Social inequity disrupts reward-based learning

Huang Ham[1,2], Adrianna C. Jenkins[1]

[1]Department of Psychology, University of Pennsylvania

[2]Department of Psychology, Princeton University

July 3, 2025

Abstract

Through reward-based learning, people learn which actions generate which outcomes in which situations. What happens to human reward-based learning when outcomes are shared? Here we show that learning is impacted by inequity in the distribution of rewards (self-other) and others' identity. In three experiments, participants could learn how different actions, in response to different stimuli, generated different monetary rewards, each split between the participant and a member of a particular social group. Overall, participants learned more slowly and less successfully when they received a smaller (vs larger) share of the total reward. Stereotypes about the partner's social group additionally modulated learning rates when cognitive load was reduced, with lower perceived warmth or competence hampering learning from the partner's share. Computational modeling showed participants' learning was best explained by adapting the standard reinforcement learning model to account for stereotypes and inequity information, demonstrating that social context modulates non-social learning processes.

1 Introduction

Through experience, people learn links between their behaviors and the outcomes they produce, including which actions lead to which kinds of costs or benefits in which situations, on average. For example, across repeated experiences, someone might learn how much seasoning they like on their pasta, which floors of a parking garage will have available spaces on the weekend, or what type of social media post gets the most engagement. This sort of learning can be modeled using a reinforcement learning framework (Sutton & Barto, 1998), which formalizes the relationship between the expected and actual rewards of an action in terms of a reward prediction error (RPE). In humans and other species, these learning signals are reflected in the activity of subcortical structures in the brain (Frank, Seeberger, & O'reilly, 2004; Langdon, Sharpe, Schoenbaum, & Niv, 2018; Schultz, Dayan, & Montague, 1997) and known to guide behavior. More positive prediction error increases the likelihood of repeating the action, whereas more negative prediction error reduces the likelihood of repeating the action. Interestingly, people often perform actions whose rewards are not theirs alone but are instead distributed across themselves and others, whether choosing what kind of cake to order for a birthday party or choosing a driving route for a group road trip. How does the social distribution of rewards affect learning? Despite progress in characterizing the reinforcement learning processes, on the one hand, and people's preferences for the social distribution of resources, on the other, whether and how the social distribution of rewards impacts reinforcement learning is not well understood.

Although research on reinforcement learning has predominantly treated rewards as fixed (i.e., rewarding versus not rewarding; punishing versus not punishing), there is growing interest in characterizing contextual influences on valuation during reinforcement learning in a more flexible and continuous manner (Bavard, Lebreton, Khamassi, Coricelli, & Palminteri, 2018; Palminteri & Lebreton, 2021; Spektor, Gluth, Fontanesi, & Rieskamp, 2019; Suzuki & O'Doherty, 2020). In particular, a number of studies have documented effects of aspects of social context on learning. For example, social contexts can serve as information sources for learning, including through observational learning or advice-taking (Charpentier & O'Doherty, 2018; Hertz, Bell, & Raihani, 2021; Vélez & Gweon, 2019; Witt, Toyokawa, Lala, Gaissmaier, & Wu, 2024). There is also evidence that social context can serve as reward or punishment itself (e.g. smiling or positive feedback; Bhanji & Delgado, 2014; Heerey, 2014; Jones et al., 2011; Lindström, Selbing, Molapour, & Olsson, 2014), modulate the value one places on rewards received by others (Hackel, Zaki, & Van Bavel, 2017; Nafcha & Hertz, 2024), and shape the degree to which people learn vicariously from others' reward or punishment (Christopoulos & King-Casas, 2015; Lockwood, Apps, Valton, Viding, & Roiser, 2016; Sul et al., 2015). Still, little is known about (whether and) how social context influences basic non-social reinforcement learning process when rewards are shared.

In parallel, research on social decision-making has characterized people's preferences for how resources should be divided across themselves and others. Overall, people display social preferences for equity and fairness by making decisions that promote equal and fair distributions of resources, respectively (Fehr & Camerer, 2007), suggesting that people generally derive greater value from equal (or fair) distributions of resources than unequal ones, but if resources are divided unequally, people generally prefer to receive the larger share. Yet this preference can be moderated by factors, such as social distance (Strombach et al., 2015) and social group membership (Hackel, Mende-Siedlecki, Loken, & Amodio, 2022; Jenkins, Karashchuk, Zhu, & Hsu, 2018). In particular, recent evidence suggests that perceptions of others' traits, such as their warmth and competence, changes the value people place on particular divisions of resources across themselves and other people: the more warm the recipient, the less value people derive from *advantageous inequity,* i.e., receiving more than the other person, whereas the more competent the recipient, the less value people derive from *disadvantageous inequity,* i.e., receiving less than the other person (Jenkins et al., 2018; Kobayashi, Kable, Hsu, & Jenkins, 2022).

To what extent do patterns of social valuation, observed during decision-making, extend to impact reward signals during learning? Outside the social domain, differences between subjective value measured during reward-based learning versus decision-making suggest the non-triviality of this question. For example, during learning, individuals show range adaptation: that is, they demonstrate no meaningful advantage to learning when the rewards at play are in the range of

\$100 in magnitude than when rewards (differ by the same proportions but) are in the range of \$10 in magnitude (Bavard, Rustichini, & Palminteri, 2021; Rustichini, Conen, Cai, & Padoa-Schioppa, 2017; Webb, Glimcher, & Louie, 2021). Yet during decision-making, people demonstrably value actions that generate \$100 over those that generate \$10. Likewise, although individuals generally value advantageous over disadvantageous inequity during decision-making, it could be the case that subjective value shows adaptation to the inequity context. If so, individuals exposed only to (various levels of) advantageous inequity and individuals exposed only to (various levels of) disadvantageous inequity would learn equally well from the best split percentage to which they are exposed, even though individuals in the advantageous context are exposed to better split percentages overall.

In the current studies, we addressed these questions by adapting a computerized, non-social reinforcement learning task from Collins and Frank (2012) to accommodate manipulations of the inequity of the reward (reward distributions across the learner and a partner, by percentage) and the social identity of the partner (Figure 1). We chose this task paradigm because it can isolate contributions of lower-level reinforcement learning processes (versus executive processes) to performance while accommodating parametric modulations of inequity along with manipulations of partner identity. In 3 studies, participants had opportunities to learn the reward of stimulus-action pairs (images and button presses) under conditions of (in)equity in the social distribution of the rewards. One button press generated the largest reward, another generated the smallest reward, and a third generated an intermediate reward. Participants played 8 independent blocks of the task. In each block, they were first shown the set of images they would encounter in that block (5 images per block in Studies 1 and 2; 2 images per block in Study 3), followed by a piece of information about the partner with whom the rewards would be split during that block (e.g., “a nurse”). Then, on each trial, the participant saw an image on the screen and chose which of 3 eligible buttons on the keyboard to press (j, k, or l), within 1.5 seconds. Following the button press, a feedback screen displayed how much money they and the other person gained (e.g., You: \$.30; Nurse: \$.70). In each block, 12 trials per image were intermixed in a random order, for (5 images x 12 trials =)

60 trials per block in Studies 1 and 2 and (2 images x 12 trials =) 24 trials per block in Study 3. The total reward to be split across the participant and the partner depended on the button press: for each image, one button deterministically generated the highest reward (e.g., $2), another deterministically generated the intermediate reward (e.g., $1), and a third deterministically generated the lowest reward (e.g., $0). Inequity type was manipulated between subjects, such that each participant was either assigned to the advantageous condition (they always gained more than 50%) or the disadvantageous condition (they always gained less than 50%). Within each inequity type, specific split percentage were manipulated within subjects. To implement this, each image corresponded to a specific reward split between the participant and the partner (unbeknownst to participants). For example, for one participant in one block, the car image was always associated with the participant receiving 30% of the money and the other person receiving 70%.

In Study 1, the total amount of reward (summing the participant and the partner's reward) was held constant across inequity conditions, varying the amount the participant personally received. In Study 2, the amount that a participant received themselves was held constant across inequity conditions, varying the total amount of reward summing across the participant and the partner. In Study 3, in light of prior research suggesting that cognitive load may dampen social information processing (Jenkins, 2019; Sullivan-Toole, Dobryakova, DePasque, & Tricomi, 2019), we reduced the total number of stimulus-action pairs in order to reduce cognitive load during learning, allowing for the possibility that effects of social information may be elevated in such cases. In all studies, we varied the identity of the partner by informing the participant of their occupation (e.g., Nurse), following past studies (Goncharova & Jenkins, submitted; Jenkins et al., 2018; Kobayashi et al., 2022) and collected participants' ratings of the perceived warmth and competence of people with these occupations, both following the learning task and in an independent set of participants (Fiske, Cuddy, & Glick, 2007). The occupation information remained the same within each learning block.

We raise 3 possible, not mutually exclusive hypotheses about the ways in which social contextual information might shape reinforcement learning. First, reinforcement learning may operate on objective rewards during learning (such as the total reward generated by different stimulus-action pairings). Second, reinforcement learning may be shaped by the degree of social inequity of the reward (i.e., how much more or less of the reward the participant receives than the partner). Third, reinforcement learning may be shaped by the perceived traits of the partners with whom rewards are shared. We formalized these hypotheses into 4 computational models and tested their fit to participants' data. Two models formalize the first hypothesis, assuming only effects of objective reward, with no effect of social inequity or identity on reinforcement learning (baseline RL models). The baseline model in Study 1 and 3 only learned from the total reward because by experimental design, the total reward did not depend on the inequity manipulation. In Study 2, the baseline model only learned from the amount of reward given to the participant because Study 2 was designed such that the total reward amount depended on the inequity manipulation while the reward given to the participant did not depend on the inequity manipulation. To formalize the second hypothesis, we constructed an inequity-weighted reinforcement learning model (IRL), which assumes that the rewards of both the learner and the target, along with the (signed) difference between them, affect learning. To formalize the third hypothesis, we created a social perception-weighted reinforcement learning model (SPRL), which assumes that the perceived traits (warmth and competence) of the social target combine with inequity to impact learning. This model builds upon our findings in social decision-making that perceptions of others' traits modulate people's preferences for (or against) inequity (Jenkins et al., 2018; Kobayashi et al., 2022). Here, we applied the utility function from our previously-developed social perception weighted model of social valuation to calculate the reward signal (Q value) for reinforcement learning (see Method). By comparing the fits of each of these models to participants' learning behavior, we can distinguish among the different hypotheses about whether and how social contextual information impacts reinforcement learning. All procedures were conducted in a manner approved by the Institutional Review Board at the University of Pennsylvania (protocol #831852).

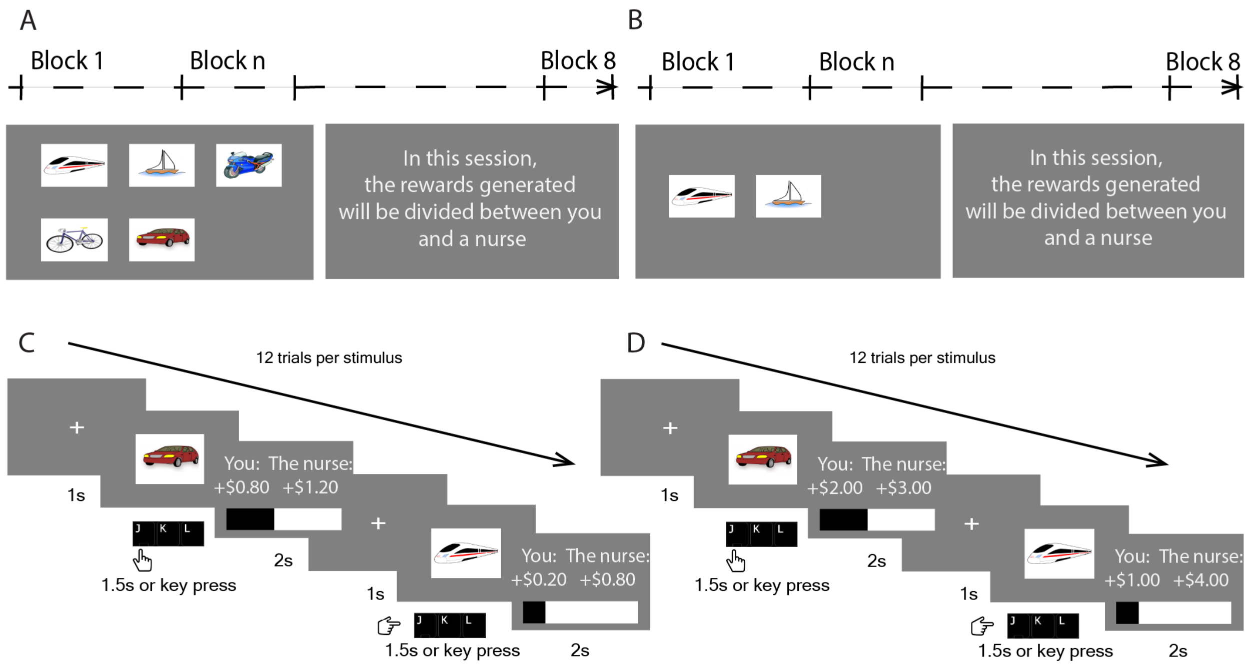

Figure 1: Example task. A, B) The task was grouped into 8 independent blocks. Before each block, participants saw all stimuli they may encounter, and the occupation of the social target that would receive a share of the reward. Figure A) shows the block structure of Study 1 & 2, where people had to learn about 5 stimuli per block. Figure B) shows Study 3, where the cognitive load was reduced by only learning about 2 stimuli per block. C, D) Trial structures: Each image stimulus would appear for 12 trials in total and corresponded to one percentage of split. The trials for all 5 stimuli were randomly shuffled. Figure C) shows the trial structure of Study 1 & 3 where each action generated a fixed total reward ($0, $1, or $2) which was then split between the participant and the social target. Figure D) shows the structure of Study 2, where each action generated a fixed reward to the participant ($0, $1, or $2). The reward went to the social target, and hence the total reward, depended on the specific percentage of split. In all studies, both the participant and the social target would gain more (hypothetical) money if the participant learned quickly to press the button that generates the highest reward.

2 Methods

2.1 Study 1

2.1.1 Participants

Participants (N = 99) were recruited through the University of Pennsylvania's SONA platform and earned undergraduate psychology class credit for their online participation. 5 participants were excluded for a response rate below 80%, leaving us with 94 participants for analysis and modeling. 41 participants were randomly assigned to the *advantageous* condition (M_{age} = 19.63, 32 women, 8 men) and 53 participants were randomly assigned to the *disadvantageous* condition (M_{age} = 19.79, 35 women, 16 men, 2 self-described). Demographic information was collected in a self-report survey after the experiment. All procedures were conducted in a manner approved by the Institutional Review Board at the University of Pennsylvania (protocol #831852). Informed consent was obtained from all participants before their participation. The study was not preregistered.

2.1.2 Experimental Procedure

After choosing to participate on the SONA system, participants proceeded to a new browser window to start the experiment which was coded in PsychoPy (Peirce et al., 2019), converted to PsychoJS, and hosted on Pavlovia (pavlovia.org). On each trial of the experiment, the participant viewed a stimulus image on the screen and chose which of three possible keys to press in response; this action would generate some amount of monetary reward. Crucially, the monetary rewards obtained would ostensibly be split between themselves and another person, making it possible to manipulate inequity. We used images from (Collins, Ciullo, Frank, & Badre, 2017) as stimuli in our task.

The task consisted of 8 blocks of trials. A trial was the smallest unit of the task where participants saw a stimulus (in this case, an image) presented on the screen and pressed a key in response. A block was a collection of trials run one after another. Our experiment had one between-subject manipulation, one between-block manipulation, and one within-block manipulation. Between subjects, we manipulated *inequity type*. We assigned each participant randomly into either an *advantageous* condition, where the participant received more than half of the reward, or a *disadvantageous* condition, where the participant received less than half of the reward. We chose to manipulate inequity type between-subjects for two main reasons, First, this allowed for a stronger test for the absolute effect of inequity type on learning. If each participant had experience with both advantageous and disadvantageous splits, any difference between the two conditions could have arisen due to anchoring effects. Second, the between-subject design kept the duration of the experiment within a range that minimized concerns about data quality. Similarly, given that learning on a non-social version of this task is already documented (Collins et al., 2017; Collins & Frank, 2012), we chose not to include a non-social learning condition in the current study. Between blocks, we manipulated the social group membership of the partner (social target). Before each block, participants were told the occupation of the social target (Taxi Driver, Nurse, Judge, Computer programmer, Dancer, Plumber, Politician, or Secretary). These 8 occupations were selected to cover as widely as possible the range of warmth and competence ratings according to Jenkins et al. (2018). Each occupation was randomly assigned to one of the 8 inequity blocks. Within a block, each stimulus image corresponded to a different *percentage of split*. For example, if the stimulus on the screen was an image of a lotus, 30% of the reward generated by the keypress would go to the participant, whereas if it was an image of a tree, the participant would get 0% of the reward. There were 5 unique stimuli in each block, corresponding to 100%, 90%, 80%, 70%, and 60% of the reward in the advantageous condition, and 0%, 10%, 20%, 30%, or 40% in the disadvantageous condition.

Each block was a learning problem independent of the others, using distinct sets of images. At the beginning of each block, the screen presented all 5 images that the participants would encounter in that block. 12 iterations of each stimulus were interleaved throughout each block.

Within each block, the trial structure was as follows: The trial began with an inter-trial interval of 1.5 seconds during which only a white cross was displayed at the center of the screen. Next, participants viewed one of the learning stimuli on the screen which was continually displayed either until the participant responded with one of the three possible actions ("j", "k", or "l"), or if 1.5 seconds had elapsed. Each possible action generated either \$0, \$1, or \$2 as monetary reward. For each particular stimulus, two different keys could not generate the same reward amount. For any two stimuli in the same block, there was at least one key that generated different rewards for either stimulus. Upon pressing the key at each trial, participants would see on the screen how much reward they had received and how much reward the other person had received. The respective reward amount was determined by multiplying the generated reward with the split percentage corresponding to the stimulus. However, if the reaction time exceeded 1.5 seconds, the screen instead displayed the message "please respond faster", and if the response was faster than 0.15, the response would not register and the participant must press the key again. The feedback stayed for 2 seconds before transitioning to the next trial.

To become familiarized with the tasks, participants first read through the instructions, which informed participants that each key generated either $0, $1, or $2 as total reward but part of it would be split to another person. The instructions also emphasized that participants should imagine the rewards were actual money. We did not explicitly state that the goal of the task was to get as much reward as possible, in which case participants may deliberately ignore social information and solely focus on the total reward. In order to elicit how participants would naturally incorporate social information with rewards, we stated merely that paying more attention could help you get more rewards. Participants then performed one practice block which had the same structure as a regular block in the task but used a different set of images, and participants were told to receive 50% of the generated rewards while the remainder went to a singer (which was an occupation not used in the main task). After the entire experiment, participants were redirected to Qualtrics (qualtrics.com) to complete a short demographic survey where we also listed the 8 occupations that they had encountered during the task and asked them to rate from 0 to 100 how warm and how competent they perceived each of the occupations (see Figure S3). The order of these questions in which we presented to participants on Qualtircs was randomized. Because the social perception rating was always collected after the learning task, we confirmed that there was no evidence the between-subject manipulation systematically biased participants' ratings of warmth or competence in any of the 3 studies through t test (Study 1: $t_{competence}(679) = -0.81, p = .421, d = -0.06, 95\%CI = [-0.20, 0.08]$; $t_{warmth}(705) = -1.76, p = .079, d = -0.13, 95\%CI = [-0.27, 0.01]$; Study 2: $t_{competence}(595) = 0.42, p = .671, d = 0.03, 95\%CI = [-0.12, 0.18]$; $t_{warmth}(625) = 0.83, p = .407, d = 0.06, 95\%CI = [-0.09, 0.21]$; Study 3: $t_{competence}(744) = -1.05, p = .296, d = -0.08, 95\%CI = [-0.22, 0.07]$; $t_{warmth}(753) = -0.88, p = .382, d = -0.06, 95\%CI = [-0.21, 0.08]$)

2.2 Study 2

2.2.1 Participants

Participants (N = 100) were recruited through the University of Pennsylvania's SONA platform and earned undergraduate psychology class credit for their online participation. 9 participants were excluded for a response rate below 80%, leaving us with 91 participants for analysis and modeling. 55 participants were assigned to the *advantageous* condition (M_{age} = 19.87, 37 women, 16 men) and 36 participants were assigned to the *disadvantageous* condition (M_{age} = 20.03, 28 women, 8 men). Demographic information was collected in a self-report survey after the experiment. All procedures were conducted in a manner approved by the Institutional Review Board at the University of Pennsylvania (protocol #831852). Informed consent was obtained from all participants before their participation. The study was not preregistered.

2.2.2 Experimental Procedure

After choosing to participate on the SONA system, participants proceeded to a new browser window to start the experiment which was coded in PsychoPy (Peirce et al., 2019), converted to PsychoJS, and hosted on Pavlovia (pavlovia.org). The experimental design was identical to that of Study 1, with an important exception: in Study 2, the amount of reward given to the participant was held constant across split conditions. Consequently, the total amount of rewards generated may vary. Therefore we accordingly modified the instruction. We did not inform the participants that each key generates either \$0, \$1, or \$2, we just told them that each key may generate a different amount of total reward. In actuality, each key generated the amount of total reward such that the amount given to the participant was either \$0, \$1, or \$2. The total reward was thus either \$0, \$1 divided by the percentage of split to the participant, or \$2 divided by the percentage of split to the participant. Because the 0% was not mathematically possible under this new design, we replaced it with 50%.

2.3 Study 3

2.3.1 Participants

Participants (N = 100) were recruited through the University of Pennsylvania's SONA platform and earned undergraduate psychology class credit for their online participation. 5 participants were excluded for a response rate below 80%, leaving us with 95 participants for analysis and modeling. 47 participants were assigned to the *advantageous* condition (M_{age} = 19.76, 29 women, 17 men) and 48 participants were assigned to the *disadvantageous* condition (M_{age} = 19.92, 23 women, 23 men, 1 self-described). Demographic information was collected in a self-report survey after the experiment. All procedures were conducted in a manner approved by the Institutional Review Board at the University of Pennsylvania (protocol #831852). Informed consent was obtained from all participants before their participation. The study was not preregistered.

2.3.2 Experimental Procedure

After choosing to participate on the SONA system, participants proceeded to a new browser window to start the experiment which was coded in PsychoPy (Peirce et al., 2019), converted to PsychoJS, and hosted on Pavlovia (pavlovia.org). The experimental design was identical to that of Study 1, with one important exception: to reduce cognitive load, we reduced the total number of stimuli to 2 unique stimuli per block (rather than 5), where each stimulus corresponded to 90% or 70% in the advantageous condition, and 10% or 30% in the disadvantageous condition.

2.4 Regression analyses

2.4.1 Full regression predicting rewardingness of action by trial

We ran the following mixed-effect linear regression for each of the 3 studies:

$$rewardingness \sim iteration * split * inequity * warmth * competence + (1|subject) \quad (1)$$

where *rewardingness* denotes the reward amount independent of the manipulation of social distribution (\$0, \$1, or \$2). *iteration* denotes how many iterations has the stimulus appeared in a particular learning block (1, 2, ..., 12). *split* denotes the percentage of the reward given to participant themselves (0, 0.1, 0.2, ..., 1). *inequity* is a categorical variable denoting whether the participant is in the advantageous inequity condition or the disadvantageous inequity condition. *warmth* denotes participant's rating on the perceived warmth of the social target (1 - 100). *competence* denotes participant's rating on the perceived competence of the social target (1 - 100). *subject* denotes the participant ID that we use as a random intercept. Before passing into the regression model, we standardized *rewardingness*, *iteration*, *warmth*, and *competence* across the entire data set and we also standardized *split* within the advantageous or disadvantageous conditions so that it is not confounded with *inequity*.

2.4.2 Regression by separate inequity type

We ran the following mixed-effect linear regression for both the advantageous condition and the disadvantageous condition in each of the 3 studies:

$$rewardingness \sim iteration * split * warmth * competence + (1|subject) \tag{2}$$

All variables were standardized and we removed the split condition where the participant received 0% in Study 1 and the split condition where the participant received 50% in Study 2 as they might serve as edge cases that drove the results in the full regression model.

2.4.3 Regression predicting reaction time of action by trial

We ran the following mixed-effect linear regression for each of the 3 studies and the results were referenced in the supplemental (Figure S1):

$$rt \sim iteration * split + (split|subject) + (split|inequity) \tag{3}$$

rt denotes the reaction time at each trial. Both participant ID and inequity type were included as random intercept and *split* as randomly slope, allowing the effect of split to potentially differ across participants. Before passing into the regression model, we standardized *rt* and *iteration* across the entire data set and we also standardized *split* within the advantageous or disadvantageous conditions so that it is not confounded with *inequity*.

2.4.4 Regression predicting trial-by-trial model comparison

We ran the following mixed-effect linear regression for each of the 3 studies:

$$\Delta WAIC \sim iteration * inequity + (1|subject) \tag{4}$$

where *iteration* was centered by subtracting 5 from it. $\Delta WAIC$ denotes the trial-by-trial difference in WAIC between the inequity-weighted model and the baseline model, showing how much did the former outperform the latter in fitting the participant data.

2.4.5 Procedures

For the regression analyses, we tested the impact of different task variables by performing *mixed-effect linear regression* analysis using R function *mixed* in package *afex* (Brown, 2021). All numeric variables were scaled before being passed into the regression model and all interaction terms were included. The package *afex* conducts significance testing on regression coefficients by comparing the fit of the full model with the truncated model without that regressor through χ^2 test. To conduct Wilcoxon test, we used the wilcox_test function from the package *rstatix*. To conduct two-sample t test, we used the t_test function from the package *rstatix*. For both tests we averaged within participants first so the sample sizes are the number of participants in each condition and both are two-sided tests.

2.5 Computational Models

All of our candidate models were adaptations of a typical reinforcement learning model (Sutton & Barto, 1998). The model relies on two main variables representing the task environment. The first one is the state $s \in S$, where S represents the full stimulus/state space within a block (i.e., all the possible images that could appear). In our experiment, $|S| = 5$ except in Study 3 where $|S| = 2$. The second variable is the action $a \in A$, where A is the full action space. In our experiment, $|A| = 3$

because there were three possible buttons to press as a response to the instrumental learning task. The algorithm proceeds in two stages, as introduced in the introduction: the value updating stage and the policy formation stage. In the value updating stage, for stimulus s and action a on trial t, the model estimates an expected value (i.e., the Q value) $\boldsymbol{Q}(s_t, a_t)$ by performing an update using the delta rule (Equation 6; Rescorla, 1972):

$$\boldsymbol{Q}_{t+1}(s_t, a_t) = \boldsymbol{Q}_t(s_t, a_t) + \alpha\delta_t \tag{5}$$

$$\delta_t = r_t - \boldsymbol{Q}_t(s_t, a_t) \tag{6}$$

where α represents the **learning rate** and $\boldsymbol{Q}_t$ represents a $|S| \times |A|$ matrix encoding all Q values given a trial t. $\boldsymbol{Q}_0$ is initialized as a uniform matrix of the expected values of random guessing. $\delta \in \mathbb{R}$ is the reward prediction error, and $r_t \in \{0, 1, 2\}$ is the reward received at trial t.

In the policy formation stage, Q values are transformed by the *Softmax function* into a policy, i.e., a vector of probabilities of taking each action (represented by $\vec{\pi}_t$).

$$\vec{\pi}_t = p(\vec{A}|s_t) = Softmax(\boldsymbol{Q}_t(s_t), \beta) = \frac{e^{\beta \boldsymbol{Q}_t(s_t)}}{\sum_{a \in A} e^{\beta \boldsymbol{Q}_t(s_t, a)}} \tag{7}$$

where $\beta \in [0, \infty)$ represents the **inverse softmax temperature**. Finally, we allow all Q values to decay back to the initial Q values with a **decay rate** ($\phi \in [0, 1]$) to capture subjects' forgetting process.

$$\boldsymbol{Q}_{t+1}(s_t, a_t) = \boldsymbol{Q}_t(s_t, a_t) + \phi(\boldsymbol{Q}_0(s_t, a_t) - \boldsymbol{Q}_t(s_t, a_t)) \tag{8}$$

2.5.1 Naive

The naive model is just the same reinforcement learning model as described above, where the inequity and social target information are entirely discarded. This model instantiates the hypothesis that participants only learn by the total amount of reward generated and ignore how the reward is split afterward. In Study 2, just like in Study 1, the naive model is the model that only learns based on the total reward, but the total reward is dependent on the split by experimental design: $r_t \in \{0, \frac{1}{p_t^{\text{self}}}, \frac{2}{p_t^{\text{self}}}\}$ and $\forall a \in A$, $\boldsymbol{Q}_0(s_t, a) = \frac{1}{p_t^{\text{self}}}$. This model still instantiates the hypothesis that participants only learn by the total amount of reward generated and ignore how the reward is split afterward. The baseline model reported in the main task for Study 1 and 3 was the naive model since the total reward was independent of the split condition.

2.5.2 Selfish

The selfish model instantiates the hypothesis that participants only learn by the amount of reward given to themselves. It is the same as the naive model except it only uses the amount of reward given to self:

$$\delta_t = r_t * p_t^{\text{self}} - \boldsymbol{Q}_t(s_t, a_t) \tag{9}$$

where p_t denotes the percentage of the reward given to the participant at that trial, and $\boldsymbol{Q}_0$ is initialized to be the percentage of split corresponding to the trial where a new stimulus first appears: if $\forall t' < t, s_{t'} \neq s_t$, we have $\forall a \in A$,

$$\boldsymbol{Q}_0(s_t, a) = p_t^{\text{self}} \tag{10}$$

In Study 2, the concept of the selfish model is the same. Because in Study 2 we ensure the reward given to self does not depend on p_t^{self}, $r_t * p_t^{\text{self}} \in \{0, 1, 2\}$ and $\boldsymbol{Q}_0(s_t, a) = 1$. The baseline model reported in the main text for Study 2 was the selfish model.

2.5.3 IRL

The inequity-weighted model (IRL) instantiates the hypothesis that the degree of inequity of how the reward is split between the participant and the social target shapes how the reward is represented (Barnby, Raihani, & Dayan, 2022). We formalize the effect of inequity using the Fehr-Schmitt utility function (Rohde, 2010):

$$u_t = r_t * (p_t^{\text{self}} + \gamma * (p_t^{\text{target}} - p_t^{\text{self}})) \tag{11}$$

where in our task, $p_t^{\text{target}} = 1 - p_t^{\text{self}}$ is simply the percentage of reward given to the recipient, and the **inequity weight** $\gamma \in \mathbb{R}$ controls the influence of the inequity. In the disadvantageous inequity condition, positive γ means that the participant finds it more rewarding if the other person obtained more reward than themselves. Negative γ means that the participant finds it less rewarding if the other person obtained more reward than themselves. In the advantageous inequity condition, positive γ means that the participant finds it less rewarding if the other person obtained less reward than themselves. Negative γ means that the participant finds it more rewarding when the other person obtained less reward than themselves. If $\gamma = 0$, it means the participant only cares about the reward given to themselves and is indifferent to the inequity. In general, γ captures in what direction and to what extent a participant deviates from a purely selfish agent and cares about inequity in the reward distribution. The delta rule thus becomes

$$\delta_t = u_t - \boldsymbol{Q}_t(s_t, a_t) \tag{12}$$

$\boldsymbol{Q}_0$ is initialized to be the expected utility corresponding to the trial where a new stimulus first appears: if $\forall t' < t, s_{t'} \neq s_t$, we have $\forall a \in A$,

$$\boldsymbol{Q}_0(s_t, a) = p_t^{\text{self}} + \gamma * (p_t^{\text{target}} - p_t^{\text{self}}) \tag{13}$$

We initialized Q_0 differently for different split conditions because initializing Q_0 with a flat value significantly undermines the ability for the model to capture the key effects of learning. We confirmed this by simulating the winning model with a flat initialization of Q_0 on the Study 3 data and showed that it failed to produce the split effect that was produced by human participants in the disadvantageous condition (Figure S9).

2.5.4 SPRL

The social perception-weighted model (SPRL) instantiates the hypothesis that not only does the degree of inequity shape how the reward is represented, but the sensitivity to the inequity further depends on the social perception of the recipient's identity. We adopt the social preference model from (Jenkins et al., 2018) to capture the effect of warmth and competence on the inequity weighting parameter:

$$\gamma = \gamma_0 + \gamma_w * w + \gamma_c * c \tag{14}$$

where w is the warmth rating and c is the competence rating that we collected in the post-task survey. The rating scale we used for the survey was from 0 to 100. For the stability of modeling fitting, we rescaled the social perception ratings into $[-0.5, 0.5]$. $\gamma_0, \gamma_w, \gamma_c \in \mathbb{R}$ are respectively the **base weight**, **warmth weight**, and **competence weight**.

2.6 Modeling Procedure

We fitted all model parameters using hierarchical Bayesian methods. Compared to the traditional maximum likelihood estimation, not only does the Bayesian fitting method give us a full posterior distribution over the fitted parameters (instead of simply one point estimate), but it also yields a superior parameter and model recovery (Baribault & Collins, 2025; Eckstein, Master, Dahl, Wilbrecht, & Collins, 2022). The population-level priors for all model parameters were carefully tuned to be as uninformative as possible while avoiding divergence during fitting:

$$\boldsymbol{\alpha} \sim \text{Beta}(\alpha = 1, \beta = 1)$$

$$\boldsymbol{\beta} \sim \text{Gamma}(\alpha = 3, \beta = 0.5)$$

$$\boldsymbol{\phi} \sim \text{Beta}(\alpha = 2, \beta = 15)$$

$$\boldsymbol{\gamma_w}, \boldsymbol{\gamma_c} \sim \text{Normal}(\mu = 0, \sigma = 1)$$

$$\text{In Study 1: } \boldsymbol{\gamma}, \boldsymbol{\gamma_0} \sim \text{Normal}(\mu = 0.5, \sigma = 3)$$

$$\text{In Study 2: } \boldsymbol{\gamma}, \boldsymbol{\gamma_0} \sim \text{Normal}(\mu = 0, \sigma = 3)$$

$$\text{In Study 3: } \boldsymbol{\gamma}, \boldsymbol{\gamma_0} \sim \text{Normal}(\mu = 0.5, \sigma = 1)$$

We performed fitting using the python PyMC4 package version 4.1.3 (Salvatier, Wiecki, & Fonnesbeck, 2016) via the no-U-Turn sampler, which was the state-of-the-art Markov-chain Monte Carlo sampling method to estimate parameter posteriors. For each model, we ran 3 chains of 1000 tuning samples (which were discarded) and 2000 kept samples used to estimate the posterior distributions. Therefore in total 6000 samples were used to represent each parameter's posterior distribution. For diagnostic checks, we required $\hat{R} \leq 1.01$, $BFMI \geq 0.2$ for all chains, a sufficiently large effective sample size ($ESS \geq 400$) for all parameters, and that no divergences were observed. Besides these computational diagnostics, we also performed prior predictive checks to make sure that the priors had a reasonable level of informativeness, demonstrated that the fitting procedure could recover parameters that we generated from the prior distributions, and ensured that each model overall fit best to the data simulated by themselves, not by other candidate models (model recovery). For prior predictive checks and model validation (Figure 2, S2), we simulated each model 20 times per subject. For parameter recovery (Figure S5, S6, S7) and model recovery (Table S3, S4), we only simulated each model once per subject. We fit the data to various computational agents with different learning parameters and compared them using the Widely Applicable Information Criterion (WAIC; Watanabe, 2013). All point estimates of parameter values per participant were the mean of the fitted posterior distributions and different parameters are minimally correlated across subjects (Figure S8). The human data used for model fitting did not include the first iteration because the learning performance should be chance level at the first iteration and thus was not informative to model fitting.

3 Results

First, we confirmed that participants were able to learn the stimulus-action-reward mappings across the course of the experiment (Figure 2, 4). We ran a mixed-effect linear regression with the average rewardingness of actions as the dependent variable (equation 1, Table S5-S6). The average reward generated by participants in response to a stimulus increased with the number of stimulus appearances, as suggested by the significant positive main effect of stimulus iteration (Study 1: *b*

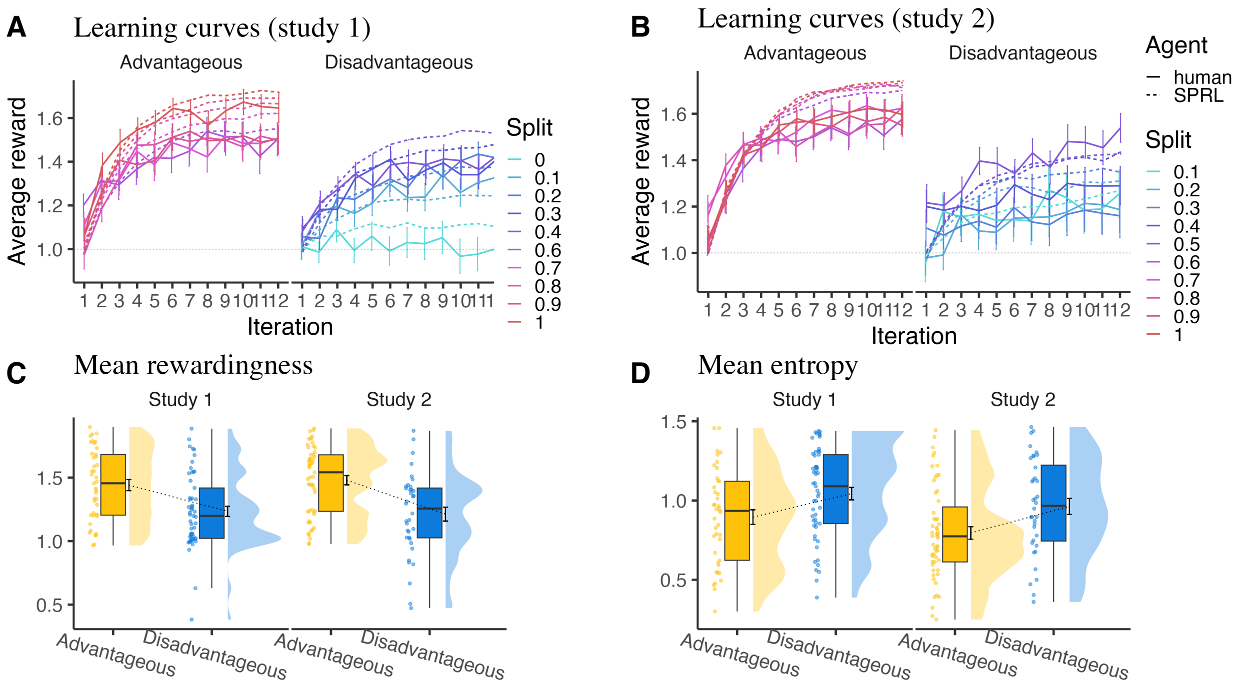

Figure 2: Main behavioral results for Study 1 & 2: A, B) learning curve of Study 1 (n = 41 in advantageous condition and n = 53 in disadvantageous condition) and Study 2 (n = 55 in advantageous condition and n = 36 in disadvantageous condition). Participants overall converged to generating higher total reward. Curves reflect the rewardingness of actions as a function of the number of times that each stimulus was presented, plotted separately for each split condition (percentage of reward given to the participant). Dashed lines are the simulated learning curve by the best-fitting SPRL model. C) overall rewardingness of actions averaged within the advantageous and disadvantageous conditions. People generated fewer rewards under disadvantageous inequity where they received the smaller share of reward. D) overall entropy of actions averaged within the advantageous and disadvantageous conditions. People's actions were less deterministic under disadvantageous inequity where they received the smaller share of reward. This suggests that people were not simply voluntarily choosing less rewarding actions but indeed had more uncertainty during learning. All error bars reflect the s.e.m.

= 0.105 , $\chi^2(1)$ =551.94 , p < 0.001, 95% CI = [0.0962, 0.1138]; Study 2: b = 0.108, $\chi^2(1)$ = 541.39, p < 0.001, 95% CI = [0.0989, 0.1171]; Study 3: b = 0.253, $\chi^2(1)$ = 1235.72, p < 0.001), 95% CI = [0.2389, 0.2671]. These results indicate that, overall, participants succeeded in learning the stimulus-response-reward mappings.

3.1 Inequity in reward distribution affects learning performance

Although participants succeeded overall in learning, so as to generate more reward across the course of the experiment, we observed a significant main effect of inequity type on reward (Figure 2, 4). Specifically, across all three studies, participants obtained significantly lower overall reward in the disadvantageous inequity condition (equation 1, Table S5-S6) than the advantageous inequity condition (Study 1: $b = 0.134$, $\chi^2(1) = 12.00$, p <0.001, 95% CI = [0.0582, 0.2098], Study 2: $b = 0.178$, $\chi^2(1) = 18.51$, p = <0.001, 95% CI = [0.0969, 0.2591], Study 3: $b = 0.143$, $\chi^2(1) = 16.83$, $p < 0.001$, 95% CI = [0.0747, 0.2113]). Additionally, we observed a significant interaction between iteration and inequity type (equation 1, Table S5-S6), such that participants learned more slowly in the disadvantageous condition (Study 1: $b = 0.022$, $\chi^2(1) = 24.47$, $p < 0.001$, 95% CI = [0.0133, 0.0307], Study 2: $b = 0.043$, $\chi^2(1) = 85.58$, $p < 0.001$, 95% CI = [0.0339, 0.0521], Study 3: $b = 0.008$, $\chi^2(1) = 1.18$, $p = .278$, 95% CI = [-0.0064, 0.0224]). There are at least two possible explanations for the observation that participants generated lower reward in the disadvantageous condition. One possibility is that participants learned less well which stimulus-action combinations were most rewarding. Another possibility is that participants deliberately chose actions that generated lower overall reward in these conditions. To distinguish these possibilities, we compared the overall Shannon entropy of participants' choice in the advantageous condition with the entropy in the disadvantageous condition. Shannon entropy is a measurement of how deterministic (or how stochastic) a random variable is. The higher the entropy is, the less deterministic participants' choices are. If the sole reason for participants generating less reward in the disadvantageous condition was that they deliberately chose the less rewarding actions, we would not see any difference in the entropy between advantageous and disadvantageous conditions. Their actions would be equally deterministic in both conditions but deterministic towards more rewarding actions in the advantageous condition. However, we did identify a significantly higher entropy in the disadvantageous condition through Wilcoxon test (Study 1: $U = 784$, $p = 0.021$, Study 2: $U = 667$, $p = 0.009$, Study 3: $U = 434$, $p < 0.001$), suggesting that participants' choices were indeed less deterministic in the disadvantageous condition (Figure 2, 4). The SPRL model was also able

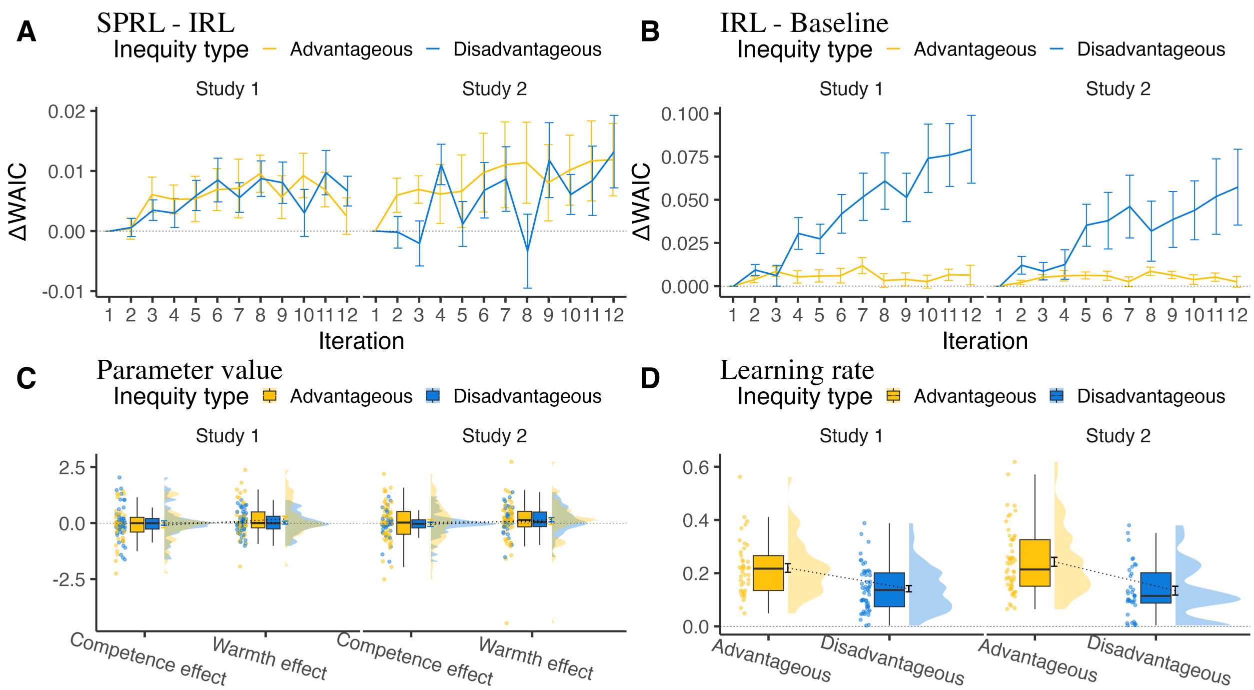

Figure 3: Modeling results of Study 1 (n = 41 in advantageous condition and n = 53 in disadvantageous condition) and Study 2 (n = 55 in advantageous condition and n = 36 in disadvantageous condition): A, B) The trial-by-trial difference in WAIC between the better fit model and worse fit model increases over the course of learning, suggesting the effect of social information enhances over learning. Figure B) also suggests that the effect of inequity was stronger under disadvantageous inequity. C) Fitted weight of the competence and warmth rating: no clear directional effect of the perceived warmth or competence in Study 1 and 2. D) Fitted learning rate of the *baseline model*: further confirms the behavioral result that people learned slower under disadvantageous inequity. Each dot represents a point estimate for one subject obtained as the mean of the posterior distribution of the parameters. All error bars reflect the s.e.m.

to simulate this effect in entropy (Figure S4). We discuss more in depth some other possible interpretations in the supplemental (section S1; Figure S1). Put together, these behavioral evidence suggest that learning was more disrupted in the disadvantageous inequity condition. We further confirmed this result by comparing the fitted learning rate parameter of the *baseline* reinforcement learning models (Figure 3, 4). Wilcoxon test supports that the learning rates of participants in the disadvantageous condition are lower than those in the advantageous condition (Study 1: U = 1578, $p < 0.001$, Study 2: U = 1520, $p < 0.001$, Study 3: U = 1823, $p < 0.001$).

3.2 Learning performance is more sensitive to the self-other difference in reward under disadvantageous inequity

We also found a significant positive main effect of the split percentage on the reward generated by participants (Study 1: b = 0.089, $\chi^2(1)$ = 393.78, p < 0.001,95% CI = [0.0802, 0.0978], Study 2: b = 0.054, $\chi^2(1)$ = 136.65, p < 0.001,95% CI = [0.0449, 0.0631], Study 3: b = 0.044, $\chi^2(1)$ = 38.68, p < 0.001,95% CI = [0.0301, 0.0579]). This suggests that participants overall learned better when they earned more percentage of the reward. We also found a significant interaction effect, suggesting that the effect of split percentage was stronger in the disadvantageous condition (Study 1: b = -0.043, $\chi^2(1)$ = 93.58, p < 0.001,95% CI = [-0.0517, -0.0343], Study 2: b = -0.047, $\chi^2(1)$ = 102.25, p < 0.001,95% CI = [-0.0561, -0.0379], Study 3: b = -0.030, $\chi^2(1)$ = 17.44, p < 0.001,95% CI = [-0.0441, -0.0159]).

To further explore this interaction effect, we fit two *mixed-effect linear regression* separately in the advantageous and disadvantageous conditions (equation 2, Table S7-S8). In the disadvantageous condition, we removed the split condition in Study 1 where the participant obtained 0% of the reward and in Study 2 where the participant obtained 50%, to make sure the effect was not solely driven by these extreme conditions. Across 3 studies, we observed a significant positive effect of split (Study 1: b = 0.058, $\chi^2(1) = 16.77, p < 0.001$,95% CI = [0.0302, 0.0858], Study 2: b = 0.068, $\chi^2(1) = 15.62, p < 0.001$,95% CI = [0.0343, 0.1017], Study 3: b = 0.064, $\chi^2(1) = 9.16, p = 0.002$,95% CI = [0.0226, 0.1054]). In the advantageous inequity condition, however, a significant effect of split is found only under reduced cognitive load, in Study 3: (Study 1: b = 0.018, $\chi^2(1) = 1.53, p = 0.216$,95% CI = [-0.0105, 0.0465], Study 2: b = -0.008, $\chi^2(1) = 0.38, p = 0.538$,95% CI = [-0.0334, 0.0174], Study 3: b = 0.046, $\chi^2(1) = 4.45, p = 0.035$,95% CI = [0.0033, 0.0887]). Through trial-by-trial comparison of the computational models (Figure 3), we confirmed in Studies 1 & 2 that the inequity-weighted model outperforms the baseline model more so in the disadvantageous condition (Study 1: $t(57.3)$ = -3.43, $p = 0.001$, d = 0.713, 95%CI = [-1.063, -0.363], Study 2: $t(36.2)$ = -2.30, $p = 0.027$), d = -0.492, 95%CI = [-0.935, -0.049]. The comparison in Study 3 is not significant, but may be due to the reduced power as a result of fewer split conditions ($t(50.7)$ = -1.51, $p = 0.137$, d = -0.310, 95%CI = [-0.714, 0.094]). Model simulation also showed that the best-fitting model (SPRL) was able to qualitatively reproduce the effect of split percentage (Figure 2, 4) but the baseline models were not able to reproduce the effect (Figure S2).

3.3 Effects of inequity arise early during learning and grow as learning continues

Through model comparison in Studies 1 & 2, we see that the inequity-weighted model outperforms the baseline model as early as the 4th iteration of the stimulus (Study 1: $t(93)$ = 3.54, $p < 0.001$, d = 0.365, 95%CI = [0.158, 0.572], Study 2: $t(90)$ = 2.25, $p = 0.027$, d = 0.236, 95%CI = [0.026, 0.446]). The effect also increases over time (Figure 3). We tested this using a *mixed-effect linear regression* with the model-fit metrics as the dependent variable (equation 4, Table S11-S12). In both Studies 1 & 2, we see a significant positive main effect of stimulus iteration (Study 1: b = 0.004, $\chi^2(1) = 86.25, p < 0.001$,95% CI = [0.0032, 0.0048], Study 2: b = 0.002, $\chi^2(1) = 49.42, p < 0.001$,95% CI = [0.0014, 0.0026]), suggesting the inequity-weighted model outperforms the baseline more in later learning trials. Additionally, the effect is stronger under disadvantageous inequity, as suggested by the significant interaction effect (Study 1: b = -0.004, $\chi^2(1) = 79.48, p < 0.001$,95% CI = [-0.0049, -0.0031], Study 2: b = -0.002, $\chi^2(1) = 42.54, p < 0.001$),95% CI = [-0.0026, -0.0014].

3.4 Social partner identity impacts learning more systematically when cognitive load is reduced

We examined whether the perceived warmth and competence of the social partner affected how the social distribution of reward influenced learning. In Study 1 and Study 2, we observed mixed evidence. On the one hand, the full SPRL model, which integrates social perception ratings, prevailed as the best-fitting computational model (Figure 3; Table S2). On the other hand, the fitted weight parameter of perceived warmth and competence was not significantly different from 0 (Figure 3; Table S1). In Study 1 and Study 2, these coefficients were not significantly different from 0 in the advantageous condition (Study 1: $t_{competence}(40) = -0.736$, $p = 0.466$, d = -0.115, 95%CI

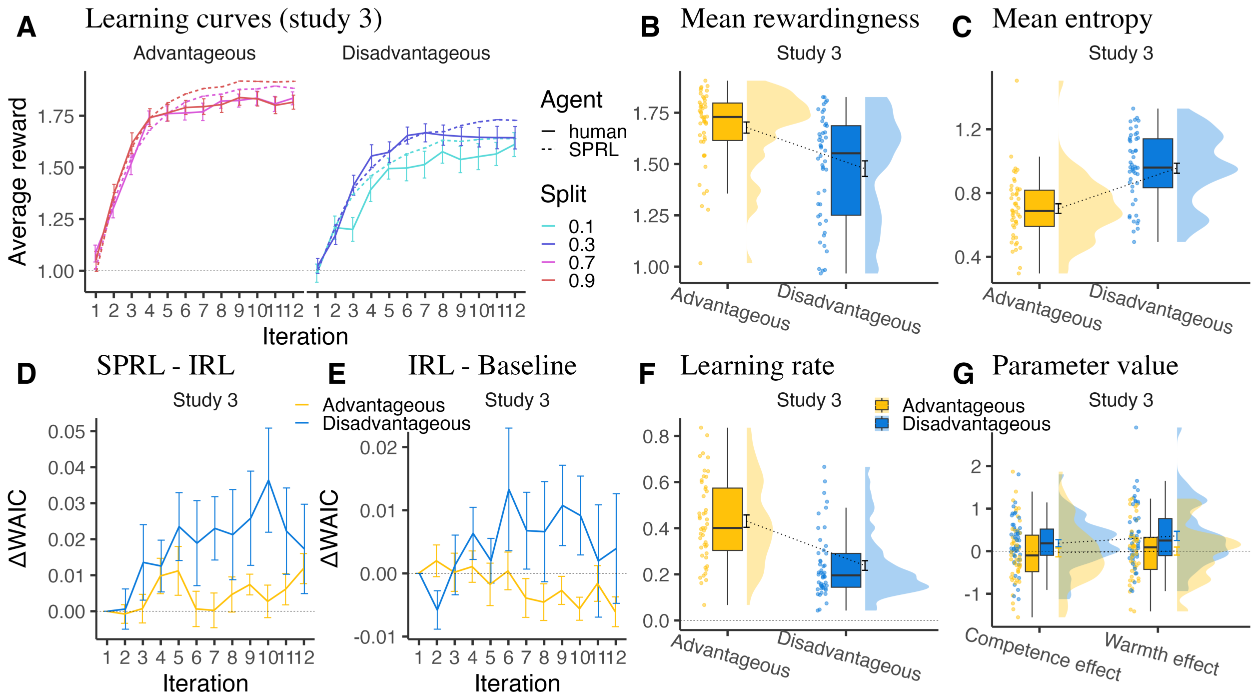

Figure 4: Results of Study 3 (n = 47 in advantageous condition and n = 48 in disadvantageous condition): A) Learning curve. B, C) Overall rewardingness of actions and action entropy averaged within the advantageous and disadvantageous conditions. Replicating Study 1 and Study 2, people learned worse under disadvantageous inequity. D, E) The trial-by-trial model comparisons. The SPRL model significantly outperformed the models without considering social perception ratings, confirming the elevated effect of social perception in Study 3. Moreover, the effect is also stronger under the disadvantageous condition. F, G) Fitted learning rate of the *baseline model* and effects of social perception of the SPRL model: replicating Study 1 and 2, the learning rate was lower under disadvantageous inequity. However, we saw a significant positive effect of both perceived warmth and perceived competence in the disadvantageous condition, suggesting that the effect of social perception enhanced under smaller cognitive load during learning. Each dot represents a point estimate for one subject obtained as the mean of the posterior distribution of the parameters. All error bars reflect the s.e.m.

= [-0.423, 0.193]; $t_{warmth}(40) = 1.45$, $p = 0.155$, d = 0.226, 95%CI = [-0.085, 0.537], Study 2: $t_{competence}(54) = 0.320$, $p = 0.75$, d = 0.043, 95%CI = [-0.222, 0.308]; $t_{warmth}(54) = 0.502$, $p = 0.618$, d = 0.068, 95%CI = [-0.197, 0.333]) or the disadvantageous condition (Study 1: $t_{competence}(52) = 0.050$, $p = 0.96$, d = 0.007, 95%CI = [-0.263, 0.277]; $t_{warmth}(52) = 0.296$, $p = 0.768$, d = 0.041, 95%CI = [-0.230, 0.312], Study 2: $t_{competence}(35) = -0.570$, $p = 0.572$, d = -0.095, 95%CI = [-0.423, 0.233]; $t_{warmth}(35) = 1.52$, $p = 0.136$, d = 0.253, 95%CI = [-0.077, 0.583]).

However, in Study 3 where the cognitive load is reduced (Figure 4), we found a significant positive effect of both perceived warmth and perceived competence (Table S1) in the disadvantageous condition on the rewardingness of actions ($t_{competence}(47) = 2.25$, $p = 0.029$, d = 0.325, 95%CI = [0.038, 0.612]; $t_{warmth}(47) = 3.34$, $p = 0.002$, d = 0.482, 95%CI = [0.194, 0.770]). Similarly, *mixed-effect linear regression* (equation 1, Table S5-S6) also revealed a significant main effect of both perceived warmth and perceived competence on the rewardingness of actions during learning ($b_{competence} = 0.031$, $\chi^2(1) = 12.10$, $p < 0.001$,95% CI = [0.0135, 0.0485]; $b_{warmth} = 0.042$, $\chi^2(1) = 25.52$, $p < 0.001$,95% CI = [0.0257, 0.0583]). The overall effect of perceived warmth was stronger in the disadvantageous condition ($b = -0.069$, $\chi^2(1) = 69.34$, $p < 0.001$,95% CI = [-0.0852, -0.0528]). The overall effect of perceived competence was also stronger in the disadvantageous condition ($b = -0.028$, $\chi^2(1) = 9.77$, $p = 0.002$,95% CI = [-0.0456, -0.0104]). Notably, these effects remained significant when we ran the same regression on simulated data from the SPRL model with the fitted parameter values (Table S13-S14). Moreover, they disappeared when using data simulated by the IRL model, which ignored information about perceived warmth and competence (Table S13-S14).

Similar to the effect of inequity, the effect of social perception on learning also increases over the course of learning, but the increase is stronger under disadvantageous inequity. We support this again with *mixed-effect linear regression* on the model comparison metrics (equation 4, Table S11-S12) where we see a positive main effect of stimulus iteration ($b = 0.001$, $\chi^2(1) = 14.65$, $p < 0.001$,95% CI = [-0.0852, -0.0528]) and also a significant effect of its interaction with inequity type ($b = -0.0008$, $\chi^2(1) = 4.53$, $p = 0.033$,95% CI = [-0.0015, -0.0001]). One possible interpretation of these results is that the perceived warmth and competence did shape how inequity affects learning across all three studies, but individual differences in the nature of this effect precluded its detection in model-free, group-level analyses in Study 1 and 2. In Study 3, where cognitive load was reduced, we found evidence for a more systematic effect of social perception on learning at the group level.

4 Discussion

People often rely on learned reward contingencies to guide their decisions, making factors that impact the learning process important precursors to decisions. Through a reinforcement learning task, we found that inequity of the distribution of reward across oneself and another person, as well as the identity of that person, shaped people's ability to learn from rewards.

First, people learned faster and more successfully overall when they received the larger share of the reward (compared to the other person) than when they received the smaller share of the reward (compared to the other person), even controlling for the overall reward size to self (Study 2). This result is especially potent given that we manipulated inequity type between-subjects, ruling out the possibility that participants could have contrasted different inequity conditions and adjusted their internal learning incentives accordingly. Moreover, this shows that the impact of inequity (advantageous vs disadvantageous) on valuation during learning does not show range adaptation to the possible range of the degree of inequity (Bavard et al., 2021). In other words, disadvantageous inequity decreases the value of the rewards without needing a separate reference condition in which the same participants experience advantageous inequity. In this way, the effect of disadvantageous versus advantageous inequity on learning can be thought of as "absolute" rather than "relative".

Second, people were more sensitive to the specific percentage of the reward given to themselves when that percentage was less than (compared to when it was more than) 50%. For example, the difference between receiving 20% and 40% of the reward (a 20% difference) was greater, in terms of its impact on learning, than the difference between receiving 60% and 80% (also a 20% difference). This could be because disadvantageous inequity prompts people to be more sensitive to the split percentage. We note that this finding is especially striking as it was replicated in Study 2, in which impaired learning under disadvantageous inequity cannot be explained by the amount of reward personally received by the participant because that amount was held constant. In this study, more total reward (across the participant and the partner) is actually generated in the disadvantageous condition than in the advantageous condition, yet participants still learn less effectively in this condition. In other words, participants' learning seems to be more driven by social comparison than by total welfare during the learning task. Future studies could explore a within-subject manipulation on the design of Study 2, where a tradeoff between inequity and total reward exists. It would be especially interesting to see to test the possibility that split percentage may have a non-continuous effect on learning across the range from 0% to 100%.

Third, a more systematic effect of social perception on learning emerged when cognitive load was reduced. We saw a significantly positive effect of perceived warmth and competence on learning under the disadvantageous condition only in Study 3, where the stimulus space was reduced from 5 to 2. Because the parameter recovery result in Study 3 did not seem to differ substantially from Studies 1 and 2 (Figure S5,S6,S7), it is unlikely that this difference arose because the social perception weight parameters were harder to recover in Studies 1 and 2 compared to Study 3. Moreover, it is worth noting that the SPRL model, which incorporates social perception as well as inequity information, emerged as the best fitting model to participants' behavior even in Studies 1 and 2. This may suggest that while perceived warmth and competence have somewhat idiosyncratic effects on learning across individuals when the task is especially taxing (perhaps due to different reliance on heuristics and/or different levels of working memory capacity), these effects are more systematic when cognitive load is reduced. These possible interpretations are preliminary and open to more direct investigation in future research.

We would like to highlight some broader implications of this study. First, the observation that disadvantageous inequity hampers learning in an absolute sense—i.e., without requiring direct comparison to advantageous inequity—speaks to the potential importance of the current findings in ecological settings, where a given instance of learning from shared rewards is likely to be characterized by a single type of inequity. Second, the finding that social contextual information shapes reinforcement learning adds on to the body of evidence that the reinforcement learning system, despite long being considered a low-level implicit cognitive system, is impacted by higher-level cognition – in the case, social cognition (Collins & Frank, 2012; Ham, McDougle, & Collins, 2024; Master et al., 2020). Third, our studies contribute to a growing trend toward integrating models of cognition and models of economic behavior (Andrade, Gaballo, Mengus, & Mojon, 2019; Andre, Pizzinelli, Roth, & Wohlfart, 2022; Barberis & Jin, 2023; Jenkins et al., 2018; Kobayashi et al., 2022). We tested experimentally how the distribution of rewards impacts reward values relevant to learning and designed formal models that make it possible to characterize this impact.

Finally, these findings point to a possible gap between subjective value generated social decision-making and social learning that warrants further investigation (Barron & Erev, 2003; Garcia, Cerrotti, & Palminteri, 2021; Hertwig, Barron, Weber, & Erev, 2004; Hertwig & Erev, 2009; Martin, Gonzalez, Juvina, & Lebiere, 2014). Although research on social decision-making has shown that others' perceived warmth and competence have different effects on people's equity preferences (Jenkins et al., 2018), the current studies found that perceived warmth and competence both had the same positive effect on reinforcement learning. This gap may be due to how group processes such as social status impact valuations differently during learning than descriptive decision-making. Future research is needed to further investigate how contextual effects on valuation during social decision-making relate to contextual effects on learning from shared rewards.

4.1 Limitations

One potential limitation of this study is the hypothetical nature of the monetary reward as well as the social partner. Participants were asked to imagine that the rewards were real and that a portion was actually given to another person. Research comparing real to hypothetical monetary rewards generally finds consistent patterns of behavior across the two contexts (Kühberger, Schulte-Mecklenbeck, & Perner, 2002; Wiseman & Levin, 1996), and when they differ, hypothetical contexts typically have smaller effects (Camerer & Hogarth, 1999). In particular, although there is sometimes an overall shift from hypothetical to real rewards (e.g., in mean levels of generosity), manipulated factors (e.g., inequity, target identity) typically have similar effects in both contexts (Jenkins et al., 2018; Kobayashi et al., 2022). We also acknowledge that all participants are young adults in the United States, leaving open questions about the degree to which findings from our sample generalize to people situated within different socioeconomic systems.

5 Data availability

All experiment materials, data, and analysis code are publicly available at `https://osf.io/xcwqd/?view_only=02df3a86b50f488a9feb285ff3d0ac93`.

6 Code availability

All experiment materials, data, and analysis code are publicly available at `https://osf.io/xcwqd/?view_only=02df3a86b50f488a9feb285ff3d0ac93`.

References

Andrade, P., Gaballo, G., Mengus, E., & Mojon, B. (2019). Forward guidance and heterogeneous beliefs. *American Economic Journal: Macroeconomics*, *11*(3), 1–29.

Andre, P., Pizzinelli, C., Roth, C., & Wohlfart, J. (2022). Subjective models of the macroeconomy: Evidence from experts and representative samples. *The Review of Economic Studies*, *89*(6), 2958–2991.

Barberis, N. C., & Jin, L. J. (2023). *Model-free and model-based learning as joint drivers of investor behavior* (Tech. Rep.). National Bureau of Economic Research.

Baribault, B., & Collins, A. G. E. (2025). Troubleshooting Bayesian cognitive models. *Psychological Methods*, *30*(1), 128–154. (Epub 2023 Mar 27) doi: 10.1037/met0000554

Barnby, J. M., Raihani, N., & Dayan, P. (2022, August). Knowing me, knowing you: Interpersonal similarity improves predictive accuracy and reduces attributions of harmful intent. *Cognition*, *225*, 105098. doi: 10.1016/j.cognition.2022.105098

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of behavioral decision making*, *16*(3), 215–233.

Bavard, S., Lebreton, M., Khamassi, M., Coricelli, G., & Palminteri, S. (2018, October). Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. *Nature Communications*, *9*(1), 4503. doi: 10.1038/s41467-018-06781-2

Bavard, S., Rustichini, A., & Palminteri, S. (2021). Two sides of the same coin: Beneficial and detrimental consequences of range adaptation in human reinforcement learning. *Science Advances*, *7*(14), eabe0340.

Bhanji, J. P., & Delgado, M. R. (2014, January). The social brain and reward: social information processing in the human striatum. *Wiley Interdisciplinary Reviews. Cognitive Science*, *5*(1), 61–73. doi: 10.1002/wcs.1266

Brown, V. A. (2021, January). An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920960351. doi: 10.1177/2515245920960351

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of risk and uncertainty*, *19*, 7–42.

Charpentier, C. J., & O'Doherty, J. P. (2018, November). The application of computational models to social neuroscience: promises and pitfalls. *Social Neuroscience*, *13*(6), 637–647. doi: 10.1080/17470919.2018.1518834

Christopoulos, G. I., & King-Casas, B. (2015). With you or against you: Social orientation dependent learning signals guide actions made for others. *NeuroImage*, *104*, 326-335. doi: https://doi.org/10.1016/j.neuroimage.2014.09.011

Collins, A. G., Ciullo, B., Frank, M. J., & Badre, D. (2017, April). Working Memory Load Strengthens Reward Prediction Errors. *The Journal of Neuroscience*, *37*(16), 4332–4342. doi: 10.1523/JNEUROSCI.2700-16.2017

Collins, A. G., & Frank, M. J. (2012, April). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024–1035. doi: 10.1111/j.1460-9568.2011.07980.x

Eckstein, M. K., Master, S. L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. (2022). Reinforcement learning and bayesian inference provide complementary models for the unique advantage of adolescents in stochastic reversal. *Developmental Cognitive Neuroscience*, *55*, 101106.

Fehr, E., & Camerer, C. F. (2007, October). Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences*, *11*(10), 419–427. doi: 10.1016/j.tics.2007.09.002

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007, February). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. doi: 10.1016/j.tics.2006.11.005

Frank, M. J., Seeberger, L. C., & O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, *306*(5703), 1940–1943.

Garcia, B., Cerrotti, F., & Palminteri, S. (2021, March). The description–experience gap: a challenge for the neuroeconomics of decision-making under uncertainty. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1819), 20190665. doi: 10.1098/rstb.2019.0665

Goncharova, E. Y., & Jenkins, A. C. (submitted). Spontaneous future imagination shapes decisions on behalf of others.

Hackel, L. M., Mende-Siedlecki, P., Loken, S., & Amodio, D. M. (2022). Context-dependent learning in social interaction: Trait impressions support flexible social choices. *Journal of personality and social psychology*, *123*(4), 655.

Hackel, L. M., Zaki, J., & Van Bavel, J. J. (2017, August). Social identity shapes social valuation: evidence from prosocial behavior and vicarious reward. *Social Cognitive and Affective Neuroscience*, *12*(8), 1219–1228. doi: 10.1093/scan/nsx045

Ham, H., McDougle, S. D., & Collins, A. G. (2024). Dual effects of dual-tasking on instrumental learning. *Available at SSRN 4801088*.

Heerey, E. A. (2014, February). Learning from social rewards predicts individual differences in self-reported social ability. *Journal of Experimental Psychology. General*, *143*(1), 332–339. doi: 10.1037/a0031511

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological science*, *15*(8), 534–539.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*, 517–523. doi: https://doi.org/10.1016/j.tics.2009.09.004

Hertz, U., Bell, V., & Raihani, N. (2021). Trusting and learning from others: immediate and long-term effects of learning from observation and advice. *Proceedings of the Royal Society B*, *288*(1961), 20211414.

Jenkins, A. C. (2019). Rethinking cognitive load: a default-mode network perspective. *Trends in Cognitive Sciences*, *23*(7), 531–533.

Jenkins, A. C., Karashchuk, P., Zhu, L., & Hsu, M. (2018, September). Predicting human behavior toward members of different social groups. *Proceedings of the National Academy of Sciences*, *115*(39), 9696–9701. doi: 10.1073/pnas.1719452115

Jones, R. M., Somerville, L. H., Li, J., Ruberry, E. J., Libby, V., Glover, G., … Casey, B. J. (2011, September). Behavioral and Neural Properties of Social Reinforcement Learning. *Journal of Neuroscience*, *31*(37), 13039–13045. doi: 10.1523/JNEUROSCI.2972-11.2011

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

Kühberger, A., Schulte-Mecklenbeck, M., & Perner, J. (2002). Framing decisions: Hypothetical and real. *Organizational Behavior and Human Decision Processes*, *89*(2), 1162–1175.

Langdon, A. J., Sharpe, M. J., Schoenbaum, G., & Niv, Y. (2018). Model-based predictions for dopamine. *Current opinion in neurobiology*, *49*, 1–7.

Lindström, B., Selbing, I., Molapour, T., & Olsson, A. (2014, March). Racial Bias Shapes Social Reinforcement Learning. *Psychological Science*, *25*(3), 711–719. doi: 10.1177/0956797613514093

Lockwood, P. L., Apps, M. A. J., Valton, V., Viding, E., & Roiser, J. P. (2016, August). Neurocomputational mechanisms of prosocial learning and links to empathy. *Proceedings of the National Academy of Sciences*, *113*(35), 9763–9768. doi: 10.1073/pnas.1603198113

Martin, J. M., Gonzalez, C., Juvina, I., & Lebiere, C. (2014). A Description–Experience Gap in Social Interactions: Information about Interdependence and Its Effects on Cooperation. *Journal of Behavioral Decision Making*, *27*(4), 349–362. doi: 10.1002/bdm.1810

Master, S. L., Eckstein, M. K., Gotlieb, N., Dahl, R., Wilbrecht, L., & Collins, A. G. (2020). Disentangling the systems contributing to changes in learning during adolescence. *Developmental cognitive neuroscience*, *41*, 100732.

Nafcha, O., & Hertz, U. (2024). Asymmetric cognitive learning mechanisms underlying the persistence of intergroup bias. *Communications Psychology*, *2*(1), 14.

Palminteri, S., & Lebreton, M. (2021, October). Context-dependent outcome encoding in human reinforcement learning. *Current Opinion in Behavioral Sciences*, *41*, 144–151. doi: 10.1016/j.cobeha.2021.06.006

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019, February). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. doi: 10.3758/s13428-018-01193-y

Rescorla, R. (1972). A theory of Pavlovian conditioning : Variations in the effectiveness of reinforcement and nonreinforcement..

Rohde, K. I. (2010). A preference foundation for fehr and schmidt's model of inequity aversion. *Social Choice and Welfare*, *34*, 537–547.

Rustichini, A., Conen, K. E., Cai, X., & Padoa-Schioppa, C. (2017). Optimal coding and neuronal adaptation in economic decisions. *Nature communications*, *8*(1), 1208.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016, April). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55. (Publisher: PeerJ Inc.) doi: 10.7717/peerj-cs.55

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Spektor, M. S., Gluth, S., Fontanesi, L., & Rieskamp, J. (2019, January). How similarity between choice options affects decisions from experience: The accentuation-of-differences model. *Psychological Review*, *126*(1), 52–88. doi: 10.1037/rev0000122

Strombach, T., Weber, B., Hangebrauk, Z., Kenning, P., Karipidis, I. I., Tobler, P. N., & Kalenscher, T. (2015, February). Social discounting involves modulation of neural value signals by temporoparietal junction. *Proceedings of the National Academy of Sciences*, *112*(5), 1619–1624. doi: 10.1073/pnas.1414715112

Sul, S., Tobler, P. N., Hein, G., Leiberg, S., Jung, D., Fehr, E., & Kim, H. (2015, June). Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. *Proceedings of the National Academy of Sciences*, *112*(25), 7851–7856.

Sullivan-Toole, H., Dobryakova, E., DePasque, S., & Tricomi, E. (2019, February). Reward circuitry activation reflects social preferences in the face of cognitive effort. *Neuropsychologia*, *123*, 55–66. doi: 10.1016/j.neuropsychologia.2018.06.011

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 135). MIT press Cambridge.

Suzuki, S., & O'Doherty, J. P. (2020, June). Breaking human social decision making into multiple components and then putting them together again. *Cortex*, *127*, 221–230. doi: 10.1016/j.cortex.2020.02.014

Vélez, N., & Gweon, H. (2019). Integrating Incomplete Information With Imperfect Advice. *Topics in Cognitive Science*, *11*(2), 299–315.

Watanabe, S. (2013). A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, *14*(27), 867–897.

Webb, R., Glimcher, P. W., & Louie, K. (2021). The normalization of consumer valuations: Context-dependent preferences from neurobiological constraints. *Management Science*, *67*(1), 93–125.

Wiseman, D. B., & Levin, I. P. (1996). Comparing risky decision making under conditions of real and hypothetical consequences. *Organizational behavior and human decision processes*, *66*(3), 241–250.

Witt, A., Toyokawa, W., Lala, K. N., Gaissmaier, W., & Wu, C. M. (2024). Humans flexibly integrate social information despite interindividual differences in reward. *Proceedings of the*

National Academy of Sciences, *121*(39), e2404928121. doi: 10.1073/pnas.2404928121

7 Acknowledgments

We thank Beth Baribault, Maria Eckstein, Sunny Liu, Alex Witt, Harry Wang, and Dilara Berkay for their helpful suggestions and feedback. This work was supported by NSF CAREER award 2339853 to ACJ. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

8 Author information

8.1 Authors and Affiliations

1. **Department of Psychology, University of Pennsylvania**

Adrianna C. Jenkins, Huang Ham

2. **Department of Psychology, Princeton University**

Huang Ham

8.2 Contributions

Huang Ham was responsible for project conceptualization, experimental design, data collection, data processing and analysis, and manuscript preparation. Adrianna C. Jenkins was involved in conceptualization, design, funding acquisition, and manuscript preparation.

8.3 Corresponding author

Correspondence to Huang Ham

9 Ethics declarations

9.1 Competing interests

The authors declare no competing interests.