejudice may have a more consistent general influence across a wider range of novel contexts (see also Amodio, 2025). These findings may also have implications for perpetuating wealth inequality. Given the same degree of proportional generosity, a wealthier individual provides others with greater reward value than a less wealthy counterpart. Past work demonstrates that people therefore reciprocate more with wealthier individuals in cooperative games, even when there is no strategic reason to do so, by sharing more with them in return and giving them a better reputation (Hackel & Zaki, 2018). However, if reward learning casts a positive glow across multiple domains of person perception, then people may be more likely to select wealthy individuals for a range of opportunities. Furthermore, future research could consider how easily these different sorts of learned associations can change. Just as more associative evaluations are posited to endure more than proposition-based ones (Gawronski & Bodenhausen, 2007), perhaps reward learning is more difficult to extinguish than trait learning. This could mean that prejudice is less susceptible to change than are stereotypes. Nevertheless, trait impressions vary in their likelihood of change (see Luttrell et al., 2022)

7. Conclusion

Across four studies, we found evidence that learning to associate people with rewards leads to general positive regard. After learning that other individuals provided large rewards—regardless of their generous character—participants preferred these individuals as trivia players, actors in a play, and dodgeball teammates, in addition to rating them as more attractive, generous, and intelligent. Reward-based decision-making was thus relatively insensitive to the conceptual content of a decision. In contrast, when other individuals demonstrated generous character by sharing a large proportion of their resources—regardless of the absolute rewards they provided—participants especially preferred these individuals for situations relevant to generosity. These findings

arose when participants learned about other individuals' through making inferences about their character based on their behavior in a game and directly experiencing differentially rewarding outcomes, without ever being given explicit information about whether these individuals were rewarding or generous.

This work highlights how distinct forms of learning contribute to and interact in decision-making. The findings also explain why we might feel an inexplicable pull toward some individuals without being able to trace this feeling to a specific instance that makes them a good fit for a task. Although trait inferences tend to have a larger impact on individual decisions, the smaller-but-consistent influences of reward associations over time and among our whole social networks provide a meaningful influence on our perceptions of those around us and on the individuals toward whom we gravitate.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used ChatGPT in order to generate stimuli and to contribute to analysis and task code. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Open practices

Study materials, analysis code, and de-identified data are publicly available at https://osf.io/k95cy/overview?view_only=d9f18f98fc37450cb4173b5960d01e03. The studies were pre-registered prior to data collection, and these pre-registrations can be accessed at https://researchbox.org/5570?PEER_REVIEW_passcode=QZETGG.

CRediT authorship contribution statement

Kira Harris: Writing – original draft, Visualization, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Andrew Luttrell:** Writing – review & editing, Supervision, Resources, Methodology, Conceptualization. **Peter Mende-Siedlecki:** Writing – review & editing, Supervision, Resources, Methodology, Conceptualization. **Leor M. Hackel:** Writing – review & editing, Visualization, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

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.jesp.2026.104968.

References

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience, 15*(10), 670–682. https://doi.org/10.1038/nrn3800

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

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

Babür, B. G., Leong, Y. C., Pan, C. X., & Hackel, L. M. (2024). Neural responses to social rejection reflect dissociable learning about relational value and reward. *Proceedings of the National Academy of Sciences, 121*(49), Article e2400022121. https://doi.org/10.1073/pnas.2400022121

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed effects models. *Frontiers in Psychology, 4*, 328.

Berridge, K. C., Robinson, T. E., & Aldridge, J. W. (2009). Dissecting components of reward: 'Liking', 'wanting', and learning. *Current Opinion in Pharmacology, 9*(1), 65–73. https://doi.org/10.1016/j.coph.2008.12.014

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

Bürkner, P. C. (2017). brms: An R Package for Bayesian multilevel models using stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Cho, H. J., & Hackel, L. M. (2022). Instrumental learning of social affiliation through outcome and intention. *Jounal of Expeirmental Psychology: General, 151*(9), 2204–2221. https://doi.org/10.1037/xge0001190

Clore, G. L., & Byrne, D. (1974). A reinforcement-affect model of attraction. In T. L. Huston (Ed.), *Foundations of interpersonal attraction* (pp. 143–170). New York: Academic Press.

Connor, P., Nicolas, G., Antonoplis, S., & Koch, A. (2026). Unconstrained descriptions of Facebook profile pictures support high-dimensional models of impression formation. *Personality and Social Psychology Bulletin, 52*(1), 53–69.

Cushman, F. (2024). Computational social psychology. *Annual Review of Psychology, 75*(1), 625–652.

Daniel, R., & Pollmann, S. (2014). A universal role of the ventral striatum in reward-based learning: Evidence from human studies. *Neurobiology of Learning and Memory, 114*, 90–100.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204–1215.

De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes beyond associations: On the role of propositional representations in stimulus evaluation. In , *61. Advances in experimental social psychology* (pp. 127–183). Academic Press.

Dermer, M., & Thiel, D. L. (1975). When beauty may fail. *Journal of Personality and Social Psychology, 31*(6), 1168–1176. https://doi.org/10.1037/h0077085

Derreumaux, Y., Elder, J., Suri, G., Ben-Zeev, A., Quimby, T., & Hughes, B. L. (2023). Stereotypes disrupt probabilistic category learning. *Journal of Experimental Psychology: General, 152*(6), 1622–1638.

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience, 18*, 767–772. https://doi.org/10.1038/nn.3981

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but…: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin, 110*(1), 109.

Elder, J., Cheung, B., Davis, T., & Hughes, B. (2023). Mapping the self: A network approach for understanding psychological and neural representations of self-concept structure. *Journal of Personality and Social Psychology, 124*(2), 237–263. https://doi.org/10.1037/pspa0000315

Fazio, R. H., & Olson, M. A. (2014). The MODE model: Attitude-behavior processes as a function of motivation and opportunity. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual process theories of the social mind* (pp. 155–171). New York: Guilford Press.

FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences, 115*(7), E1690–E1697. https://doi.org/10.1073/pnas.1715227115

Fischer, A. G., Bourgeois-Gironde, S., & Ullsperger, M. (2017). Short-term reward experience biases inference despite dissociable neural correlates. *Nature Communications, 8*, Article 1690. https://doi.org/10.1038/s41467-017-01703-0

Frolichs, K. M., Rosenblau, G., & Korn, C. W. (2022). Incorporating social knowledge structures into computational models. *Nature Communications, 13*(1), 6205.

Gawronski, B, & Bodenhausen, G. V (2007). Unraveling the processes underlying evaluation: Attitudes from the perspective of the APE model. *Social Cognition, 25*(5), 687–717.

Gawronski, B., & Bodenhausen, G. V. (2011). The associative–Propositional evaluation model. In , *44. Advances in experimental social psychology* (pp. 59–127). Elsevier. https://doi.org/10.1016/B978-0-12-385522-0.00002-0.

Gershman, S. J., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences, 5*, 43–50. https://doi.org/10.1016/j.cobeha.2015.07.007

Hackel, L. M., Berg, J. J., Lindström, B. R., & Amodio, D. M. (2019). Model-based and model-free social cognition: Investigating the role of habit in social attitude formation and choice. *Frontiers in Psychology, 10*, 2592. https://doi.org/10.3389/fpsyg.2019.02592

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience, 18*(9), 1233–1235. https://doi.org/10.1038/nn.4080

Hackel, L. M., & Mende-Siedlecki, P. (2023). Two modes of social impressions and their effects on choice. *Journal of Experimental Psychology: General, 152*(11), 3002–3020. https://doi.org/10.1037/xge0001432

Hackel, L. M., Mende-Siedlecki, P., & Amodio, D. M. (2020). Reinforcement learning in social interaction: The distinguishing role of trait inference. *Journal of Experimental Social Psychology, 88*, Article 103948. https://doi.org/10.1016/j.jesp.2019.103948

Hackel, L. M., Mende-Siedlecki, P., Loken, S., & Amodio, D. M. (2022). Context-dependent learning in social interaction: Trait impressions support flexible social choices. *Journal of Personality and Social Psychology, 123*(4), 655–675. https://doi.org/10.1037/pspa0000296

Hackel, L. M., & Zaki, J. (2018). Propagation of economic inequality through reciprocity and reputation. *Psychological Science, 29*(4), 604–613. https://doi.org/10.1177/0956797617741720

Han, D. E., & Laurent, S. M. (2023). Beautiful seems good, but perhaps not in every way: Linking attractiveness to moral evaluation through perceived vanity. *Journal of Personality and Social Psychology, 124*(2), 264–286. https://doi-org.libproxy2.usc.edu/10.1037/pspa0000317.

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins, & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133–168). New York, NY: Guilford Press.

Higgins, E. T. (1998). The aboutness principle: A pervasive influence on human inference. *Social Cognition, 16*(1), 173–198. https://doi.org/10.1521/soco.1998.16.1.173

Hsu, M., Anen, C., & Quartz, S. R. (2008). The right and the good: Distributive justice and neural encoding of equity and efficiency. *Science, 320*(5879), 1092–1095.

Izuma, K., & Adolphs, R. (2013). Social manipulation of preference in the human brain. *Neuron, 78*(3), 563–573.

Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. *Journal of Personality and Social Psychology, 89*, 899–913.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1), 54069. https://doi.org/10.1037/a0028347

Kahnt, T., Park, S. Q., Burke, C. J., & Tobler, P. N. (2012). How glitter relates to gold: Similarity-dependent reward prediction errors in the human striatum. *The Journal of Neuroscience, 32*(46), 16521–16529. https://doi.org/10.1523/JNEUROSCI.2383-12.2012

Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology, 110*(5), 675.

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

Laham, S. M., & Forgas, J. P. (2022). Halo effects. In R. F. Pohl (Ed.), *Cognitive illusions* (3rd ed., pp. 259–271). Routledge. https://doi.org/10.4324/9781003154730-19.

Lin, C., Keles, U., & Adolphs, R. (2021). Four dimensions characterize attributions from faces using a representative set of English trait words. *Nature Communications, 12*(1), 5168.

Lin, C., & Thornton, M. (2023). Evidence for bidirectional causation between trait and mental state inferences. *Journal of Experimental Social Psychology, 108*, Article 104495.

Liu, Y., Charlesworth, T., Koch, A., Luttrell, A., & Jackson, J. C. (2025). *The content, structure, and history of English trait words*. https://doi.org/10.31234/osf.io/7qkg8_v1

Lovibond, P. F., & Colagiuri, B. (2013). Facilitation of voluntary goal-directed action by reward cues. *Psychological Science, 24*(10), 2030–2037.

Lu, J., & Lin, C. (2025). Network models reveal high-dimensional social inferences in naturalistic settings beyond latent construct models. *Communications Psychology, 3*(1), 98.

Luo, J., Mende-Siedlecki, P., & Hackel, L. M. (2025). Rewards bias self-evaluations of ability. *Communications Psychology, 3*(1), 143.

Luttrell, A., Sacchi, S., & Brambilla, M. (2022). Changing impressions in competence-oriented domains: The primacy of morality endures. *Journal of Experimental Social Psychology, 98*, Article 104246.

Melnikoff, D. E., & Strohminger, N. (2024). Bayesianism and wishful thinking are compatible. *Nature Human Behaviour, 8*(4), 692–701.

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.

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

Mende-Siedlecki, P., & Todorov, A. (2016). Neural dissociations between meaningful and mere inconsistency in impression updating. *Social Cognitive and Affective Neuroscience, 11*(9), 1489–1500. https://doi.org/10.1093/scan/nsw058

Moran, T., Nudler, Y., & Bar-Anan, Y. (2023). Evaluative conditioning: Past, present, and future. *Annual Review of Psychology, 74*(1), 245–269.

Mullin, G., Mende-Siedlecki, P., & Hackel, L. M. (2025). *Conceptual maps of social knowledge support trait generalization* (Manuscript submitted for publication).

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35*(4), 250–256. https://doi.org/10.1037/0022-3514.35.4.250

Niv, Y., Joel, D., & Dayan, P. (2006). A normative perspective on motivation. *Trends in Cognitive Sciences, 10*(8), 375–381. https://doi.org/10.1016/j.tics.2006.06.010

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*(32), 11087–11092.

Ottati, V. C., & Isbell, L. M. (1996). Effects of mood during exposure to target information on subsequently reported judgments: An on-line model of misattribution and correction. *Journal of Personality and Social Psychology, 71*, 39–53.

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience, 5*(2), 97–98.

Pham, M. T. (1998). Representativeness, relevance, and the use of feelings in decision making. *Journal of Consumer Research, 25*(2), 144–159.

Raghunathan, R., & Pham, M. T. (1999). All negative moods are not equal: Motivational influences of anxiety and sadness on decision making. *Organizational Behavior and Human Decision Processes, 79*(1), 56–77.

Rothman, A. J., & Schwarz, N. (1998). Constructing perceptions of vulnerability: Personal relevance and the use of experiential information in health judgments. *Personality and Social Psychology Bulletin, 24*, 1053–1064.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decisionmaking in the ultimatum game. *Science, 300*(5626), 1755–1758.

Schultner, D. T., Stillerman, B. S., Lindström, B. R., Hackel., L. M., Hagen, D. R., Jostmann, N. B., & Amodio, D. M. (2024). Transmission of societal stereotypes to individual-level prejudice through instrumental learning. *Proceedings of the National Academy of Sciences of the United States of America, 12*(45), Article e2414518121. https://doi.org/10.1073/pnas.2414518121

Schwarz, N. (2012). Feelings-as-information theory. In P. Van Lange, A. Kruglanski, & E. Higgins (Eds.), *Handbook of theories of social psychology: Volume 1* (pp. 289–308). SAGE Publications Ltd.. https://doi.org/10.4135/9781446249215.n15

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45*, 513–523.

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

Siemer, M., & Reisenzein, R. (1998). Effects of mood on evaluative judgements: Influence of reduced processing capacity and mood salience. *Cognition & Emotion, 12*, 783–805.

Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *, 31(1). Proceedings of the AAAI conference on artificial intelligence.* , February.

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., et al. (2019). A tutorial on Bayes factor design analysis using an informed prior. *Behavior Research Methods, 51*, 1042–1058. https://doi.org/10.3758/s13428-018-01189-8

Stolier, R. M., Hehman, E., & Freeman, J. B. (2020). Trait knowledge forms a common structure across social cognition. *Nature Human Behaviour, 4*(4), 361–371. https://doi.org/10.1038/s41562-019-0800-6

Tamir, D. I., & Hughes, B. L. (2018). Social rewards: From basic social building blocks to complex social behavior. *Perspectives on Psychological Science, 13*(6), 700–717. https://doi.org/10.1177/1745691618776263

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: Evidence from a false recognition paradigm. *Journal of Personality and Social Psychology, 83*(5), 1051–1065.

Todorov, A., & Uleman, J. S. (2003). The efficiency of binding spontaneous trait inferences to actors' faces. *Journal of Experimental Social Psychology, 39*(6), 549–562.

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

Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology, 47*(2), 237–252. https://doi.org/10.1037//0022-3514.47.2.237

Wood, W. (2017). Habit in personality and social psychology. *Personality and Social Psychology Review, 21*(4), 389–403. https://doi.org/10.1177/1088868317720362

Zaki, J., Schirmer, J., & Mitchell, J. P. (2011). Social influence modulates the neural computation of value. *Psycholical Science, 22*(7), 894–900.

Zanna, M. P., & Hamilton, D. L. (1972). Attribute dimensions and patterns of trait inferences. *Psychonomic Science, 27*(6), 353–354. https://doi.org/10.3758/BF03328989