Check for updates

Research Article

Psychological Science
1–22
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09567976251351304
www.psychologicalscience.org/PS

Reward Association With Mental States Shapes Empathy and Prosocial Behavior

Yi Zhang and Leor Hackel
Department of Psychology, University of Southern California

Abstract

Valuing the welfare of others is a fundamental aspect of empathy and prosocial behavior. How do people develop this valuation? Theories of associative learning suggest that people can associate social cues, such as smiles, with personal reward, thus feeling good when others thrive. Yet people often display *generalized* concern for others' welfare, regardless of the specific cues present. We propose that Pavlovian conditioning allows people to associate reward directly with others' abstract mental states, learning that another's happiness predicts their own reward. In four online experiments with 1,500 U.S.-based adults recruited from CloudResearch, participants' monetary outcomes were congruently or incongruently predicted by a target's mental states. Participants who experienced congruent learning reported more empathic feelings toward the target in novel situations. The values attached to mental states further influenced participants' prosocial choices. These results demonstrate how associative learning of abstract mental states can give rise to generalizable empathy and influence moral behavior.

Keywords

Pavlovian conditioning, associative learning, empathy, vicarious reward, prosocial behavior

Received 6/7/24; Revision accepted 5/20/25

Although humans value rewards such as food and money, people also derive reward from the positive outcomes of their fellows. This phenomenon, known as "vicarious reward," is a key component of empathy: Because people value others' welfare, they feel good when others do well and feel bad when others suffer (Davis, 1983; Mobbs et al., 2009; Morelli et al., 2015). These feelings can motivate prosocial behavior: When people value others' welfare, they endure costs to improve that welfare (Contreras-Huerta et al., 2023; Gęsiarz & Crockett, 2015; Kleiman-Weiner et al., 2017; Lockwood et al., 2017). How do people come to derive reward from the welfare of others such that their emotional responses align with the gains and losses of their fellows?

One account of this ability comes from reward-related learning mechanisms, including Pavlovian conditioning (Pavlov, 1927; Rescorla, 1988). When a neutral stimulus is paired with an intrinsically rewarding or aversive stimulus, the neutral stimulus will acquire similar rewarding or aversive values and become capable of eliciting conditioned responses by itself. This logic has been used to explain the emergence of empathy in children: A mother's smile that consistently precedes a child's own comfort, for example, can become a predictive cue for reward, leading the child to experience vicarious positive affect on seeing their mother smile (Hoffman, 1985). This mechanism has also been used to explain the emergence of counterempathy among adults. In "zero-sum" competitions, in which the success of one person predicts the failure of others, individuals may attach reward to the negative outcomes of others and experience counterempathic emotions such as schadenfreude (Cikara, 2018). Indeed, conditioning a target's smiling face with negative outcomes can lead an observer to respond counterempathically (Englis et al., 1982; Yamada et al., 2011).

These examples illustrate how people come to experience vicarious reward toward *concrete* outcomes

Corresponding Author:
Leor Hackel, Department of Psychology, University of Southern California
Email: lhackel@usc.edu

involving others (e.g., a smiling face). Yet people often show *generalized* empathy, experiencing vicarious reward across a vast array of distinct contexts, such as a coworker getting engaged or being promoted, regardless of the specific cues present. If people learn that one outcome (e.g., a coworker's promotion) relates to their own reward, how might they generalize this learning to new outcomes (e.g., the coworker's engagement) that also constitute that person's abstract welfare but involve different concrete cues?

Here, we identify a psychological process that can promote generalizable empathy. Although people can associate concrete stimuli with rewarding or aversive outcomes, Pavlovian conditioning can also apply to abstract concepts (e.g., "mammals"), leading people to generalize associations to new category members (Dunsmoor & Murphy, 2015). In social interactions, humans easily construe social scenes in terms of abstract concepts and can associate these concepts with reward (Hackel & Kalkstein, 2023; Hackel et al., 2024). We propose that the affective states of others constitute one such type of abstract information that can be associated with reward. When seeing a social scene, people readily infer the mental states of others, including positive or negative affect (Heider & Simmel, 1944; Meeren et al., 2005; Tracy & Robins, 2008; Wagner et al., 2011). These affect representations can then serve as a common input to Pavlovian conditioning processes and become associated with reward. As a result, people may intuitively feel good or bad when others experience positive or negative affect in novel situations, thereby empathizing with their general welfare.

In turn, Pavlovian conditioning may shape not only empathic feelings but also behaviors. First, the presence of a Pavlovian cue (e.g., food) can facilitate instrumental actions to obtain reward—a phenomenon called Pavlovian-to-instrumental transfer (PIT; Lovibond, 1983). Accordingly, people might become more reward-seeking in the presence of others' feelings associated with reward. Second, Pavlovian learning can also impact instrumental choice in a more specific way: It leads people to choose cues previously associated with Pavlovian reward, even when instrumental reward contingencies differ from prior Pavlovian contingencies (Lindström et al., 2019). Through this influence, once people associate the positive feelings of another person with reward, they might more readily choose actions that elicit positive feelings within the target, resulting in prosocial behavior, even when these actions induce cost for themselves.

In four experiments, we tested whether Pavlovian conditioning can influence people's valuation of others' abstract mental states, thereby influencing empathy and prosocial choices. Participants completed a Pavlovian conditioning task in which a target's outcomes predicted their own monetary outcomes in either a congruent or incongruent manner such that the target's positive outcomes predicted the participant's gains and the target's negative outcomes predicted the participant's losses, or vice versa. Crucially, each stimulus seen during learning depicted a unique event in which the target experienced positive or negative affect. Stimuli thus shared no concrete visual cues; only an abstract representation of the target's affect could bind together stimuli that predicted gain or loss. We hypothesized that conditioning would influence participants' empathic feelings toward the target in novel scenarios: When seeing new representations of the target's positive or negative affect, participants would experience more congruent or incongruent emotions depending on their training, even when participants could earn no further reward. We further tested whether this learning influenced instrumental behavior, including traditional measures of PIT and prosocial decisions that would impact the other's happiness.

Research Transparency Statement

General disclosures

Conflicts of interest: All authors declare no conflicts of interest. **Funding:** This research received no specific funding. **Artificial intelligence:** No AI-assisted technologies were used in this research or the creation of this article. **Ethics:** This research received approval from the University of Southern California Institutional Review Board (ID: UP-19-00404). **Open Science Framework:** To ensure long-term preservation, all OSF files were registered at https://osf.io/58wsu.

Experiment 1a disclosures

Preregistration: The hypotheses, method, and analysis plans were preregistered (https://aspredicted.org/hv4g2.pdf) on December 9, 2022, prior to data collection, which began on December 12, 2022. There were deviations from the preregistration (for details, see the Method section for Experiment 1a below and Table S13 in the Supplemental Material available online). **Materials:** All study materials are publicly available (https://osf.io/58wsu). **Data:** All primary data are publicly available (https://osf.io/58wsu). **Analysis scripts:** All analysis scripts are publicly available (https://osf.io/58wsu). **Computational reproducibility:** The computational reproducibility of the results has been independently confirmed by the journal's STAR team.

Experiment 1b disclosures

Preregistration: The hypotheses, method, and analysis plans were preregistered (https://aspredicted.org/9wi5x.pdf) on July 17, 2023, prior to data collection, which began on August 7, 2023. There were deviations from the preregistration (for details, see the Method section for Experiment 1b and Table S13). **Materials:** All study materials are publicly available (https://osf.io/58wsu). **Data:** All primary data are publicly available (https://osf.io/58wsu). **Analysis scripts:** All analysis scripts are publicly available (https://osf.io/58wsu). **Computational reproducibility:** The computational reproducibility of the results has been independently confirmed by the journal's STAR team.

Experiment 2 disclosures

Preregistration: The hypotheses, method, and analysis plans were preregistered (https://aspredicted.org/bmst-6n2z.pdf) on October 31, 2024, prior to data collection, which began on November 2, 2024. There were deviations from the preregistration (for details, see the Method section for Experiment 3 and Table S13). **Materials:** All study materials are publicly available (https://osf.io/58wsu). **Data:** All primary data are publicly available (https://osf.io/58wsu). **Analysis scripts:** All analysis scripts are publicly available (https://osf.io/58wsu). **Computational reproducibility:** The computational reproducibility of the results has been independently confirmed by the journal's STAR team.

Experiment 3 disclosures

Preregistration: The hypotheses, method, and analysis plans were preregistered (https://aspredicted.org/zb4cs.pdf) on December 3, 2023. Because the current experiment used a novel paradigm, we checked data from a small batch of participants (n = 50) to ensure they understood the task and followed the instructions. All remaining data were collected beginning on December 5, 2023, after the preregistration, and no further analyses were conducted before data collection was completed. In Section 8 of the preregistration, we explain why we still consider this to be a valid preregistration. There were deviations from the preregistration (for details, see Table S13). **Materials:** All study materials are publicly available (https://osf.io/58wsu). **Data:** All primary data are publicly available (https://osf.io/58wsu). **Analysis scripts:** All analysis scripts are publicly available (https://osf.io/58wsu). **Computational reproducibility:** The computational reproducibility of the results has been independently confirmed by the journal's STAR team.

Experiments 1a and 1b: Pavlovian Conditioning Promotes Empathic Feelings

Method

Overview. In Experiments 1a and 1b, we asked whether Pavlovian conditioning can influence people's valuation of another person's abstract mental states and whether this process influences empathic feelings. Experiments 1a and 1b had the same general structure, which was pretested in a pilot experiment (see the Supplemental Material). Participants first completed a Pavlovian conditioning task, learning that a target's mental states (the conditioned stimuli; CSs) predicted their own monetary outcomes (the unconditioned stimuli; USs) in either a congruent or incongruent manner. Next, to link this learning to a classic marker of Pavlovian conditioning, participants completed a PIT task that measured how the rewards attached to the target's mental states influenced their instrumental reward-seeking behavior (Cartoni et al., 2016). Last, participants reported their feelings toward novel positive or negative target outcomes, which served as our measure of empathic feelings.

In a pilot study (N = 226), participants had learned that the target's concrete economic outcomes (i.e., earning bonus payment) predicted their own gains and losses; this study provided initial evidence that conditioning impacted empathy and gave rise to PIT (see the Supplemental Material). In contrast, Experiments 1a and 1b used images depicting the target feeling positive or negative emotion in multiple distinct ways, linking rewards to an abstract representation of affect.

Experiment 1a manipulated outcome congruence between subjects. Each participant learned about one target whose affect was either congruent or incongruent with the participants' own outcomes. Experiment 1b manipulated outcome congruence within subjects. Each participant learned about two targets; one of them had congruent outcomes with participants, and the other one had incongruent outcomes. All experiments were approved by the University of Southern California Institutional Review Board to ensure adequate protection of participants.

Participants. Participants were recruited on the CloudResearch platform and participated in exchange for payment in both Experiments 1a and 1b. For Experiment 1a, a power analysis indicated that 250 participants would provide 90% power to detect a small to medium effect size (η_p^2 = .02). For Experiment 1b, a power analysis indicated that 350 participants would provide more than 95% power for detecting the effect size observed in Experiment 1a (η_p^2 = .03). We therefore aimed to recruit 300 participants for Experiment 1a and 400 participants for

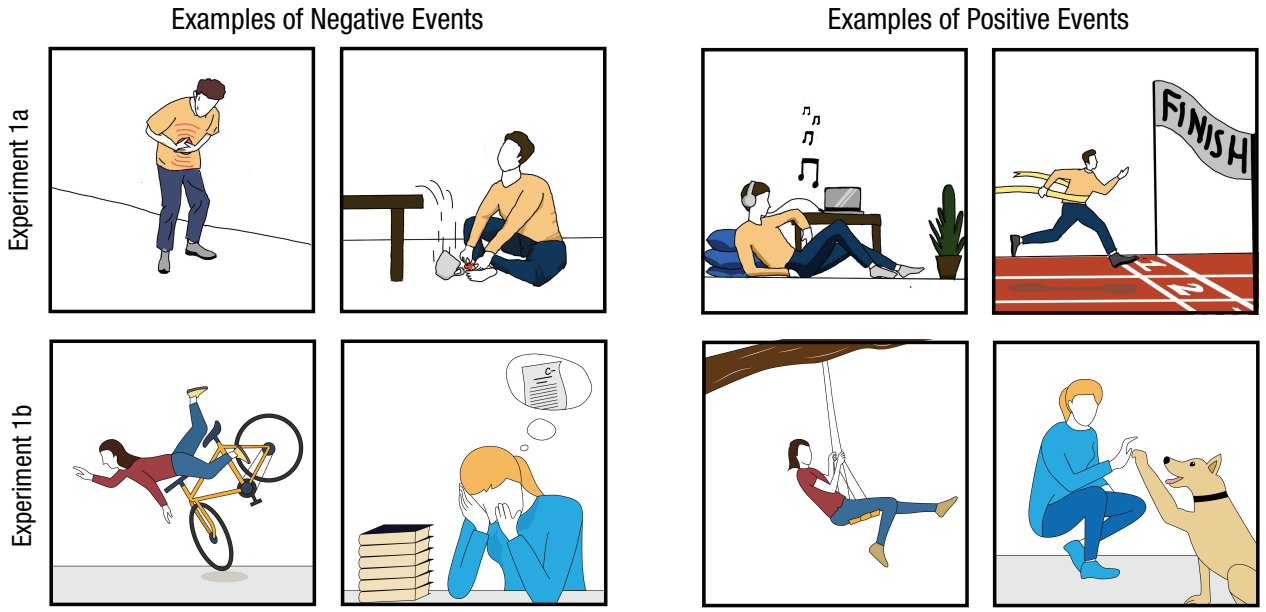

Fig. 1. Examples of the conditioned stimuli used in Experiments 1a and 1b. In Experiment 1a (top row), each image represents a unique life event that happened to a single target, characterized by either positive or negative emotions. Experiment 1b (bottom row) used similar images, with the protagonist being one of two females. In both experiments, the face of the target was intentionally left undrawn so that participants could not learn by relying on concrete facial features (e.g., smile). A complete set of stimuli can be found on OSF.

Experiment 1b to account for potential exclusions and received 299 and 396 responses, respectively. Because of technical errors, Pavlovia data were not saved for five participants in Experiment 1a and 11 participants in Experiment 1b, leaving 294 participants in Experiment 1a and 385 participants in Experiment 1b. For both experiments, power was calculated using G*Power (Faul et al., 2007). Informed consent was obtained from all participants.

To ensure that participants were actively engaged in the task, we asked participants to press the space key every time the US was revealed in the Pavlovian conditioning phase. Following exclusion rules used in prior work, we removed data from any participant who missed 20% or more of the trials (Hackel & Kalkstein, 2023). Additionally, at the end of the study, we asked participants the extent to which they paid attention during the study on a scale from 1 to 7 (1 = *not at all*, 4 = *moderate attention*, 7 = *full attention*) and excluded participants who gave a rating of 4 or lower. These preregistered criteria left 268 participants for analysis in Experiment 1a (M_{age} = 39.9 years, *SD* = 11.0; 127 women, 137 men, two nonbinary, two undisclosed) and 342 participants in Experiment 1b (M_{age} = 41.8 years, *SD* = 12.1; 172 women, 164 men, three nonbinary, three undisclosed).

Stimuli. To examine whether Pavlovian conditioning can influence people's valuation of another person's abstract mental states, we created images depicting unique life events of another person that were either positive (e.g., playing with one's dog) or negative (e.g., falling off a bike; Fig. 1). As a cover story, participants were told that these life events were collected from a previous research subject's descriptions of personal life events, which we then turned into images. To ensure participants would not learn to associate concrete features such as smiles or frowns with reward, the target's face was left undrawn in all images. As a result, no concrete visual features were common across images of the same valence; each image featured a different concrete event that suggested positive or negative affect. Accordingly, the valence of the event and the target's corresponding mental states (i.e., feeling good/bad) served as the CSs. For Experiment 1a, we created 24 images featuring a male target (12 positive scenes and 12 negative scenes). For Experiment 1b, we created 36 images featuring two female targets (nine positive scenes and nine negative scenes for each target).

Procedure. As a cover story that explained to participants why they would see the CSs (i.e., target images) and the USs (i.e., participants' own monetary outcomes) during the experiments, participants were told that the study was about how people memorize different types of information, including social information about another person and personal information about their own outcomes. In particular, they needed to observe and try to memorize images of another person's life events, as well

as how the price of their "stock" changes over time. Participants were told that a higher stock price would result in a higher chance of winning an Amazon gift card in a raffle at the end of the study. Therefore, an increasing stock price would be a rewarding US and a decreasing stock price an aversive US.

Participants first completed a Pavlovian conditioning task programmed in PsychoPy (Version 2021.2.3) and hosted on Pavlovia (Peirce et al., 2019). In this task, participants learned how a target's affect in the images (the CSs) predicted their own monetary outcomes (the USs; Fig. 2). In each round, participants first saw an image depicting a target's life event and were instructed to imagine how the other person felt in the scene. The image was displayed for 5 s. Next, participants observed their own stock outcomes, which either increased or decreased by 50 points. Participants were asked to press the space key to acknowledge seeing it within 3 s. The task moved on after 3 s. If no key was pressed, then participants would see "No response" for 1 s. Each round ended with a 1.5-s intertrial interval, during which a fixation cross was displayed at the center of the screen.

Crucially, we manipulated the congruence between the CSs and the USs. In Experiment 1a, congruence was manipulated between subjects. The conditioning task consisted of 20 trials, with a unique image displayed in each trial. As a result, a random subset of 20 of the 24 images (10 positive, 10 negative) were displayed during conditioning, and the remaining four images (two positive, two negative) were used as novel stimuli during the PIT and empathy rating phases. In the congruent condition, the CSs and USs had the same valence in 90% of the trials (i.e., 18 of 20 trials). That is, positive target scenes predicted increases in participants' stock price 90% of the time, and negative target scenes predicted decreases in participants' stock price 90% of the time. By contrast, in the incongruent condition, the reverse was true; the CSs and USs had the opposite valence in 90% of the trials. As a result, participants in the congruent condition would learn that positive feelings of the target were rewarding, whereas for participants in the incongruent condition, negative feelings of the target were rewarding.

In Experiment 1b, congruence was manipulated within subjects. Each participant learned about two targets simultaneously. For each target, a random subset of 14 images (seven positive, seven negative) was displayed, and four were reserved for the testing phase. Each image was displayed twice, resulting in a total of 56 conditioning trials. One target was randomly selected to be the "congruent target" such that their scenes predicted congruent changes in participants' stock price 93% of the time (i.e., 26 of 28 trials). The other target was the "incongruent target," and their scenes predicted incongruent changes in participants' stock price 93% of the time. Which target was congruent was counterbalanced across participants. In both Experiments 1a and 1b, the order of presentation for the CSs was randomized.

After the conditioning phase, participants completed an adapted PIT task (Allman et al., 2010; Huys et al., 2011), which we pretested in a pilot experiment (see the Supplemental Material). The purpose of the task was to test whether the value associated with the CSs would influence instrumental behavior, providing a classic marker of Pavlovian conditioning; if people associate reward with a cue, then they should show stronger reward-seeking behavior when that cue is visible. First, participants were instructed to press the "L" key to see what happens. After a given number of key presses (randomly selected from 10 to 15), participants saw that their stock price went up. This procedure was repeated five times to help participants learn that pressing "L" was rewarding. Next, participants were instructed to continue pressing "L" without receiving feedback. Meanwhile, a CS (i.e., a novel image featuring the target) was displayed at the center of screen, and participants were told to ignore the CS and focus on pressing the key. After some time (6 s in Experiment 1a and 5 s in Experiment 1b), the screen moved on. In Experiment 1a, participants completed four rounds of key pressing, with two rounds displaying a positive scene and two rounds displaying a negative scene featuring a single target. In Experiment 1b, participants completed eight rounds of the task, which consisted of positive and negative scenes (two for each) for both the congruent and the incongruent target.

After the PIT task, participants were told that their stock price would no longer change. This instruction ensured that any impact of conditioning on empathic feelings would not reflect participants' explicit expectations that the CSs would help them win additional money but instead would reflect participants' intrinsic affective responses to the CSs. Participants then saw each novel CS from the PIT task again and indicated how good and bad these images made them feel on two separate rating scales (from "Not at all" to "Very much"). We used separate scales to capture potentially ambivalent feelings. This approach has been used in past work to measure empathic feelings, with empathic feelings defined as feelings congruent with the target's outcomes (i.e., good feelings toward positive target scenes and bad feelings toward negative target scenes) and counterempathic feelings defined as feelings incongruent with the target's outcomes (Cikara et al., 2014). After the empathy ratings, participants indicated how pleasant or unpleasant they found the target (from

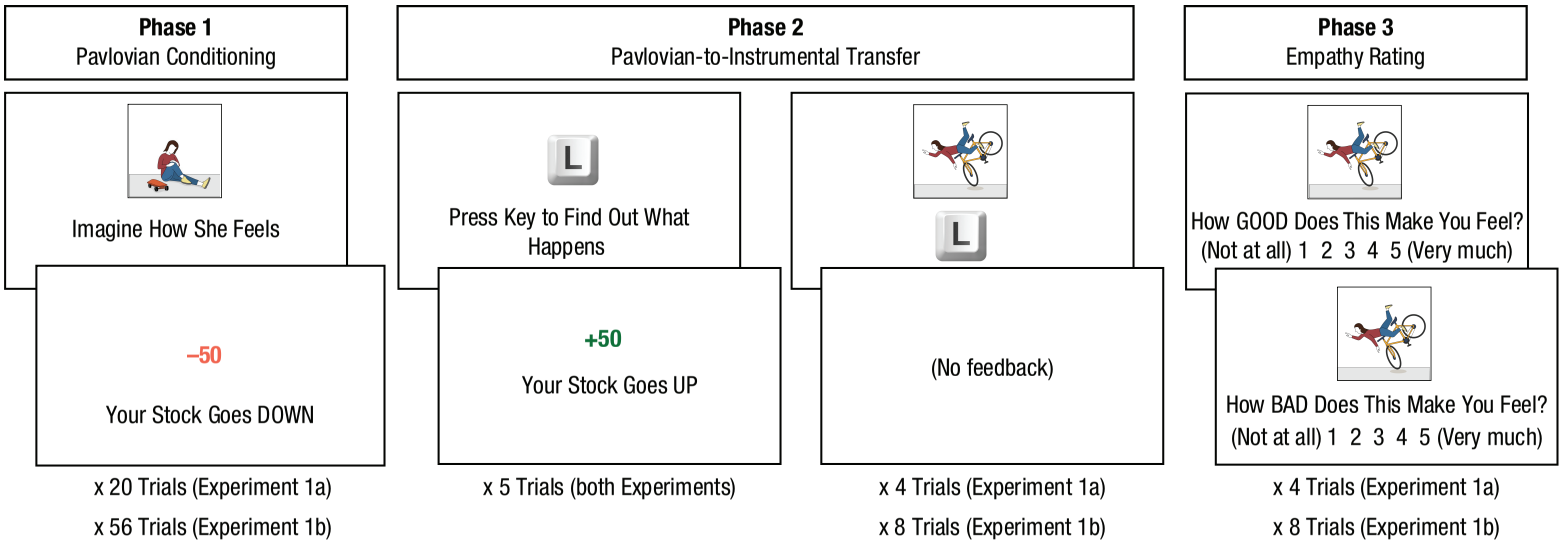

Fig. 2. Schematic of the computer task in Experiments 1a and 1b. Participants first learned the pairing between the conditioned stimuli (target scenes) and the unconditioned stimuli (stock outcomes). Next, they performed an instrumental action (i.e., key press) in the presence of novel conditioned stimuli. Last, they reported their feelings toward the novel conditioned stimuli, which served as our measure of empathy. Experiments 1a and 1b were based on the same paradigm, with different stimuli and minor changes to the instructions and rating scales.

"Very unpleasant" to "Very pleasant"). Each trial moved on when participants made their response. Whereas Experiment 1a used five-point rating scales (1–5) for both questions, Experiment 1b replaced these scales with continuous sliders (0–100) to allow more fine-grained responses.

At the end of the computer task, participants were directed to an online survey hosted on Qualtrics. Given that past research has highlighted a role for contingency awareness in Pavlovian conditioning (Jeffs & Duka, 2017; Lovibond & Shanks, 2002), participants completed three questions measuring their awareness of the contingency during Pavlovian conditioning. First, we asked participants whether they noticed a relationship between the target's images and their own stock outcomes. Second, we asked participants how their stock price would change when the target experienced a positive event and, separately, a negative event. These measures provide a measure of the strength of the perceived contingency (for full details, see the Supplemental Material). Next, we collected the following measures for exploratory purposes: In Experiment 1a, participants rated the extent to which the target experienced various emotions (e.g., anger, excitement) in a random subset of five images; participants also retrospectively reported how they felt when their stock price increased and decreased; in both Experiments 1a and 1b, participants rated the extent to which they considered the target's thoughts and feelings when viewing the images and the extent to which they found the events in the images relatable. Finally, both experiments measured trait-level empathy using the Interpersonal Reactivity Index (IRI; Davis, 1983) and the AQ-10 Autism Spectrum Quotient (Allison et al., 2012). We report exploratory analyses involving these measures in the Supplemental Material.

Analytic procedures. To test whether outcome congruence during conditioning influenced participants' valuation of the target(s)'s mental states and participants' empathy for the target(s), we conducted repeated measures analyses of variance (ANOVAs) using the ez package in R (Lawrence, 2016). In each experiment, we conducted two separate ANOVAs, predicting participants' good and bad feelings toward the novel scenes respectively using (a) outcome congruence during conditioning, (b) scene valence, and (c) their interaction. In Experiment 1a, outcome congruence was a between-subjects predictor and scene valence was a within-subjects predictor, whereas in Experiment 1b, both predictors were at the within-subjects level.

To test whether outcome congruence influenced PIT, we examined participants' numbers of key presses during the PIT phase. After ruling out overdispersion of count data (Bolker et al., 2009; Zeileis et al., 2008), we fitted the numbers of key presses to mixed-effects Poisson regression models with the following predictors: target's outcome congruence (−1 = incongruent, 1 = congruent), scene valence (−1 = negative, 1 = positive), and the interaction between outcome congruence and scene valence. In Experiment 1a, by-subjects random intercept and slope for scene valence were included. In Experiment 1b, random slopes were removed to allow model convergence (Barr et al., 2013). Both models were fitted using the lme4 package in R (Bates et al., 2015).

Notably, while examining the distribution of the number of key presses during the PIT phase, we noticed a few trials with unrealistically high numbers of key presses (e.g., > 10 per second). These trials likely resulted from participants holding down instead of repeatedly pressing the "L" key during a given trial and thus did not meaningfully reflect participants' behavior. Therefore, we detected and excluded outlier trials on the basis of the median absolute deviation (MAD) rule, which is a method that is more robust than detecting outliers using the mean and the standard deviation (Leys et al., 2013). Specifically, we first computed the absolute deviation of each observation from the median and then computed the median of these deviations. Next, we removed key presses that exceeded the median by more than 3 MADs. This led to the exclusion of 2.3% of the trials and three additional participants in Experiment 1a (final $N = 265$) and the exclusion of 3.0% of the trials and four additional participants in Experiment 1b (final $N = 338$). These exclusions deviated from the preregistration. However, the inclusion of these trials did not alter the direction or significance of the results (see Table S2).

Moreover, an examination of the posttask survey responses suggested that about half of the participants (Experiment 1a: $n = 118$, 44.0%; Experiment 1b: $n = 175$, 51.2%) reported not noticing a relationship between the target images and their own outcomes during the conditioning phase. Given that contingency awareness may be important for Pavlovian conditioning (Jeffs & Duka, 2017; Lovibond & Shanks, 2002), we also conducted secondary analyses in both Experiments 1a and 1b, asking whether contingency awareness moderated the empathy and PIT effects by adding participants' binary contingency awareness score (−1 = unaware, 1 = aware) and its interactions with outcome congruence and outcome valence as additional predictors in each model. This analysis was not preregistered in Experiment 1a but was noted in the preregistration for Experiment 1b. We briefly describe the results below and include full results in the Supplemental Material.

Results

Empathy ratings. As hypothesized, outcome congruence during Pavlovian conditioning influenced participants' empathic and counterempathic feelings for the

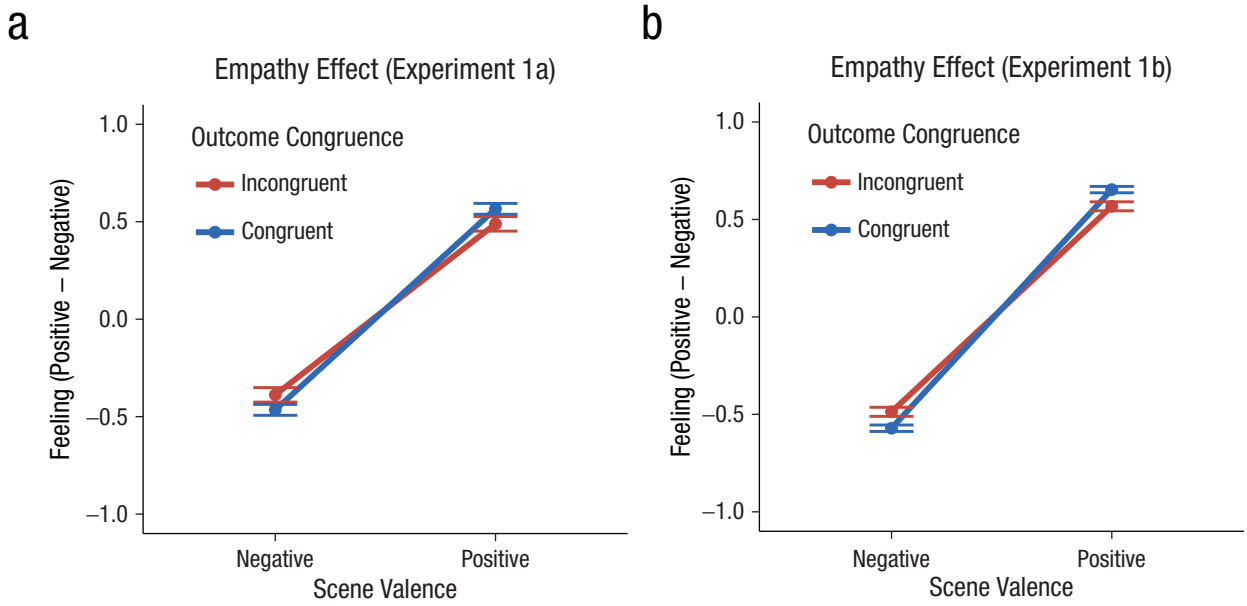

Fig. 3. Empathy ratings as a function of outcome congruence and scene valence in Experiments 1a and 1b. For visualization purposes, a composite feeling score was computed by subtracting participants' bad feelings from good feelings and then standardizing scores on a scale from −1 to 1. A score above 0 represents an overall positive feeling toward the images, whereas a score below 0 represents an overall negative feeling toward the images. Visualizations for the original analyses are included in Figure S4 in the Supplemental Material. The error bars represent ± 1 *SE* of the mean with within-participants adjustment (Morey, 2008).

target(s). In Experiment 1a, scene valence had a significant main effect on both good feelings, $F(1, 266) = 1071.54$, $p < .001$, and bad feelings, $F(1, 266) = 389.48$, $p < .001$, toward the target's scenes. Participants generally felt better about the target's positive than negative events, indicating the general presence of empathy. Importantly, scene valence also interacted with outcome congruence to predict participants' good feelings, $F(1, 266) = 8.05$, $p = .005$, $\eta_p^2 = 0.029$. Participants in the congruent condition reported less good feelings toward negative images and more good feelings toward positive images relative to participants in the incongruent condition. No significant interaction effect was found for bad feelings, $F(1, 266) = 2.04$, $p = .154$, $\eta_p^2 = 0.008$, although the pattern of means reflected the hypothesized direction. These results suggest that Pavlovian conditioning influenced participants' empathy toward the target such that individuals who experienced congruent (vs. incongruent) contingencies during learning reported more good feelings when witnessing novel positive (vs. negative) outcomes for the target (Fig. 3a).

These results from Experiment 1a could reflect a change in empathy specific to the target or a general change in the value of any other person's mental states, or even the abstract notions of "good event" and "bad event." By using a within-subjects design, Experiment 1b allowed us to test whether Pavlovian conditioning shapes empathy in a target-specific way. Indeed, Experiment 1b found patterns similar to those of Experiment 1a across the two targets presented. Outcome congruence interacted with scene valence to predict participants' good feelings, $F(1, 341) = 16.10$, $p < .001$, $\eta_p^2 = 0.045$, as well as bad feelings, $F(1, 341) = 15.56$, $p < .001$, $\eta_p^2 = 0.044$. Participants reported more empathy for the congruent target than the incongruent target—reporting less good feelings and more bad feelings toward the congruent target's negative images and more good feelings and less bad feelings toward their positive images (Fig. 3b).

Pavlovian-to-instrumental transfer. In traditional tests of PIT, people exhibit greater reward-seeking behavior (i.e., faster button presses to gain reward) when a reward-predictive cue is displayed—even if that cue is no longer relevant. We therefore tested whether the values participants associated with the target's mental states influenced subsequent reward-seeking instrumental behavior during the PIT phase, which occurred directly after the conditioning phase and before the empathy phase. Notably, positive images may start out as more intrinsically rewarding than negative ones because of a lifetime of experience people have with empathic feelings. Nonetheless, conditioning could shift any relative benefit of positive over negative images in promoting instrumental behavior. Indeed, in Experiment 1a, outcome congruence interacted with scene valence to predict participants' numbers of key presses, $b = 0.018$, $SE = 0.006$, $z = 2.86$, $p = .004$, 95% confidence interval (CI) = [0.006, 0.030]. Seeing positive

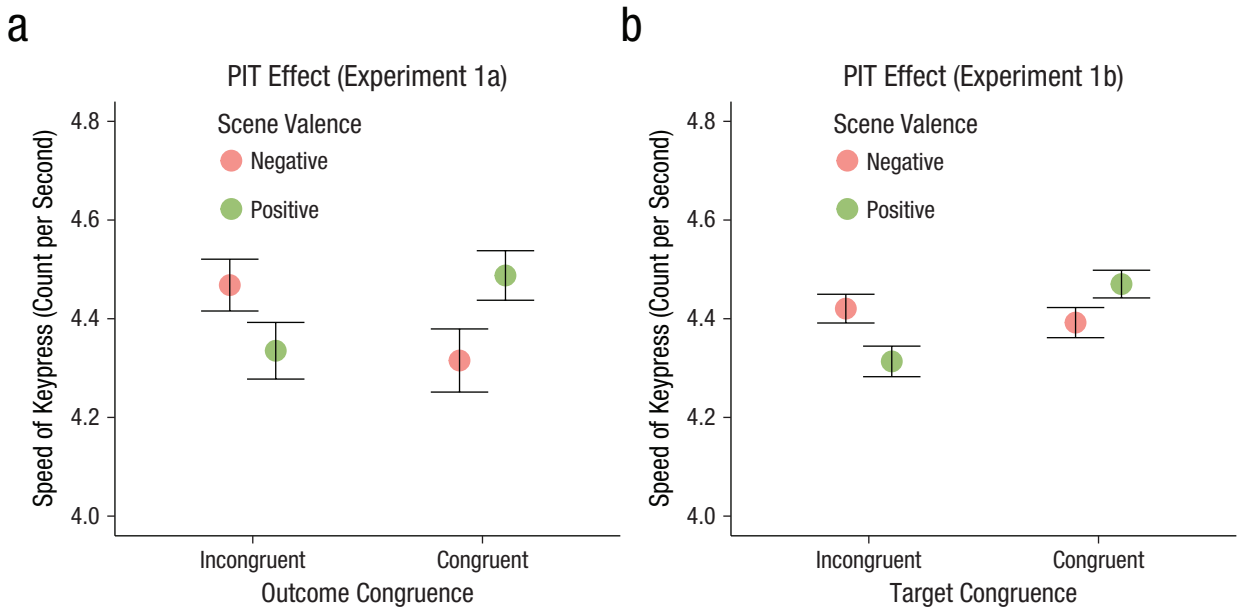

Fig. 4. PIT effects in Experiments 1a and 1b. For visualization purposes, the speed (instead of total numbers) of key presses was used as the dependent variable on the *y*-axis. The error bars represent ± 1 *SE* of the mean with within-participants adjustment (Morey, 2008). PIT = Pavlovian-to-instrumental transfer.

images versus negative images thus had differential impacts on participants' reward-seeking behavior in the congruent versus incongruent conditions. Participants in the congruent condition pressed the "L" key faster when seeing positive versus negative images (simple effect contrast = 0.042, $SE = 0.019$, $z = 2.27$, $p = .023$), whereas pressing speed in the incongruent condition did not significantly differ between positive and negative images (simple effect contrast = −0.030, $SE = 0.019$, $z = -1.55$, $p = .122$; Fig. 4a, Table S1). This result indicates that outcome congruence influenced the rewards participants attached to the mental states of the target, which in turn influenced their instrumental behavior.

Results from Experiment 1b replicated this finding. The interaction between outcome congruence and scene valence was significant, $b = 0.011$, $SE = 0.004$, $z = 2.58$, $p = .010$, 95% CI = [0.003, 0.019], indicating that seeing images of the congruent versus incongruent targets had differential impacts on participants' reward-seeking behavior when viewing positive versus negative scenes. For the incongruent target, participants pressed the key faster for negative images than positive images (simple effect contrast = −0.025, $SE = 0.012$, $z = -2.14$, $p = .033$), whereas for the congruent target, pressing speed did not significantly differ between positive and negative images (simple effect contrast = 0.018, $SE = 0.018$, $z = 1.51$, $p = .131$; Fig. 4b, Table S1). Notably, the PIT effects in Experiments 1a and 1b replicated the PIT effect in the pilot experiment, which used concrete economic stimuli as the USs, indicating that PIT effects were robust across different kinds of outcomes (see Table S2).

Contingency awareness. Next, we tested whether contingency awareness moderated the empathy and PIT effects in Experiments 1a and 1b because past work has suggested contingency awareness may be important for Pavlovian conditioning (Jeffs & Duka, 2017; Lovibond & Shanks, 2002). We observed mixed results across studies and measures. In Experiment 1a, contingency awareness did not significantly moderate the effect of conditioning on either participants' positive empathy ratings, $b = 0.045$, $SE = 0.041$, $t = 1.10$, $p = .271$, 95% CI = [−0.035, 0.124], or negative empathy ratings, $b = -0.081$, $SE = 0.055$, $t = -1.49$, $p = .137$, 95% CI = [−0.19, 0.026], nor did contingency awareness significantly moderate the PIT effect, $b = 0.007$, $SE = 0.006$, $t = 1.09$, $p = .275$, 95% CI = [−0.006, 0.019].

In Experiment 1b, contingency awareness moderated the effect of outcome congruence on both positive and negative empathy ratings—good feelings: $b = 2.24$, $SE = 0.37$, $t = 6.08$, $p < .001$, 95% CI = [1.52, 2.96]; bad feelings: $b = -1.87$, $SE = 0.38$, $t = -4.97$, $p < .001$, 95% CI = [−2.61, −1.13]. Participants who were aware of the contingencies showed a greater empathy gap between the congruent versus incongruent targets compared with those who were unaware. Likewise, contingency awareness moderated the PIT effect, $b = 0.016$, $SE = 0.004$, $z = 3.87$, $p < .001$, 95% CI = [0.008, 0.024], such that the interaction effect of outcome congruence and scene valence on key pressing was stronger among

participants who were aware of the contingencies. These findings should be interpreted with caution, as discussed in the Supplemental Material, because of limitations in the sensitivity of our contingency awareness measure.

Experiment 2: Pavlovian Conditioning Through Mental State Representations

Method

Overview. In Experiments 1a and 1b, participants learned to associate reward with events that would lead a target to feel positive or negative mental states, consistent with the hypothesis that reward is attached to representations of another's feelings. However, an alternative explanation is that participants might have simply categorized the stimuli in the conditioning phase as "positive events" or "negative events," or as events that would make participants themselves feel good or bad, without considering how the target was feeling in the images. Although these abstractions would still offer a root for empathic feelings in associative reward learning, Experiment 2 aimed to more directly test whether people attach reward to mental state representations. To do so, the target's good or bad feelings in Experiment 2 depended on winning gift cards to restaurants they liked or disliked. Which restaurants were liked or disliked was randomized across participants, meaning that events could not be intrinsically categorized by participants as "good" or "bad" in the absence of the target's preferences. Moreover, the target's preferences were not systematically related to participant preferences (see the Supplemental Material). We tested whether participants would nonetheless show PIT and empathy effects in this situation.

In addition, participants in Experiments 1a and 1b responded to the same stimuli during both the PIT and empathy phases; it was therefore possible that participants' actions during PIT may have influenced their feelings when making the empathy ratings through action as input to affect (Cacioppo et al., 1993; Kawakami et al., 2007). In Experiment 2, we used distinct stimuli across the PIT and empathy phases to rule out this possibility (Fig. 5).

Participants. Participants were recruited on CloudResearch in exchange for payment. On the basis of a pilot study, a simulation suggested that 550 participants were required to achieve 80% power for the PIT analysis (Green & MacLeod, 2016). We therefore aimed to recruit 650 participants to account for potential exclusion and received 637 responses. Two participants did not complete the computer task as a result of technical errors, leaving us with a total of 635 participants. Using the same criteria as in Experiments 1a and 1b, we excluded 78 participants, leaving 557 participants for analysis (M_{age} = 38.3 years, SD = 13.1; 269 women, 278 men, 9 nonbinary, 1 undisclosed).

Stimuli. To ensure that the stimuli in the conditioning phase represented the target's mental states, we created illustrations of the target's *arbitrary preferences* for various restaurants by combining restaurant logos with symbols of liking and disliking. We used 32 unique restaurant logos—16 during the conditioning phase, eight during the PIT phase, and eight during the empathy rating phase.

Procedures. As a cover story, we told participants they would learn about a target who had received various restaurant gift cards in a previous study. Before doing so, participants saw gift cards for different restaurants and rated how much they would like each one on a scale from 1 (*strongly dislike*) to 10 (*strongly like*). This step was included to familiarize participants with the scale supposedly seen by the targets. In addition, by measuring participant preferences, this stage allowed us to ensure that participant preferences were unrelated to target preferences on average (see the Supplemental Material).

Next, participants completed the learning phase, which was adapted from the task in Experiments 1a and 1b. Participants were told that they should observe and try to memorize in each round of the task which gift card was sent to the target. In each round, two restaurant logos were displayed on screen, with a number below each logo indicating how the target had supposedly rated that restaurant, using the same scale previously viewed by the target. Of the two restaurants displayed onscreen, the target always liked one restaurant (rating ranging between 8 and 10) and disliked the other (rating ranging between 1 and 3). Three seconds later, one restaurant logo was highlighted with a blue box, indicating the corresponding gift card had been sent to the target. This outcome served as the CSs and was displayed for 5 s. Next, participants saw a change in their own stock price (+50 or −50 points), which served as the USs. Following Experiment 1a, we manipulated the congruence between the CSs and USs between subjects. The CSs and USs had the same (opposite) valence in 94% of the trials (30 of 32) in the congruent (incongruent) condition.

In each round, the liked and disliked restaurants were randomly selected from a list of 16 restaurants. Therefore, on average, each restaurant logo appeared four times. Crucially, which gift cards were liked and which were disliked was randomized across participants. As a result, no features of the outcomes could indicate whether the event was "good" or "bad" beyond

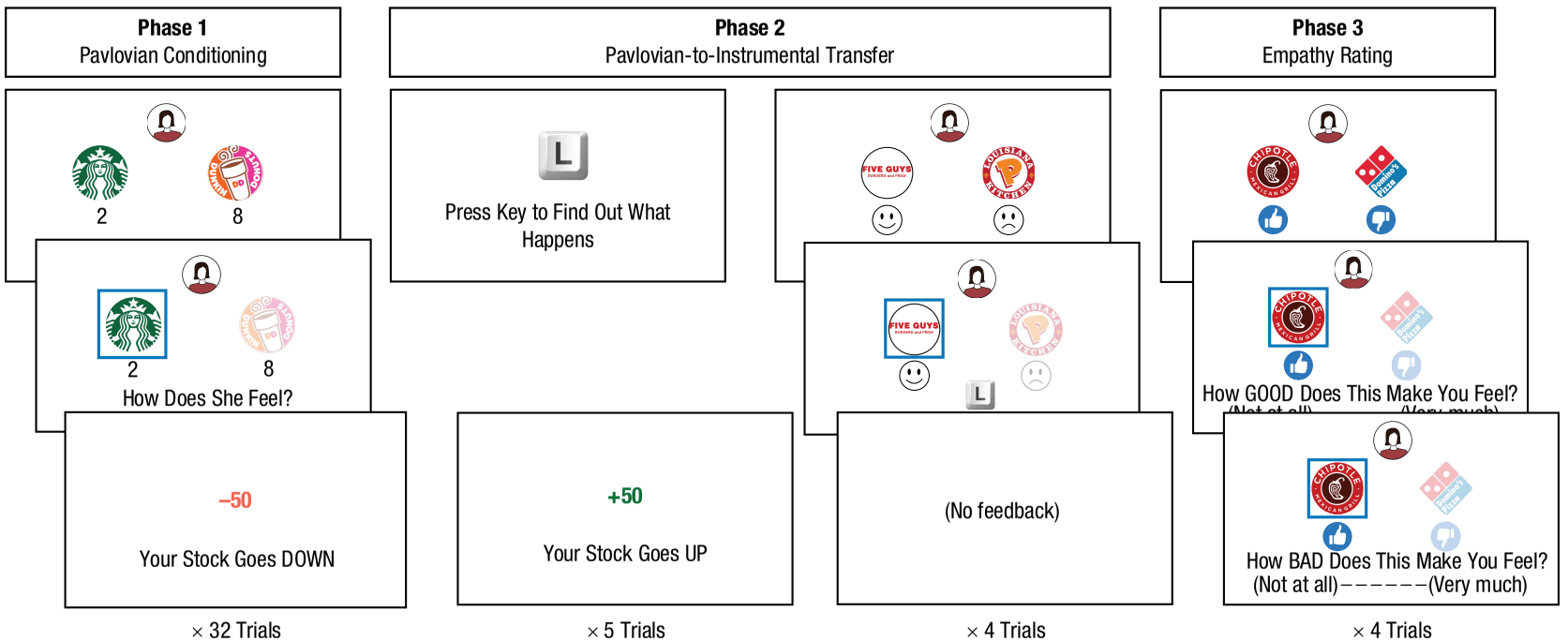

Fig. 5. Schematic of the computer task in Experiment 2. The task structure remained the same as Experiments 1a and 1b. However, the conditioned stimuli were replaced with illustrations of gift cards being sent to the target. The target's preferences were represented by numbers in the conditioning phase, by smiling/frowning faces in the Pavlovian-to-instrumental transfer phase, and by thumbs up/down in the empathy rating phase. In addition, a different set of restaurant logos were used in each phase to ensure participants could not learn and generalize on the basis of concrete cues.

the target's preferences. Similarly, by using numbers to indicate the target's preferences, the stimuli also avoided any intrinsic valence. (For instance, had the numbers reflected rankings, 1 would be the best and 10 would be the worst.) Accordingly, the target's inferred feelings served as the CSs.

The PIT phase followed the same structure as Experiment 1a (instrumental learning followed by a PIT test) except that we replaced the images in the transfer task with illustrations of the target receiving additional gift cards they either liked or disliked. As in the conditioning phase, participants first saw two gift cards—one liked and one disliked—appear for 3 s, after which one of the two was highlighted. After this, participants were cued to press the "L" button to receive the reward. To avoid repeating visual features between conditioning and PIT, participants saw a new set of restaurant logos that had not appeared during conditioning. In addition, we replaced the numbers indicating preferences with smiling and frowning faces to represent the target's liking and disliking for each restaurant, respectively. Accordingly, no concrete visual features were shared between conditioning and PIT such that participants could not simply learn that high or low numbers predict reward. Although smiling and frowning faces have intrinsic valence, they were used only in PIT, not in conditioning; accordingly, participants would respond to these symbols in a manner congruent with their prior learning only to the extent that they had learned to associate reward with the target's preferences during the conditioning stage. The PIT task included four rounds. The target received a liked gift card in two rounds and a disliked gift card in the other rounds. Participants were instructed to continue to attend to the target's outcomes for a supposed later memory test.

After PIT, participants completed the empathy phase. Participants saw additional illustrations of the target's outcomes in the same format as the conditioning and PIT phases. Again, to ensure that no visual features were shared between PIT and empathy ratings, new restaurants were used, and the target's restaurant preferences were now represented using a thumbs up (indicating liking) and a thumbs down (indicating disliking). As soon as a gift card was highlighted, indicating which of the two onscreen the target had received, a slider appeared on the bottom of the screen asking participants how good and bad the outcome made them feel, respectively (0 = *not at all*, 100 = *very much*). Similar to PIT, this task included four rounds: The target received a liked gift card in two rounds and a disliked gift card in the other two.

Together, no common visual features were present across the conditioning, PIT, and empathy phases of the study, and it was therefore only an abstract representation of the target's mental states (liking vs. disliking) that could bind the stimuli together. If conditioning had an impact on participants' PIT and empathy ratings, then it would suggest that participants had attached rewards to the target's mental states.

Analytic procedures. To test whether outcome congruence during conditioning influenced participants' empathy for the target, we conducted the same ANOVAs as in Experiment 1a, predicting participants' good and bad feelings, respectively, on the basis of outcome congruence, outcome valence, and their interaction.

To test whether outcome congruence influenced participants' key-pressing behavior during PIT, we fitted participants' number of key presses to mixed-effects Poisson regressions with the following predictors: outcome congruence (−1 = incongruent, 1 = congruent), outcome valence (−1 = negative, 1 = positive), and their interaction. We included by-subjects random intercept and random slope for outcome valence. As in Experiments 1a and 1b, we excluded outlier trials on the basis of the MAD rule (Leys et al., 2013), which led to the exclusion of 3.5% of the total trials and 10 additional participants (final $N = 547$). However, the inclusion of the outlier trials did not alter the direction or significance of our results (see Table S2).

Last, as in Experiments 1a and 1b, we tested whether contingency awareness moderated the empathy and PIT effects by adding contingency awareness (−1 = unaware, 1 = aware) and its interactions with outcome congruence and outcome valence as additional predictors to each model.

Results

Empathy ratings. As hypothesized, outcome congruence interacted with outcome valence to predict both good feelings, $F(1, 555) = 23.18$, $p < .001$, $\eta_p^2 = 0.040$, and bad feelings, $F(1, 555) = 19.23$, $p < .001$, $\eta_p^2 = 0.033$. Compared with participants in the incongruent condition, participants in the congruent condition reported feeling worse when the target received a disliked gift card and feeling better when the target received a liked gift card (Fig. 6).

Pavlovian-to-instrumental transfer. We found a significant effect of outcome valence on key pressing such that participants responded more vigorously to positive than negative events, $b = 0.030$, $SE = 0.007$, $z = 4.25$, $p < .001$, 95% CI = [0.016, 0.044]. Importantly, replicating Experiments 1a and 1b, we also observed an interaction between outcome valence and conditioning, $b = 0.019$, $SE = 0.007$, $z = 2.67$, $p = .008$, 95% CI = [0.005, 0.032]. Participants in the congruent condition pressed the key

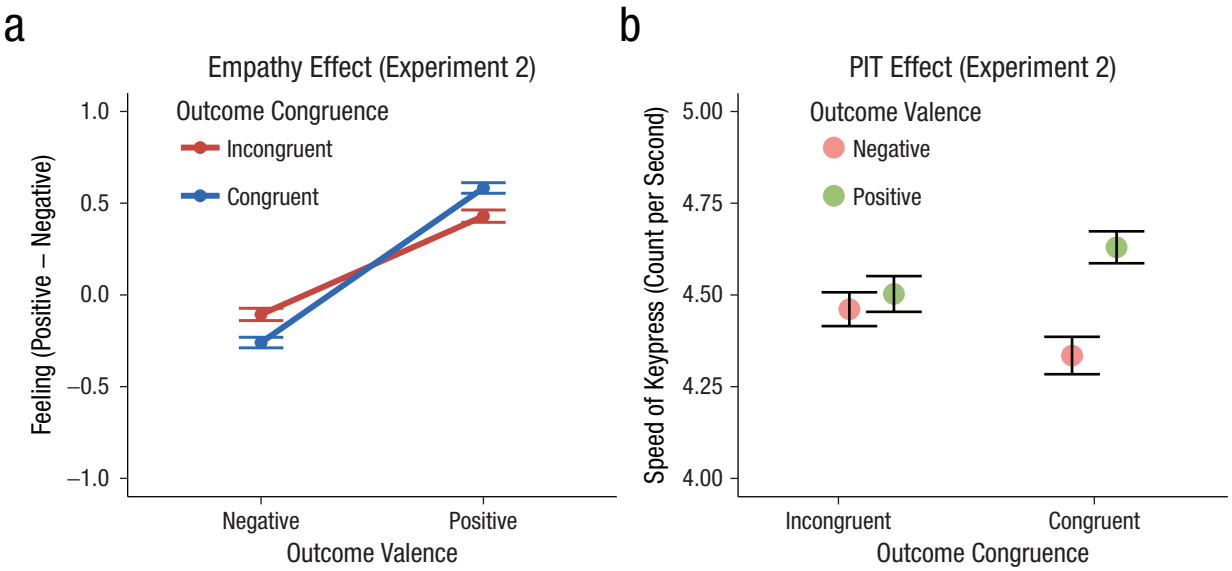

Fig. 6. The empathy effect and PIT effect in Experiment 2. To visualize (a) the empathy effect, a composite feeling score was computed by subtracting participants' bad feelings from good feelings and then standardizing scores on a scale from −1 to 1. A score above 0 represents an overall positive feeling toward the target's outcomes, whereas a score below 0 represents an overall negative feeling toward the target's outcomes. Visualizations for the original analyses are included in Figure S4 in the Supplemental Material. For (b) the PIT effect, the speed (instead of total numbers) of key presses was used as the dependent variable on the *y*-axis. The error bars represent ± 1 *SE* of the mean with within-participants adjustment (Morey, 2008). PIT = Pavlovian-to-instrumental transfer.

faster when seeing positive versus negative outcomes (simple effect contrast = 0.098, *SE* = 0.020, $z = 4.83$, $p < .001$), but this difference was nonsignificant for participants in the incongruent condition, who had learned to associate negative outcomes with reward (simple effect contrast = 0.023, *SE* = 0.020, $z = 1.19$, $p = .233$; Fig. 6). This finding suggests that conditioning influenced the value participants associated with the CSs above and beyond the original valence of the target outcomes. Together, these results replicated Experiments 1a and 1b, showing that Pavlovian conditioning can influence participants' valuation of a target's welfare, even when this process involves inferring the target's mental states.

Contingency. We again observed mixed results when examining contingency awareness. Contingency awareness significantly interacted with outcome congruence and outcome valence to predict both good and bad feelings—good feelings: $b = 3.17$, *SE* = 0.87, $t = 3.64$, $p < .001$, 95% CI = [1.46, 4.88]; bad feelings: $b = -2.78$, *SE* = 0.85, $t = -3.29$, $p = .001$, 95% CI = [−4.43, −1.12]. Specifically, outcome congruence influenced empathy to a greater extent among participants who were aware of the contingencies between target affect and self-reward. However, contingency awareness did not significantly moderate the PIT effect, $b = 0.007$, *SE* = 0.007, $z = 1.04$, $p = .297$, 95% CI = [−0.007, 0.022]. Again, these moderation effects should be interpreted with caution given measurement limitations (see the Supplemental Material).

Experiment 3: Pavlovian Conditioning Influences Prosocial Choice

Method

Overview. Experiments 1a, 1b, and 2 provided evidence that people can learn to associate reward values with the inferred mental states of another person through Pavlovian conditioning, leading them to experience different levels of empathic feelings toward the person in novel scenarios. In Experiment 3, we asked whether these changes in empathy can influence people's subsequent prosocial behavior toward a target. In nonsocial settings, Pavlovian learning can transfer to instrumental decision-making, leading people to choose or avoid stimuli that previously predicted good or bad outcomes (Lindström et al., 2019). For example, after learning that a blue square predicts electric shock through Pavlovian conditioning, people avoid choosing the blue square during later instrumental learning, even when contingencies reverse such that it is no longer optimal to do so. In Experiment 3, we hypothesized that the same would happen for social stimuli such as the inferred mental states of another person. When an individual associates a target's mental state of "feeling good" with reward value, they might be

more willing to act in ways that enhance positive mental states within the target and, as a result, act more prosocially toward the target, even when these actions can bring costs to themselves. Conversely, when one associates a target's negative feelings with reward, they might be more reluctant to act prosocially toward the target, even when doing so would also benefit themselves.

To test this hypothesis, Experiment 3 used a paradigm adapted from Lindström et al. (2019). Participants first learned the predictive value of a target's mental states via the same Pavlovian conditioning task as in Experiment 1a. We again manipulated whether the target's mental states congruently or incongruently predicted participants' monetary outcomes. Next, participants completed a decision-making task in which they saw novel images representing the target's preferences and decided whether to act prosocially toward the target by choosing to send items the target likes or dislikes (Fig. 8). After each choice, participants received a personal reward or loss, allowing instrumental learning. Importantly, for half of the participants, reward contingencies during instrumental learning remained identical to those during conditioning, in which case a Pavlovian bias should lead participants to make more optimal (i.e., self-serving) choices. For the other participants, reward contingencies reversed between Pavlovian conditioning and instrumental learning, in which case a Pavlovian bias should lead participants to make fewer optimal (i.e., self-serving) choices. We therefore asked whether reversing (vs. keeping) the contingencies between the conditioning and instrumental learning phases would influence the choices participants made on behalf of the target.

Participants. On the basis of the power simulation using the simr package in R (Green & MacLeod, 2016), we estimated that 350 participants would yield an 80% power of detecting the effect of an odds ratio of 2 and 99% power of detecting an odds ratio of 3. Therefore, we aimed to recruit 400 participants on CloudResearch to account for potential exclusions and received 390 responses. As a result of technical errors, Pavlovia data were not saved for five participants, leaving 385 complete responses. We further excluded 52 participants on the basis of our preregistered exclusion criteria (i.e., missing at least 20% of the trials during either the Pavlovian conditioning or decision-making phase of the experiment or failing an attention-check question at the end), resulting in 333 participants for analyses (M_{age} = 39.3 years, SD = 11.7; 166 women, 164 men, three nonbinary).

Stimuli. For the Pavlovian conditioning phase, we selected 20 images from the stimuli used in Experiment 1b that featured one female target (10 positive, 10 negative). As in Experiment 1b, participants were told that these images were created on the basis of the life events shared by a previous participant. For the decision-making phase, we created eight new images depicting the target's preferences for eight different restaurants, which were supposedly reported by the target and depicted in images by the experimenters. The target liked four restaurants and disliked the other four (Fig. 7). Participants were told that these images were created on the basis of the previous participant's restaurant preferences, and when participants selected an image, it would make it more likely for the target to get a gift card to the corresponding restaurant depicted in the image.

Task and procedure. The first part of the experiment followed the same procedures as Experiment 1a, in which participants learned the association between images of a target's life events and their own monetary outcomes via Pavlovian conditioning. The conditioning phase consisted of 40 trials, and each image was displayed twice. In the congruent condition, images of positive events predicted an increase in participants' stock price 80% of the time (i.e., 32 of 40 trials), whereas images of negative events predicted a decrease in participants' stock price 80% of the time. The reverse was true for participants in the incongruent condition.

During the decision-making phase, participants were told that the target indicated their preferences for different restaurants in a previous study and that they had an opportunity to send restaurant gift cards to the target by selecting images that reflected the target's preferences. In each trial, participants saw two images on the screen representing how the target would react to two different restaurants. In each trial, there was always one liked restaurant and one disliked restaurant, and the images' position (left vs. right) was counterbalanced. Participants had 5 s to respond by pressing "E" to send a gift card for the restaurant on the left or "I" to send a gift card for the restaurant on the right. Importantly, participants' choices were followed by changes to their stock price in either a congruent or incongruent manner that would continue to influence participants' chance for bonus compensation. For half of the participants, their choice for the target (e.g., sending a gift card that the target liked) predicted congruent changes in their own stock price (e.g., price increase) 75% of the time and predicted incongruent changes in stock price (e.g., price decrease) 25% of the time; for the other half, the opposite was true. As a result, the congruence between target and participant outcomes during the Pavlovian conditioning phase was either preserved or reversed (Fig. 8). Contingency reversal between Pavlovian conditioning and instrumental learning has been used to test whether Pavlovian learning biases later instrumental decision-making, making people more likely to choose options associated with greater reward during

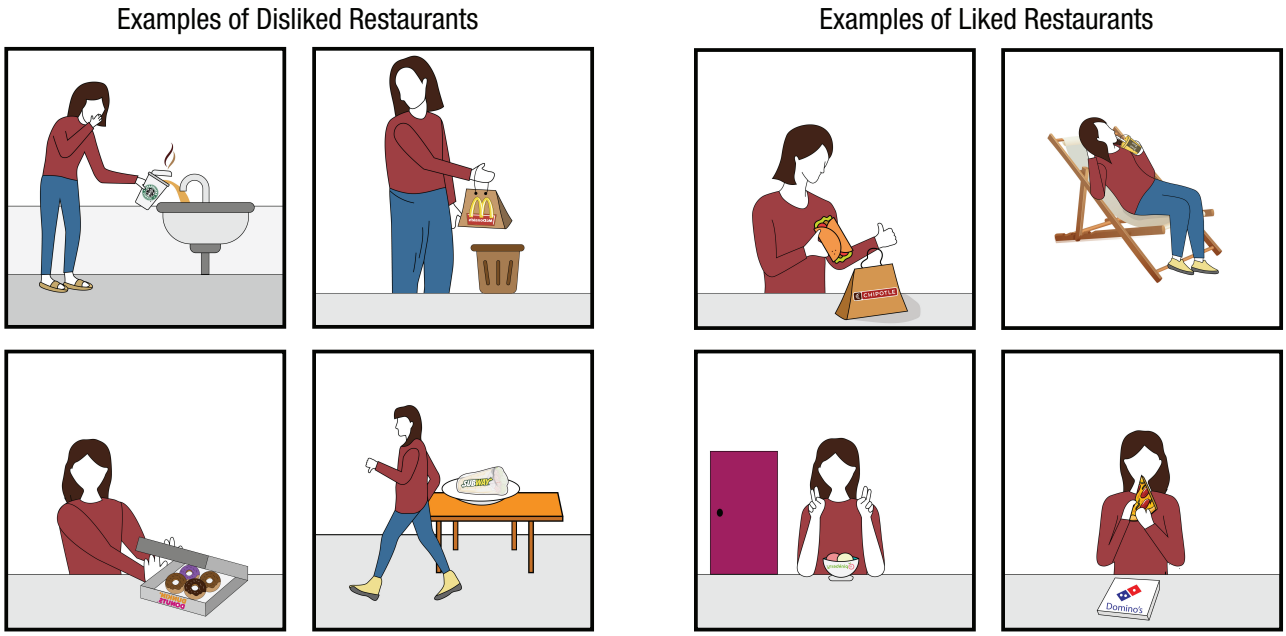

Fig. 7. Stimuli in the decision-making phase of Experiment 3. Each image represents the female target's attitude toward a restaurant (i.e., either liking or disliking).

instrumental choice if those items were previously linked to Pavlovian reward (Lindström et al., 2019).

After the decision-making phase, participants reported how pleasant or unpleasant they found the target using the same scale as Experiment 1b. Participants were then directed to a posttask survey that included the following measures: (a) participants' awareness of the contingency between the CSs and USs during Pavlovian conditioning, (b) the extent to which participants considered the target's thoughts and feelings during conditioning, (c) the extent to which participants found the events in the images relatable, and (d) trait empathy (measured using the IRI). Exploratory analyses with these measures are reported in the Supplemental Material. Notably, participants higher in trait empathic concern were more likely to send gift cards liked by the target, indicating the task indeed reflected prosocial concern for the target (see the Supplemental Material).

Analytic procedures. If Pavlovian learning biases instrumental choice, then participants who attached higher reward to the positive feelings of the target during the conditioning phase should be more likely to continue choosing positive outcomes in the decision-making phase, sending gift cards liked by the target, even when the contingency has been reversed and selecting the positive images no longer led to personal monetary gain. By contrast, when participants attached rewards to the negative images, they should be more likely to send gift cards disliked by the target, even when it no longer offered personal gain. Accordingly, we predicted that regardless of outcome congruence during conditioning, the reversal of contingency between the conditioning and decision-making phases would lead participants to make less personally rewarding choices (i.e., less "optimal") instrumental choices during the decision-making phase (Lindström et al., 2019). To test this question, we recoded participants' choices during the decision-making phase to reflect whether they predicted monetary gain for the self based on instrumental reward contingencies (e.g., sending a disliked gift card when it had a high likelihood of incurring monetary gain would be coded as an optimal choice). We then fitted a mixed-effects logistic regression model predicting participants' choices in the decision-making phase (0 = nonoptimal, 1 = optimal) using contingency-reversal status (−1 = no change, 1 = reversal). We included a by-subjects random intercept because the trials were nested within subjects.

In addition, we tested as an exploratory analysis whether contingency reversal influenced participants' reaction time for choosing which gift card to send. Accordingly, we fitted a mixed-effects logistic regression predicting participants' reaction time in each trial (log transformed) using contingency-reversal status, and we included a by-subjects random intercept for the predictor.

Notably, a few trials contained unrealistically fast reaction times (e.g., < 200 ms). These trials likely resulted from participants holding down or pressing the buttons without processing the gift card options on the screen and, like the outlier trials during the PIT phase of Experiments 1a, 1b, and 2, did not

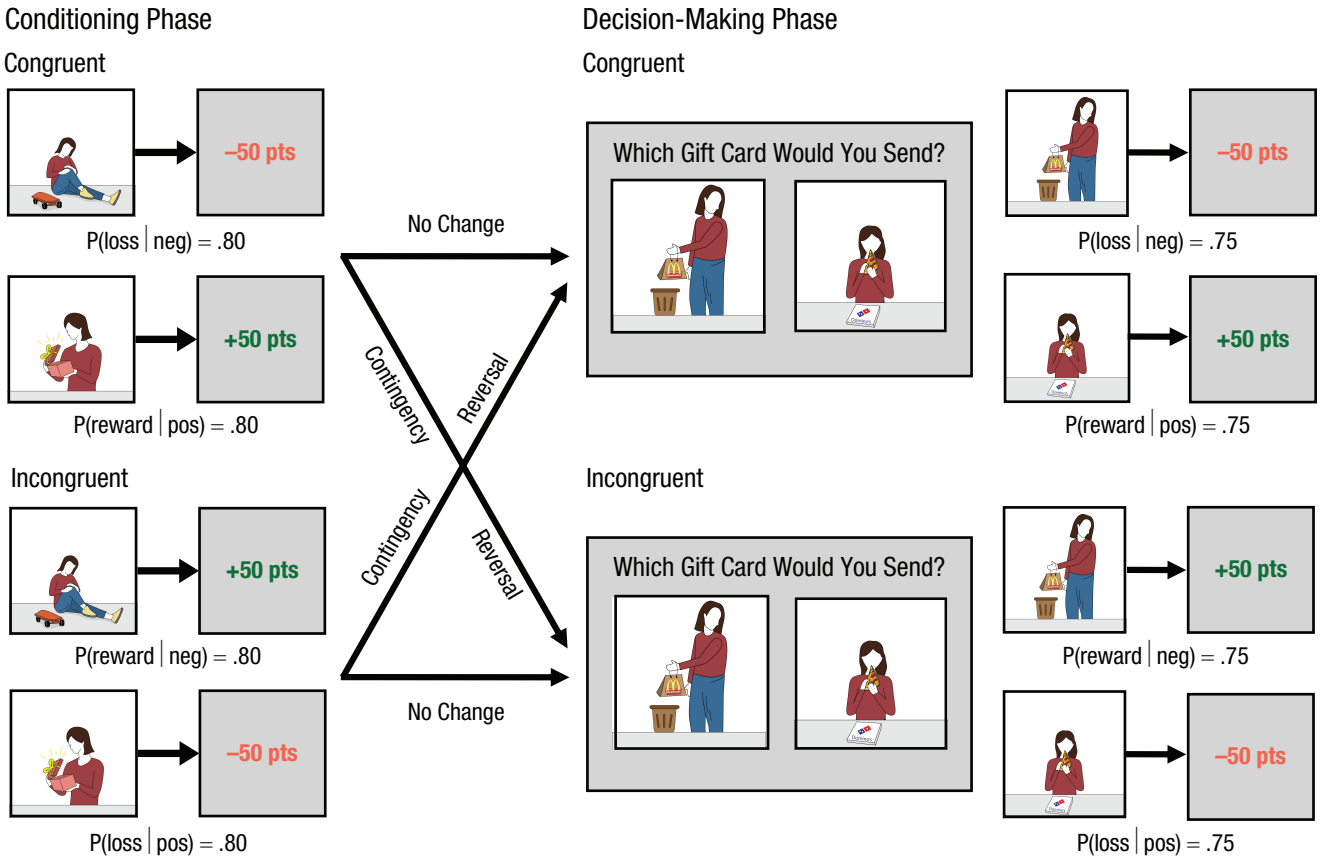

Fig. 8. Schematic of Experiment 3. In the conditioning phase, participants learned the association between images of a target's emotional events and their own monetary outcomes, which was either congruent or incongruent. Next, participants made 32 decisions between two novel images at a time, representing the target's preference for two restaurants. Participants were told that their choice would increase the target's chance of receiving a gift card for the corresponding restaurant. Each choice was probabilistically followed by either monetary reward or loss for the participant. For half of the participants, giving a liked gift card led to monetary gain and giving a disliked gift card led to monetary loss, whereas for the other half, the reverse was true. As a result, for half of the participants, the congruence between target and self outcomes remained the same as in initial conditioning ("no-change" group), and for the other half the congruence was reversed ("contingency-reversal" group).

meaningfully reflect participants' behavior. Therefore, we excluded from analyses trials with a reaction time of 200 ms or faster following the criterion used in past work (Semmelmann & Weigelt, 2017). This led to the exclusion of 1.0% of all trials (106 of 10,532). Given that we did not describe this analysis in detail in the preregistration, we consider this analysis exploratory and report full results in the Supplemental Material. The inclusion of the outlier trials did not alter the direction or significance of most analyses; however, it rendered the effect of contingency reversal on reaction time non-significant (see the Supplemental Material).

Results

Probability of optimal choices. We first asked whether participants' likelihood of making optimal choices during the decision-making phase depended on whether contingency was reversed or preserved between the conditioning and decision-making phases. Contrary to our prediction, contingency reversal did not have a significant main effect on the probability of optimal choice, $b = -0.19$, $SE = 0.13$, $z = -1.51$, $p = .131$, 95% CI = [–0.44, 0.06] (Fig. 9a).

However, past work suggests Pavlovian conditioning may depend on the extent to which participants form awareness of the contingency between the CSs and USs (Jeffs & Duka, 2017; Lovibond & Shanks, 2002). An examination of the posttask survey responses suggested that more than half of the participants ($n = 187$, 56.1%) reported that they did not notice a relationship between the target's outcomes and their own outcomes during the conditioning phase. We therefore conducted a secondary analysis specified as an exploratory test in our preregistration that examined whether contingency awareness (–1 = unaware, 1 = aware) moderated the effect of Pavlovian conditioning. This analysis revealed a significant interaction between contingency reversal

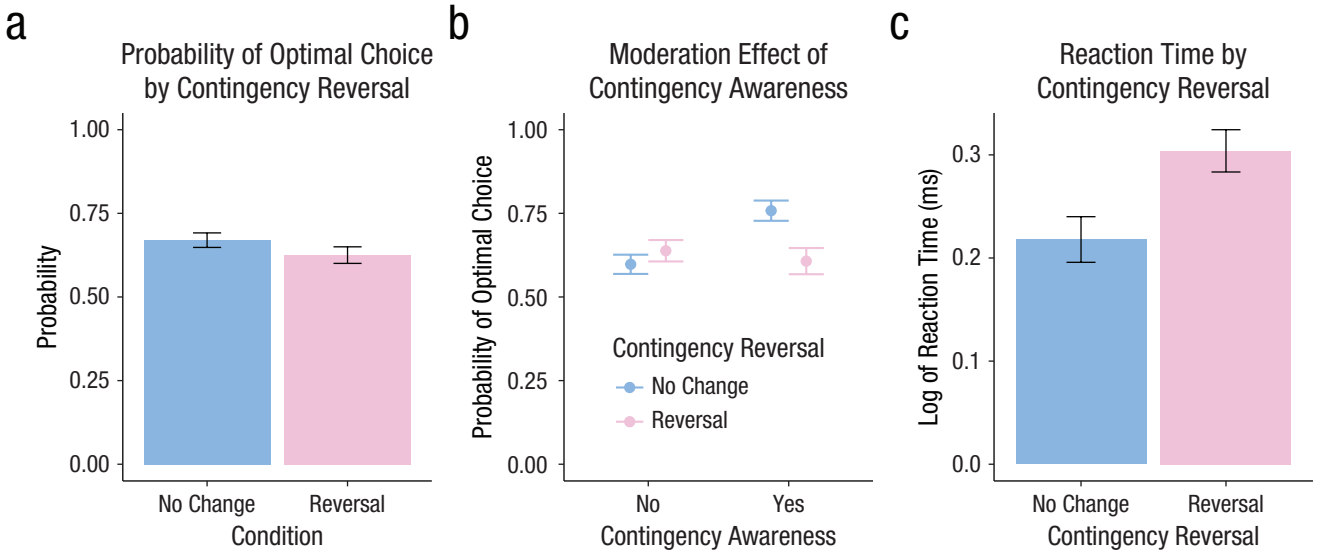

Fig. 9. Choices during the decision-making phase as a function of contingency reversal. The graphs show (a) the probability of selecting the optimal action (i.e., stimulus leading to the highest expected reward for oneself) during the decision-making phase for preserved- versus reversed-contingency participants, (b) the moderation effect of contingency awareness during Pavlovian conditioning, and (c) the log-transformed reaction time during the decision-making phase as a function of contingency reversal. Error bars represent ± 1 *SE* of the mean with within-participants adjustment (Morey, 2008).

and contingency awareness in predicting optimal choice, $b = -0.38$, $SE = 0.12$, $z = -3.05$, $p = .002$, 95% CI = [−0.62, −0.14]. For participants who were aware of the contingency during Pavlovian conditioning, reversal of the contingency during decision-making impaired optimal choice (simple effect contrast = −1.24, $SE = 0.38$, $z = -3.29$, $p = .001$). However, for participants who were unaware of the contingency, contingency reversal did not have a significant effect on optimal choice (simple effect contrast = 0.28, $SE = 0.33$, $z = 0.87$, $p = .385$; Fig. 9b, Table S3). Pavlovian values associated with mental states thus biased later instrumental decision-making among participants who strongly encoded the Pavlovian associations during learning.

Reaction time. Next, we conducted another secondary analysis specified in our preregistration that tested whether contingency reversal influenced participants' reaction time during the decision-making phase. Past work suggests that people are generally faster when they make choices with high expected value and slower when they make choices with lower expected value (Krajbich et al., 2015). We therefore hypothesized that a reversal of contingencies would lead to slower decisions given that Pavlovian values conflicted with instrumental values. In this manner, beyond shaping which choices people make, Pavlovian conditioning might influence the readiness with which people make prosocial choices. Indeed, reversal (vs. preservation) of the contingency led participants to make slower choices, $b = 0.043$, $SE = 0.015$, $t = 2.86$, $p = .005$, 95% CI = [0.014, 0.073] (Fig. 9c, Table S4). When the contingency was reversed, participants were less ready to select a choice. However, this effect did not significantly depend on whether participants were aware of the contingency, $b = -0.013$, $SE = 0.015$, $t = -0.87$, $p = .385$, 95% CI = [−0.043, 0.017] (Table S5).

Together, the findings of Experiment 3 provide further evidence that people can associate value with the inferred mental states of another person and that this process can influence their subsequent decision-making when making prosocial choices on behalf of the target. As with prior studies, the moderation effects of contingency awareness should be interpreted with caution given measurement limitations (see the Supplemental Material).

General Discussion

Empathy and the valuation of others' welfare are important sources of moral behavior (Mobbs et al., 2009; Zaki, 2018; Zaki & Mitchell, 2011). How do people come to form affective responses to the abstract welfare of others? Our findings show that this phenomenon can arise from Pavlovian conditioning of a target's affect with reward. In four experiments, when depictions of a target's positive affect reliably predicted participants' own reward, participants reported more empathic feelings toward the target in novel scenarios, even though no additional reward was available to participants. These findings held true when categorizing the target's

outcomes as good or bad required inferring the target's preferences, suggesting that people can learn empathy by attaching reward to others' abstract mental states.

In addition, when seeing the target in affective states that had been previously rewarding, participants also worked harder to obtain reward for themselves—a classic marker of Pavlovian learning. This behavioral evidence extends beyond self-report, further suggesting that Pavlovian conditioning influenced participants' reward association with the target's welfare. Given that the PIT measure was ostensibly unrelated to empathy and the hypotheses of the PIT task were more difficult to discern, this finding also reduces the likelihood of demand characteristics influencing behavior; at the same time, given that the PIT outcomes were ultimately under participant control, future work could further address this possibility using more implicit measures of affect (Payne & Lundberg, 2014).

Finally, this learning also influenced participants' prosocial choices toward the target in novel scenarios: Stimuli associated with Pavlovian reward not only promote instrumental reward seeking in general but also can motivate people to choose those Pavlovian cues themselves, even when doing so is personally suboptimal (Lindström et al., 2019). Accordingly, we found that when the Pavlovian and instrumental values of a target's mental states differed, participants who were aware of the Pavlovian contingency made fewer choices that were personally optimal in light of instrumental contingencies present during prosocial decision-making. Altogether, reward association with mental states systematically influenced participants' empathy and prosocial choices.

These findings illuminate how simple learning mechanisms can shape complex social behavior. Basic principles of associative learning—including extinction, cue competition, and stimulus generalization—can influence social preferences, such as learning which individuals to trust or which moral acts to perform (Crockett, 2013; FeldmanHall & Dunsmoor, 2019). Yet associative learning can also lead people to mentally link positive or aversive outcomes to abstract categories (e.g., "mammals") and social roles (e.g., "helpers"), helping people generalize learning to novel situations (Dunsmoor & Murphy, 2015; Hackel & Kalkstein, 2023). The current findings indicate that abstract learning can likewise occur for affective mental state concepts: People intuitively infer others' positive or negative affect and can associate these abstract representations with reward.

In this manner, associative learning can give rise to generalizable empathy. Past research shows that associative learning can give rise to empathy by linking concrete outcomes to reward, which can explain the development of empathy in children and the emergence of counter-empathy in adults (Berger, 1962; Englis et al., 1982; Hoffman, 2008). Our findings extend these works by showing that reward can also be attached to abstract representations of others' well-being—namely positive and negative affect. Because these affective states can represent others' "general welfare" across distinct situations, this learning mechanism can lead to generalizable empathic feelings and prosocial preferences in situations novel to people's learning history.

Similar learning processes might contribute to intergroup empathy biases. People often empathize more with in-group than out-group members (Bruneau et al., 2017; Cikara et al., 2014), partly motivated by explicit goals to empathize with teammates over opponents (Weisz & Cikara, 2021; Zaki, 2014). Our findings support a complementary possibility: Members of the same group often share interdependent outcomes, and over time, people might learn that in-group gains predict their own rewards whereas out-group gains predict their own losses (Cikara, 2018, 2021). These differential values may contribute to intergroup empathy gaps alongside goal-directed processes related to cooperation or competition.

Although the current work focused on Pavlovian conditioning, other kinds of associative learning can also shape empathy. First, observational learning can produce similar effects on feelings and decision-making as direct conditioning (Olsson et al., 2007). For instance, seeing someone else empathizing with a target increases one's own empathy toward the target (Zhou et al., 2024) and can fine-tune moral concern (Kleiman-Weiner et al., 2017). Second, through instrumental learning, empathizers can tune their empathic responses on the basis of feedback from the target of empathy (Shamay-Tsoory & Hertz, 2022). Future work should explore whether observational and instrumental learning can attach reward to abstract mental states, providing additional routes to generalizable empathy. Additionally, in evaluative conditioning, positive attitudes can transfer from a positively valenced US to a neutral CS (Walther et al., 2005), which may have contributed to the current findings. That said, evaluative conditioning is argued to produce semantic valence associations, whereas Pavlovian conditioning produces more direct affective associations and reward expectancies (Amodio, 2019), suggesting the current findings may specifically depend on Pavlovian conditioning. Future research can further characterize and dissociate the roles of these forms of learning in empathy.

Notably, although our research supports an influence of Pavlovian conditioning on empathy, the effect sizes were relatively small. Across Experiments 1a, 1b, and 2, most variance in participants' empathic feelings was explained by the valence of the target's outcomes (η_p^2 = 0.59–0.81) rather than the interaction of valence with outcome congruence (η_p^2 = 0.008–0.045). One likely

reason is that our participants, as adults, have developed a generalized valuation of others' welfare through a lifetime of experiences and observations, leading them to generally report emotions congruent with others' outcomes and to choose prosocial options. Accordingly, the room for shifting these empathic feelings and prosocial choices within a short lab experiment may be quite limited. Additionally, the clear-cut and unambiguous emotions in the CSs (e.g., falling off a bike) may have enhanced participants' baseline empathy, further reducing our effect sizes.

Another possible reason for the small effects is the incidental nature of the contingencies between target affect and participant reward. In our paradigm, participants learned how a stranger's life events related to their "stock values" in a computer task. By contrast, reward contingencies in real life, such as during cooperation or competition, often involve clearer causal links and higher stakes, which may lead to stronger conditioning effects. That said, meaningful contingencies could influence empathy even in the absence of reward learning because of competitive motives (Cikara et al., 2014), empathic goals (Cameron et al., 2022; Zaki, 2014), and social norms (Nook et al., 2016). We therefore used a Pavlovian conditioning paradigm in a controlled environment featuring no competition to test whether learning even under these minimal conditions can impact empathy. Despite the small effects, the current results highlight Pavlovian conditioning as one process that can align one individual's reward experiences with the rewards of another. Although our task was somewhat artificial, it complements recent work showing that affective congruence in daily life is associated with increased empathy (Ringwald et al., 2025) and presents a potential learning mechanism underlying this phenomenon. Together, this work offers an initial step toward understanding how learning shapes empathy and more complex social dynamics in naturalistic contexts.

Finally, boundary conditions may have shaped the generalizability of our effects. In our experiments, stimuli depicted familiar concepts, allowing participants to easily recognize abstract mental states and tag them with reward. However, this learning may not occur as easily for more complex scenarios that require goal-directed reasoning to identify abstract relationships (Hackel & Kalkstein, 2023)—a possibility future research can examine. Moreover, how much people generalize learning may depend on the diversity of CSs: In daily life, people may experience distinct contingencies across individuals, groups, or social contexts and learn to empathize in a more context-dependent manner. Last, our reliance on CloudResearch's online convenience samples may limit the generalization of the current findings.

In sum, we found that people can associate the positive or negative feelings of others with reward and loss, which in turn influences their empathic feelings and choices impacting others in novel situations. These findings inform a route through which people come to value the general welfare of others, a process fundamental to empathy and moral decision-making. Given that machine learning often uses reward learning algorithms, the current findings may also guide how artificial intelligence forms humanlike moral intuitions (Leshinskaya et al., 2023). More broadly, these findings highlight how learning mechanisms can inform our understanding of human empathy and morality, identifying how simple reward association can produce abstract and generalizable moral preferences.

Transparency

Action Editor: Tom Beckers
Editor: Simine Vazire

Author Contributions

Yi Zhang: Conceptualization; Data curation; Formal analysis; Methodology; Project administration; Visualization; Writing – original draft.

Leor Hackel: Conceptualization; Methodology; Project administration; Resources; Validation; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research received no specific funding.

Artificial Intelligence

No AI-assisted technologies were used in this research or the creation of this article.

Ethics

This research received approval from the University of Southern California Institutional Review Board (ID: UP-19-00404).

Open Practices

Experiment 1a disclosures. Preregistration: The hypotheses, method, and analysis plans were preregistered (https://aspredicted.org/hv4g2.pdf) on December 9, 2022, prior to data collection, which began on December 12, 2022. There were deviations from the preregistration (for details, see the Method section for Experiment 1a below and Table S13 in the Supplemental Material available online). Materials: All study materials are publicly available (https://osf.io/58wsu). Data: All primary data are publicly available (https://osf.io/58wsu). Analysis scripts: All analysis scripts are publicly available (https://osf.io/58wsu). Computational reproducibility: The computational reproducibility of the results has been independently confirmed by the journal's STAR team. Experiment 1b disclosures. Preregistration: The hypotheses, method, and analysis plans were preregistered (https://aspredicted.org/9wi5x.pdf) on July 17, 2023, prior to data collection, which began

on August 7, 2023. There were deviations from the preregistration (for details, see the Method section for Experiment 1b and Table S13). Materials: All study materials are publicly available (https://osf.io/58wsu). Data: All primary data are publicly available (https://osf.io/58wsu). Analysis scripts: All analysis scripts are publicly available (https://osf.io/58wsu). Computational reproducibility: The computational reproducibility of the results has been independently confirmed by the journal's STAR team. Experiment 2 disclosures. Preregistration: The hypotheses, method, and analysis plans were preregistered (https://aspredicted.org/bmst-6n2z.pdf) on October 31, 2024, prior to data collection, which began on November 2, 2024. There were deviations from the preregistration (for details, see the Method section for Experiment 3 and Table S13). Materials: All study materials are publicly available (https://osf.io/58wsu). Data: All primary data are publicly available (https://osf.io/58wsu). Analysis scripts: All analysis scripts are publicly available (https://osf.io/58wsu). Computational reproducibility: The computational reproducibility of the results has been independently confirmed by the journal's STAR team. Experiment 3 disclosures. Preregistration: The hypotheses, method, and analysis plans were preregistered (https://aspredicted.org/zb4cs.pdf) on December 3, 2023

Because the current experiment used a novel paradigm, we checked data from a small batch of participants (n = 50) to ensure they understood the task and followed the instructions. All remaining data were collected beginning on December 5, 2023, after the preregistration, and no further analyses were conducted before data collection was completed. In Section 8 of the preregistration, we explain why we still consider this to be a valid preregistration. There were deviations from the preregistration (for details, see Table S13). Materials: All study materials are publicly available (https://osf.io/58wsu). Data: All primary data are publicly available (https://osf.io/58wsu). Analysis scripts: All analysis scripts are publicly available (https://osf.io/58wsu). Computational reproducibility: The computational reproducibility of the results has been independently confirmed by the journal's STAR team.

ORCID iD

Yi Zhang https://orcid.org/0009-0002-1737-4044

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976251351304

References

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "red flags" for autism screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(2), 202–212.

Allman, M. J., DeLeon, I. G., Cataldo, M. F., Holland, P. C., & Johnson, A. W. (2010). Learning processes affecting human decision making: An assessment of reinforcer-selective Pavlovian-to-instrumental transfer following reinforcer devaluation. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(3), 402–408.

Amodio, D. M. (2019). Social cognition 2.0: An interactive memory systems account. *Trends in Cognitive Sciences*, *23*(1), 21–33.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., Krivitsky, P. N., Tanaka, E., Jagan, M., & Bolker, M. B. (2015). *Package 'lme4'* (Version 1.1-37) [Computer software]. CRAN. https://cran.r-project.org/web/packages/lme4/lme4.pdf

Berger, S. M. (1962). Conditioning through vicarious instigation. *Psychological Review*, *69*(5), 450–466.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135.

Bruneau, E. G., Cikara, M., & Saxe, R. (2017). Parochial empathy predicts reduced altruism and the endorsement of passive harm. *Social Psychological and Personality Science*, *8*(8), 934–942.

Cacioppo, J. T., Priester, J. R., & Berntson, G. G. (1993). Rudimentary determinants of attitudes: II. Arm flexion and extension have differential effects on attitudes. *Journal of Personality and Social Psychology*, *65*(1), 5–17.

Cameron, C. D., Scheffer, J. A., Hadjiandreou, E., & Anderson, S. (2022). Motivated empathic choices. *Advances in Experimental Social Psychology*, *66*, 191–279.

Cartoni, E., Balleine, B., & Baldassarre, G. (2016). Appetitive Pavlovian-instrumental transfer: A review. *Neuroscience & Biobehavioral Reviews*, *71*, 829–848.

Cikara, M. (2018). Pleasure in response to outgroup pain as a motivator of intergroup aggression. In K. Gray & J. Graham (Eds.), *Atlas of moral psychology* (pp. 193–200). Guilford Press.

Cikara, M. (2021). Causes and consequences of coalitional cognition. In B. Gawronski (Eds.), *Advances in experimental social psychology* (Vol. 64, pp. 65–128). Academic Press.

Cikara, M., Bruneau, E., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of Experimental Social Psychology*, *55*, 110–125.

Contreras-Huerta, L. S., Coll, M. P., Bird, G., Yu, H., Prosser, A., Lockwood, P. L., Murphy, J., Crockett, M. J., & Apps, M. A. (2023). Neural representations of vicarious rewards are linked to interoception and prosocial behaviour. *NeuroImage*, *269*, Article 119881. https://doi.org/10.1016/j.neuroimage.2023.119881

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126.

Dunsmoor, J. E., & Murphy, G. L. (2015). Categories, concepts, and conditioning: How humans generalize fear. *Trends in Cognitive Sciences*, *19*(2), 73–77.

Englis, B. G., Vaughan, K. B., & Lanzetta, J. T. (1982). Conditioning of counter-empathetic emotional responses. *Journal of Experimental Social Psychology*, *18*(4), 375–391.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

FeldmanHall, O., & Dunsmoor, J. E. (2019). Viewing adaptive social choice through the lens of associative learning. *Perspectives on Psychological Science*, *14*(2), 175–196.

Gęsiarz, F., & Crockett, M. J. (2015). Goal-directed, habitual and Pavlovian prosocial behavior. *Frontiers in Behavioral Neuroscience*, *9*, Article 135. https://doi.org/10.3389/fnbeh.2015.00135

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498.

Hackel, L. M., & Kalkstein, D. A. (2023). Social concepts simplify complex reinforcement learning. *Psychological Science*, *34*(9), 968–983.

Hackel, L. M., Kalkstein, D. A., & Mende-Siedlecki, P. (2024). Simplifying social learning. *Trends in Cognitive Sciences*, *28*(5), 428–440. https://doi.org/10.1016/j.tics.2024.01.004

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259.

Hoffman, M. L. (1985). Interaction of affect and cognition in empathy. In C. E. Izard, J. Kagan, & R. B. Zajonc (Eds.), *Emotions, cognition, and behavior* (pp. 103–131). Cambridge University Press.

Hoffman, M. L. (2008). Empathy and prosocial behavior. *Handbook of Emotions*, *3*, 440–455.

Huys, Q. J., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. *PLOS Computational Biology*, *7*(4), Article e1002028. https://doi.org/10.1371/journal.pcbi.1002028

Jeffs, S., & Duka, T. (2017). Predictive but not emotional value of Pavlovian stimuli leads to pavlovian-to-instrumental transfer. *Behavioural Brain Research*, *321*, 214–222.

Kawakami, K., Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the heart grow fonder: Improving implicit racial attitudes and interracial interactions through approach behaviors. *Journal of Personality and Social Psychology*, *92*(6), 957–971.

Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition*, *167*, 107–123.

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications*, *6*(1), Article 7455. https://doi.org/10.1038/ncomms8455

Lawrence, M. A. (2016). *ez: Easy analysis and visualization of factorial experiments* (Version 4.4-0) [Computer software]. CRAN. https://cran.r-project.org/web/packages/ez/index.html

Leshinskaya, A., & Chakroff, A. (2023, December 10–16). *Value as semantics: Representations of human moral and hedonic value in large language models* [Paper presentation]. 37th Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans, LA, United States.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766.

Lindström, B., Golkar, A., Jangard, S., Tobler, P. N., & Olsson, A. (2019). Social threat learning transfers to decision making in humans. *Proceedings of the National Academy of Sciences, USA*, *116*(10), 4732–4737.

Lockwood, P. L., Hamonet, M., Zhang, S. H., Ratnavel, A., Salmony, F. U., Husain, M., & Apps, M. A. (2017). Prosocial apathy for helping others when effort is required. *Nature Human Behaviour*, *1*(7), Article 0131. https://doi.org/10.1038/s41562-017-0131

Lovibond, P. F. (1983). Facilitation of instrumental behavior by a Pavlovian appetitive conditioned stimulus. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*(3), 225–247.

Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*(1), 3–26.

Meeren, H. K., Van Heijnsbergen, C. C., & De Gelder, B. (2005). Rapid perceptual integration of facial expression and emotional body language. *Proceedings of the National Academy of Sciences, USA*, *102*(45), 16518–16523.

Mobbs, D., Yu, R., Meyer, M., Passamonti, L., Seymour, B., Calder, A. J., Schweizer, S., Frith, C. D., & Dalgleish, T. (2009). A key role for similarity in vicarious reward. *Science*, *324*(5929), 900.

Morelli, S. A., Sacchet, M. D., & Zaki, J. (2015). Common and distinct neural correlates of personal and vicarious reward: A quantitative meta-analysis. *NeuroImage*, *112*, 244–253.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*(2), 61–64.

Nook, E. C., Ong, D. C., Morelli, S. A., Mitchell, J. P., & Zaki, J. (2016). Prosocial conformity: Prosocial norms generalize across behavior and empathy. *Personality and Social Psychology Bulletin*, *42*(8), 1045–1062.

Olsson, A., Nearing, K. I., & Phelps, E. A. (2007). Learning fears by observing others: The neural systems of social fear transmission. *Social Cognitive and Affective Neuroscience*, *2*(1), 3–11.

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. Oxford University Press.

Payne, K., & Lundberg, K. (2014). The affect misattribution procedure: Ten years of evidence on reliability, validity, and mechanisms. *Social and Personality Psychology Compass*, *8*(12), 672–686.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*, 195–203.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, *43*(3), 151–160.

Ringwald, W. R., Vize, C. E., & Wright, A. G. (2025). Do you feel what I feel? The relation between congruence of perceived affect and self-reported empathy in daily life social situations. *Emotion*. doi:10.1037/emo0001531

Semmelmann, K., & Weigelt, S. (2017). Online psychophysics: Reaction time effects in cognitive experiments. *Behavior Research Methods*, *49*, 1241–1260.

Shamay-Tsoory, S. G., & Hertz, U. (2022). Adaptive empathy: a model for learning empathic responses in response to feedback. *Perspectives on Psychological Science*, *17*(4), 1008–1023.

Tracy, J. L., & Robins, R. W. (2008). The automaticity of emotion recognition. *Emotion*, *8*(1), 81–95.

Wagner, D. D., Kelley, W. M., & Heatherton, T. F. (2011). Individual differences in the spontaneous recruitment of brain regions supporting mental state understanding when viewing natural social scenes. *Cerebral Cortex*, *21*(12), 2788–2796.

Walther, E., Nagengast, B., & Trasselli, C. (2005). Evaluative conditioning in social psychology: Facts and speculations. *Cognition and Emotion*, *19*(2), 175–196.

Weisz, E., & Cikara, M. (2021). Strategic regulation of empathy. *Trends in Cognitive Sciences*, *25*(3), 213–227.

Yamada, M., Lamm, C., & Decety, J. (2011). Pleasing frowns, disappointing smiles: An ERP investigation of counterempathy. *Emotion*, *11*(6), 1336–1345.

Zaki, J. (2014). Empathy: A motivated account. *Psychological Bulletin*, *140*(6), 1608–1647.

Zaki, J. (2018). Empathy is a moral force. In K. Gray & J. Graham (Eds.), *Atlas of moral psychology* (pp. 49–58). Guilford Press.

Zaki, J., & Mitchell, J. P. (2011). Equitable decision making is associated with neural markers of intrinsic value. *Proceedings of the National Academy of Sciences, USA*, *108*(49), 19761–19766.

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, *27*(8), 1–25.

Zhou, Y., Han, S., Kang, P., Tobler, P. N., & Hein, G. (2024). The social transmission of empathy relies on observational reinforcement learning. *Proceedings of the National Academy of Sciences, USA*, *121*(9), Article e2313073121. https://doi.org/10.1073/pnas.2313073121