COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13326

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13326

Social Expectations are Primarily Rooted in Reciprocity: An Investigation of Fairness, Cooperation, and Trustworthiness

Paul C. Bogdan,[a,b,*] Florin Dolcos,[a,b,c,*] Matthew Moore,[a] Illia Kuznietsov,[a,d] Steven A. Culpepper,[a,e] Sanda Dolcos[a,b]

[a]*Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign*
[b]*Department of Psychology, University of Illinois at Urbana-Champaign*
[c]*Neuroscience Program, University of Illinois at Urbana-Champaign*
[d]*Department of Human and Animal Physiology, Lesya Ukrainka Volyn National University*
[e]*Department of Statistics, University of Illinois at Urbana-Champaign*

Received 10 March 2023; received in revised form 13 June 2023; accepted 16 July 2023

Abstract

Social expectations guide people's evaluations of others' behaviors, but the origins of these expectations remain unclear. It is traditionally thought that people's expectations depend on their past observations of others' behavior, and people harshly judge atypical behavior. Here, we considered that social expectations are also influenced by a drive for reciprocity, and people evaluate others' actions by reflecting on their own decisions. To compare these views, we performed four studies. Study 1 used an Ultimatum Game task where participants alternated Responder and Proposer roles. Modeling participants' expectations suggested they evaluated the fairness of received offers via comparisons to their own offers. Study 2 replicated these findings and showed that observing selfish behavior (lowball offers) only promoted acceptance of selfishness if observers started acting selfishly themselves. Study 3 generalized the findings, demonstrating that they also arise in the Public Goods Game, emerge cross-culturally, and apply to antisocial punishment whereby selfish players punish generosity. Finally, Study 4 introduced the Trust Game and showed that participants trusted players who reciprocated their behavior, even if it was selfish, as much as they trusted generous players. Overall, this research shows that social expectations and evaluations are rooted in drives for reciprocity. This carries theoretical

*Shared first-authorship.
Correspondence should be sent to Paul C. Bogdan, Florin Dolcos, and Sanda Dolcos; SCoPE Neuroscience Lab (https://dolcoslab.beckman.illinois.edu/), Beckman Institute for Advanced Science & Technology, University of Illinois at Urbana-Champaign, 405 North Mathews Avenue, Urbana, IL 61801, USA. E-mail: pbogda2@illinois.edu, fdolcos@illinois.edu, and sdolcos@illinois.edu

implications, speaking to a parallel in the mechanisms driving both decision-making and social evaluations, along with practical importance for understanding and promoting cooperation.

Keywords: Economic games; Prediction error; Theory of Mind; Social learning; Conformity

1. Introduction

Much of humanity's success comes from our ability to cooperate (Henrich, 2017), which permits achievements far beyond what is possible from working alone. However, to maintain this cooperation, people must continuously evaluate the social behavior of others, assessing whether it is fair and identifying who should be trusted (Fehr & Schurtenberger, 2018; Gächter, Herrmann, & Thöni, 2004). While social evaluations can be complex, converging evidence suggests that they are influenced by our social expectations, and individuals negatively evaluate those who violate these expectations. This cognitive system is effective and flexible, but when it goes haywire, the consequences can be dire, promoting punishment and distrust toward others whose actions may simply be unfamiliar. Hence, it is critical to clarify the nature of expectations and how they form.

Expectations have long been a primary focus of social cognition research, due to their close link to social norms (Horne & Mollborn, 2020). *Expectations* are the subjective standards that individuals use to evaluate the appropriateness of another person's behavior (i.e., how one determines right and wrong when judging others). Most recent research argues that expectations depend on descriptive norms, such that individuals expect the behavior that they have previously encountered (Eriksson, Strimling, & Coultas, 2015; Goldring & Heiphetz, 2020; Hetu, Luo, D'Ardenne, Lohrenz, & Montague, 2017; Irwin & Horne, 2013; Kawamura & Kusumi, 2020; Lindström, Jangard, Selbing, & Olsson, 2018; Xiang, Lohrenz, & Montague, 2013). Under this lens, atypical behavior elicits expectation-violations, which encourage negative evaluations and punishments. However, expectations also depend on individuals' own tendencies to behave generously or selfishly. For example, participants who behave selfishly/competitively in economic games generally predict that others will do the same (Kelley & Stahelski, 1970b; Van Lange, 1992). Under this alternative view, expectation-violations arise from deviations relative to the evaluator's own behavior.

No study has yet directly compared these perspectives to investigate whether individuals' expectations depend more on their past observations of others' behavior or on their own behavior. We sought to address this issue, which is illustrated by possible scenarios in response to the following question: What happens when a generous person frequently observes selfish behavior? Will they begin to expect selfishness and accept others' selfish behavior? If evaluations primarily involve comparisons with expectations based on previous observations of others' selfish behavior, the answer is "yes," but if evaluations primarily involve comparisons with their own generous behavior, the answer is "no." We refer to these two perspectives as the "Observational" and "Experiential" views (see Fig. 1). Our present research compared these views. Specifically, we tested whether participants' previous observations or their own behavior contribute more to their subjective thresholds of when an

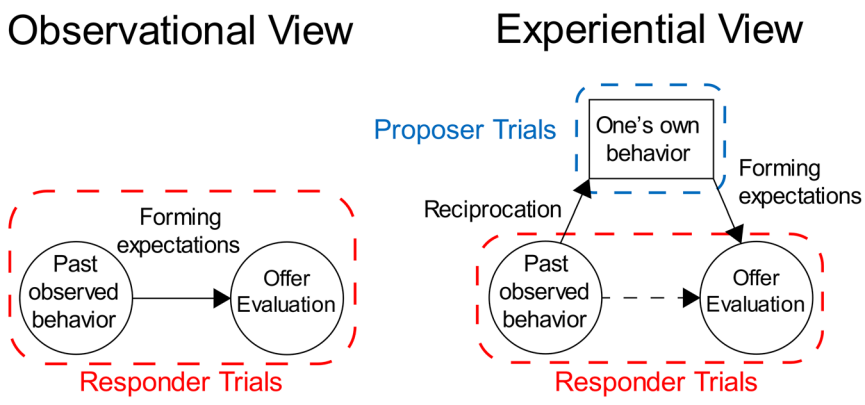

Fig. 1. Two views on how expectations are formed, linked to Ultimatum Game roles. Both views agree that expectations influence fairness evaluations and hence players' choices to accept/reject offers. However, the two models differ in how they posit that expectations are formed. The *Observational view* considers that participants expect to receive the same offers that they encountered in past Responder trials and will punish atypical behavior. On the other hand, the *Experiential view* suggests that players expect to receive the offer that they themselves have made and will punish dissimilarity. This latter view holds that past observations may impact offer evaluations but only insofar as they also prompt reciprocation and impact participants' Proposer behavior.

expectation is violated and examined which factor bears greater weight on whether a behavior is evaluated negatively (e.g., punished and distrusted).

One aspect that complicates this question is that observing selfish behavior influences people's tendencies to behave selfishly themselves—a process often referred to as "reciprocity" or "assimilation" (Bardsley & Sausgruber, 2005; Kelley & Stahelski, 1970b). Thus, there may be an indirect effect of previous observations on expectations, mediated by changes in people's own selfish/generous tendencies. The generous person may punish selfishness initially, but then may gradually conform to those around them and possibly become selfish themselves. In turn, this may motivate them to accept selfishness. In this way, the two perspectives on the origins of expectations are not mutually exclusive. Hence, aside from comparing the two perspectives on expectations, our research also performed tests to integrate these theories. Below, we review the literature supporting our rationale and approach.

1.1. Expectations and observations impact punishment

One economic game that has proven useful for studying evaluations and expectations is the Ultimatum Game (UG). In this two-player game, the *Proposer* decides how to split a pot of money with another player (e.g., take $6 from a $10 pot and give the other player $4). After receiving the offer, the *Responder* chooses whether to accept or reject it. Acceptance causes the money to be distributed as proposed, whereas rejection causes neither player to receive any money. The choice to accept or reject an offer is influenced by both participants' motivations to increase their payout and their evaluations of the offers' fairness. Rejection can be interpreted as an instance where the Responder's affective response to the perceived unfairness of an offer overwhelms the possible earnings from accepting it (Chang & Sanfey,

2013; Grecucci, Giorgetta, Van't Wout, Bonini, & Sanfey, 2013). Rejection is typically also seen as a form of punishment because the Responder assumes a cost to impose a loss on the Proposer.

Participants' fairness evaluations and choices to reject are influenced by their beliefs about typical behavior. For example, participants who believe most people will propose selfishly are less likely to reject selfish offers (Chang & Sanfey, 2013; Vavra, Chang, & Sanfey, 2018). Further, by having participants repeatedly play the Responder role, their expectations can be modeled as a function of their previously received/observed offers, which reveals that expectation-violations relative to these previous offers predict rejection likelihood (see the Observational View in Fig. 1 left; Hetu et al., 2017; Xiang et al., 2013). Similar results emerge across other economic games and social psychology paradigms, which has led to theories that consider the detection of expectation-violation to be a heuristic contributing to social evaluation—see evidence from the Public Goods Game (PGG) on the Common is Moral heuristic (Lindström et al., 2018). However, earlier research examining expectations and the effects of people's previous observations on their later evaluations has not examined their links to participants' own selfish/generous behavior.

1.2. One's own behavior impacts one's expectations

People's tendencies to behave generously themselves also contribute to their expectations (see the Experiential View in Fig. 1 right). Early work on this perspective has been summarized as the Triangle Hypothesis, which posits that cooperative players predict others to be cooperative or competitive, whereas competitive players predict others to all be competitive (Aksoy & Weesie, 2012; Bogaert, Boone, & Declerck, 2008; Kelley & Stahelski, 1970b; Van Lange, 1992). For example, competitive participants tend to see others as behaving competitively, even when others are trying to cooperate (Kelley & Stahelski, 1970a), and participants who behave selfishly in economic games believe others will also behave selfishly (Aksoy & Weesie, 2012). Within-subject effects also arise and occur even among young children, who after behaving generously, predict others will behave generously toward them (Leimgruber, 2018; Myslinska Szarek & Tanas, 2022).

However, it remains unclear whether this perspective on expectations tracks how participants evaluate fairness. In other words, to what degree do evaluators perceive expectation-violation as dissimilarity relative to their own behavior? This can be tested in a manner analogous to the expectation-modeling research noted above. Using the UG and having participants repeatedly play as both Proposers and Responders allows modeling their expectations as a function of their previously proposed offers. Then, analyses can examine the extent to which these modeled expectations predict subsequent rejection likelihood, testing whether participants become more accepting of others' selfish behavior as Responders after they themselves begin to behave selfishly as Proposers. This Proposer-based expectations model can also be compared to the model formulated by earlier UG studies, which defined expectations as a function of the offers received in previous Responder trials (Hetu et al., 2017; Xiang et al., 2013). Testing which expectation model better predicts later rejections would shed insight into the effects of previous observations versus personal behavior on evaluations.

1.3. The role of reciprocity

We must consider that individuals adjust their behavior and selfish/generous tendencies based on the people they interact with. Cooperative and generous individuals become more competitive and selfish after interacting with competitive/selfish players (De Cremer & Van Lange, 2001; Kelley & Stahelski, 1970b). Here, we refer to this process as "reciprocity," but it has also been called "assimilation" and is closely related to conformity, tit-for-tat, and conditional cooperation (Bardsley & Sausgruber, 2005; Fischbacher, Gächter, & Fehr, 2001; Kelley & Stahelski, 1970b). Reciprocity manifests in multiple ways: Participants reciprocate both when they repeatedly interact with the same person (direct reciprocity) or when they are matched with new partners (indirect reciprocity). For example, if Alice treats Bob selfishly, Bob will feel compelled to act selfishly toward Alice, and Bob will also "pay it forward" by acting selfishly toward Charlie (Jung, Seo, Han, Henderson, & Patall, 2020; Nowak & Sigmund, 2005). The drive to reciprocate is seen across the globe and studies on reciprocity regularly show large behavioral effects (Allidina, Arbuckle, & Cunningham, 2019; Curry, Mullins, & Whitehouse, 2019; Wedekind & Milinski, 1996).

These reciprocity-related effects may cause observed selfish/generous behavior to indirectly impact expectations and thus evaluations. That is, observing selfishness may compel people to become selfish themselves (reciprocity), which, in turn, prompts them to expect selfishness and punish selfish behavior less frequently. Notably, this possibility integrates the two perspectives on expectations noted above (Fig. 1). However, in this case, if people refuse to reciprocate and refuse to behave selfishly, then observing selfishness will have little effect on people's willingness to accept selfish behavior.

1.4. The present approach

To clarify how participants evaluate others' behavior, our research examined how participants' expectations shift depending on their previous observations of others' selfish/generous behavior (Observational view) and on their own tendencies to behave selfishly or generously (Experiential view). Study 1 assessed which factor had a larger role by using a version of the UG where participants alternated between the Proposer and Responder roles. Fig. 1 illustrates our analytic strategy, which involved modeling participants' expectations to assess whether previously received offers or previously proposed better predict later rejection likelihood. This study also investigated the possibility that the effect of previous observations on rejection likelihood is mediated by changes in participants' own behavior (i.e., the mediation illustrated in Fig. 1). Study 2 aimed to replicate these initial findings and also tested whether they emerge when participants are paired with particularly selfish players. Study 3 examined whether the patterns identified in the first two studies apply to also understanding punishment in the PGG (Fehr & Gächter, 2002). This study reanalyzed the dataset provided by Herrmann, Thöni, and Gächter (2008), which was notably collected cross-culturally, allowing further investigation of generalizability. Study 4 used the UG but now to examine evaluations of trustworthiness, which unfold over multiple UG trials. This study also manipulated participants' UG partners to test whether interacting with partners who reciprocate their decisions enhances perceived trustworthiness (measured using the Trust Game).[1]

2. Study 1

Study 1 used the role-swap UG as noted just above. Participants' expectations were modeled on a trial-by-trial basis as latent variables, either based on previously received offers (*E[Received]*) or previously proposed offers (*E[Proposed]*). The two models were compared in terms of which better predicted rejection likelihood. Consistent with the idea that the two views are not mutually exclusive, further analyses examined the sequence of trials to test whether changes in participants' Proposer behavior mediate possible effects of previously received offers.

2.1. Methods

2.1.1. Participants

A total of 40 individuals were recruited from the local university and surrounding community for Study 1 ($M_{Age} = 23.1$ [18, 39], $SD_{Age} = 5.3$; 50% female, 50% male). This sample size was chosen because the study recorded electroencephalography (EEG) data (Bogdan et al., 2022), which although not the present focus, influenced the recruitment goals (Boudewyn, Luck, Farrens, & Kappenman, 2018). Of the 40 participants, one was excluded for excessive drowsiness during testing. Additionally, as we primarily focused on within-subject patterns, six participants were excluded because they almost never ($< 3\%$) rejected received offers. As these participants virtually never shifted their Responder behavior, they cannot provide evidence on the within-subject factors influencing rejection likelihood. Earlier UG studies with within-subject goals have also excluded such participants (Fatfouta, Meshi, Merkl, & Heekeren, 2018; Kubota, Li, Bar-David, Banaji, & Phelps, 2013). All participants accepted at least 3% of offers, and thus none were excluded for lacking acceptances. For completeness, we also report the results with all participants included. Post hoc power analyses were conducted for the multilevel modeling analyses using Monte Carlo simulations via the *simr* package (Green & MacLeod, 2016; see our code in Data Availability), and for the across-subject correlation ($r = .37$) using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009). These revealed that the sample did not provide sufficient power (80%, $\alpha = .05$) for all of the results (see results below). Therefore, caveats related to the sample size were addressed in follow-up studies. All participants provided informed consent under a protocol approved by the Institutional Review Board and received monetary compensation.

2.1.2. Task design

Participants played the UG for 384 trials, separated into 48 trials per block. Most trials (88%) involved alternations between the Proposer and Responder roles (Fig. 2). Participants were told that their UG performance would determine their monetary payment but, in reality, all participants received equal pay. Participants were told that they would change partners throughout the task, but the instructions did not suggest that they were playing with 384 people, as this would be unrealistic. Participants were instructed that they were playing with a large group and that their decisions would not meaningfully impact others' behavior in future

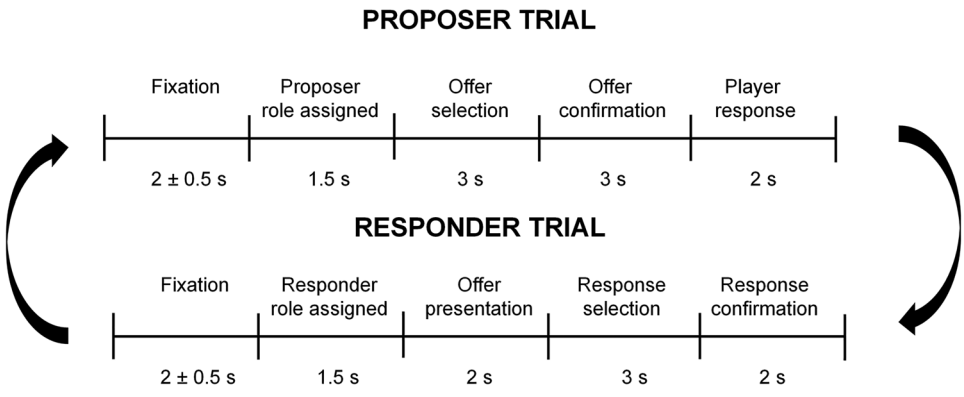

Fig. 2. Task diagram showing the Role-Switch Ultimatum Game procedure. Participants alternated between playing as Proposers and Responders, which involved them deciding what offers to propose and whether to accept or reject received offers, respectively.

trials. After the task, questionnaires were administered, but these are beyond the present focus (Supplementary Materials 1.1).

For the *Proposer* role trials, participants decided how to split a $10 pot; either $5:$5 (equity), $6:$4, $7:$3, $8:$2, or $9:$1 (maximal selfishness). Following a short delay, participants were informed of whether their partner accepted or rejected the offer. For the *Responder* role trials, participants received offers and could respond with either *Strongly Reject, Reject, Pass, Accept*, or *Strongly Accept*. Participants were told to select among these options based on the degrees of their responses, ranging from somewhat sure to extremely sure. *Pass* was included so that participants had an equal number of response options for the Proposer and Responder roles and to help avoid moving their hands across trials. This was particularly relevant for the EEG collection in this study, which we did not focus on here. *Pass* was described as an option for when participants were totally unsure of what to select, and this response option was similar to giving no response, as both led to neither player receiving money for a given trial. *Pass* was selected in only 3% of the trials and their inclusion in the analyses does not significantly affect the results. The "Strongly" versus "not-Strongly" distinction did not impact earnings. Analyses collapsed *Strongly Reject* and *Reject*, collapsed *Strongly Accept* and *Accept*, and excluded *Pass* trials. However, see Supplementary Materials 1.2 for additional tests distinguishing Strongly Reject versus Reject and Accept versus Strongly Accept, showing how the conclusions below also apply to understanding the magnitude of negative evaluations. Every included participant responded, in time, to at least 96% of trials. Trials with missed responses were excluded from the analysis. Like several previous UG studies (e.g., Chang & Sanfey, 2013; Fatfouta et al., 2018; Xiang et al., 2013), participants were told they were playing with other people, but they actually played with a computer. The computer's behavior emulated human play from earlier UG studies (Oosterbeek, Sloof, & Van de Kuilen, 2004). When the participant played as Proposer, the computer rejected offers of $5:$5 in 1% of trials, $6:$4 in 12% of trials, $7:$3 in 45% of trials, $8:$2 in 67% of trials, and $9:$1 in 75% of trials. When the participant played as Responder, they received offers of $5:$5 in

33% of trials, $4:$6 in 17% of trials, $3:$7 in 17% of trials, $2:$8 in 17% of trials, and $1:$9 in 17% of trials.

2.1.3. Expectation modeling

The Observational and Experiential views were compared by modeling expectations on a trial-by-trial basis. To represent the Observational view, expectations were calculated as the mean of previously received offers (*E[Received]*). To represent the experiential view, expectations were calculated as the mean of previously proposed offers (*E[Proposed]*). These expectation variables were calculated using only the previous trials within the same block, which increases variability relative to using all trials across the whole task. Variability could alternatively be induced via a temporal weighting procedure, whereby more recent trials contribute more to expectations than ones farther away, which was done as a confirmatory analysis (Supplementary Materials 1.3).

After calculating these two latent variables for each trial, their predictive powers were compared using multilevel logistic regressions, predicting rejection likelihood, using *E[Received]* and *E[Proposed]* as predictors. Because participants' expectations are just one factor that contributes to their decision to punish (Chang & Sanfey, 2013), two other variables were included in the regression as covariates: (1) the offer amount they received, as selfish offers are rejected more often; (2) whether participants' own offers were just rejected in the previous trial, as this also increases participants' rejection likelihood (i.e., encourages "counter-punishment," Denant-Boemont, Masclet, & Noussair, 2007; Nikiforakis, 2008). In sum, the regression predicted rejection likelihood using four predictors: (i) *E[Received]*, (ii) *E[Proposed]*, (iii) the offer amount, (iv) and what response participants previously encountered.

Notably, by using regression to investigate these phenomena, this accounts for potential collinearity between *E[Received]* and *E[Proposed]* when assigning statistical significance (such collinearity leads to greater standard error associated with their coefficients). Additionally, the inclusion of random intercepts accounts for differences in participants' average rejection likelihoods. The regression excluded trials where participants received $5:$5 offers, as $5:$5 offers were virtually never rejected. For completeness, an across-subject correlation was also performed, testing the link between the average amount that participants proposed and their average likelihood of rejecting offers of $3 or less.

2.1.4. Temporal mediation

To understand the temporal flow of forming social expectations, multilevel regressions were performed to measure links between single trials. First, a multilevel linear regression tested whether the amount participants received in Responder trial[n-2] influenced the amount they proposed in Proposer trial[n-1] (i.e., reciprocity; Fig. 3A). This regression controlled for whether participants accepted/rejected the trial[n-2] offer, as it impacts the amount they subsequently propose. Then, a second multilevel regression tested whether the amount participants proposed in trial[n-1] predicted their rejection likelihood in trial[n] (Fig. 3B). This regression controlled for the offer amount received in trial[n] and whether the offer participants proposed in trial[n-1] was accepted/rejected. Mediation was then tested, linking trial[n-2], trial[n-1], and trial[n] (Fig. 3C), using Monte Carlo simulations (Selig & Preacher, 2008).

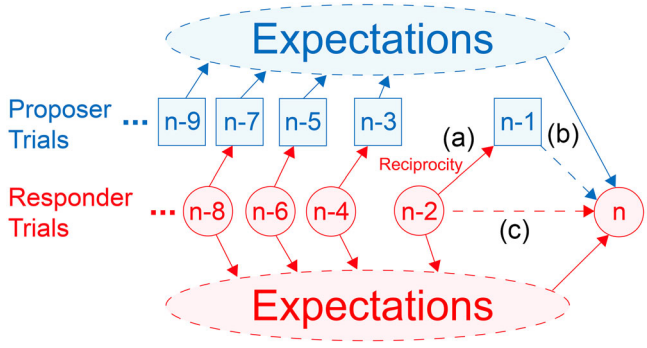

Fig. 3. Illustration of the relations tested in the present research. The left side represents chronologically more distant trials, and the right side represents more recent trials. (a) The reciprocity/conformity link reflects how offers received in the Responder trials are expected to impact subsequent Proposer behavior. (b; blue dashed line) The amount proposed in trial[n-1] is expected to predict rejection likelihood in trial[n] and mediate (c; red dashed line) an indirect effect of the offer amount received in trial[n-2] on rejection likelihood in trial[n]. For consistency, the present work always describes trial[n-2] as being a Responder trial, trial[n-1] as being a Proposer trial, and trial[n] as being the current Responder trial.

2.1.5. *Software and modeling details*

Statistics were performed using R (R Core Team, 2013). Multilevel regressions were fit using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2014) and adhered to best practices, including the use of a maximal random effects structure (Meteyard & Davies, 2020). To avoid convergence errors, predictors were mean-centered and scaled before analysis.

2.2. *Results*

First, descriptive statistics were assessed to validate that this role-swap version of the UG yielded results similar to earlier UG designs focusing on a single role (see meta-analysis by Oosterbeek et al., 2004). As expected, the participants showed typical rejection likelihoods (\$5:\$5 = 1% rejection rate; \$4:\$6 = 14%; \$3:\$7 = 41%, \$2:\$8 = 72%; \$1:\$9 = 82%; Fig. 4A) and typical proposed offers (M = \$6.3, SD = \$0.45; Fig. 4B).

Modeling expectations revealed that participants' previously proposed offers had a larger bearing on rejection likelihood than their previously received offers. Specifically, *E[Proposed]* significantly predicted rejection likelihood (β = 0.39 [0.03, 0.76]; p = .03), whereas *E[Received]* did not (β = .02 [−0.11, 0.14], p = .82). Additionally, correlations showed that participants who proposed more selfish offers, on average, were also less likely to reject selfish offers, on average (r = −.37 [.03, .63], p = .03; scatterplot shown in Fig. S1). The results in this paragraph are no longer significant when participants who accepted all offers are not excluded (Supplementary Materials 1.4), and post hoc power analyses show the analyses are underpowered (power ≤ 66%). Nonetheless, confirmatory analyses using temporal weighting speak to the robustness of the *E[Proposed]* findings (Supplementary Materials 1.3), and Study 2 sought to replicate these findings with a larger sample.

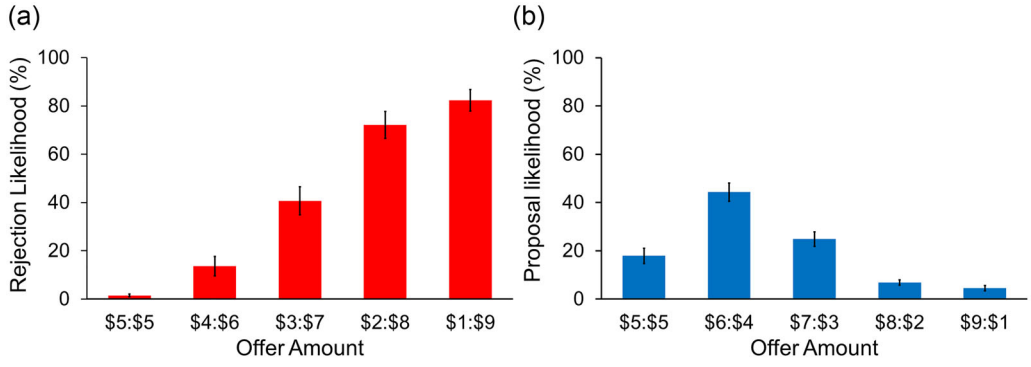

Fig. 4. Basic statistics on Responder and Proposer behavior. The histograms show participants' (a) Rejection likelihoods as Responders and (b) probabilities of offering a given amount as Proposers.

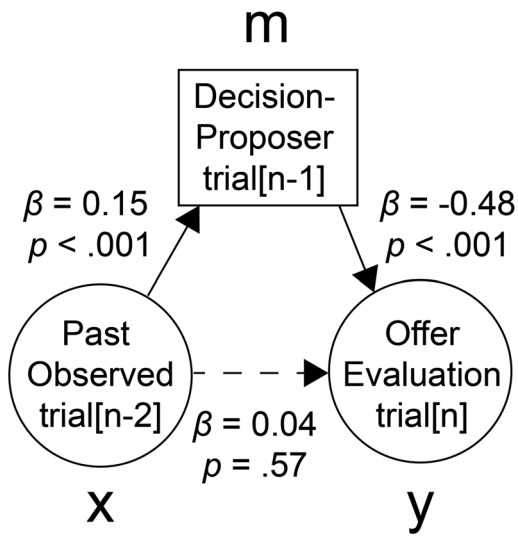

Fig. 5. Participants' decision-making in the Proposer role mediated the effect of past observed offers on offer evaluation. The offer amount that participants proposed in trial[n-1] mediated the effect of the offer received in trial[n-2] on rejection likelihood in trial[n].

Finally, examining the effect of reciprocity showed that it mediated an indirect effect of past received offers on future rejection likelihood (Fig. 5). Receiving a selfish trial[n-2] led to participants proposing more selfishly in trial[n-1] ($\beta = 0.15$ [0.07, 0.23], $p < .001$), and proposing selfishly in trial[n-1] predicted decreased rejection likelihood in trial[n] ($\beta = -0.48$ [−0.28, 0.68], $p < .001$). Taken together, participants were driven to reciprocate the offers they just received, which yielded a significant mediation ($\beta = -0.07$ [−0.13, −0.03]; $p < .001$). These results remain significant even when no participants are excluded, and each analysis was well-powered (power $\geq 98\%$).

2.3. Discussion

Overall, participants directed punishments (rejections) primarily toward offers more selfish than their own, regardless of what offer they previously observed. Hence, participants' expectations are mostly linked to their own selfish/generous tendencies, supporting the Experiential view. However, the present findings do not refute earlier evidence that previous observations (received offers) influence later rejection likelihood (Cooper & Dutcher, 2011; Hetu et al., 2017; Xiang et al., 2013). Rather, the present findings expand previous findings by demonstrating that such effects are mediated by changes in participants' own selfishness/generosity, which in turn influence expectations and rejections.

3. Study 2

Study 2 used a similar role-swap UG design but now included both a Replication condition and a Selfish condition, which were both expected to replicate the initial findings. The Selfish condition examined whether *E[Proposed]* predicts participants' fairness evaluations in situations with economic incentives to accept selfishness while still behaving generously. Finding that *E[Proposed]* remains more predictive in this context would be strong evidence of its validity.

3.1. Methods

3.1.1. Participants

A total of 106 undergraduate students were recruited from the local university for this study (M_{age} = 19.9 [18, 25], SD_{age} = 1.3; 54% female, 28% male, 2% other; 16% no response). The study was collected online due to restrictions related to the COVID-19 pandemic. Data from eight participants were excluded due to limited Responder variability (accepted fewer than 3% of offers or rejected fewer than 3% of offers). This criterion excluded participants with extremely low response rates (e.g., excluding one participant who only responded to two trials and rejected both offers). Excluding these yielded a total of 98 usable participants, 53 in the Replication condition and 45 in the Selfish condition. This sample size was motivated by power analyses (80% power, $\alpha = .05$; Faul et al., 2009; Green & MacLeod, 2016) using the Study 1 data and procedures, which revealed that a minimum of 42 participants would be needed to replicate the effects in each condition. For the across-subject correlation, a minimum sample size of 52 participants is needed, which is surpassed by combining the Replication and Selfish conditions. All participants provided informed consent under a protocol approved by the Institutional Review Board and received course credit in exchange for participation.

3.1.2. Task design

Identical trial protocols were used as in Study 1. However, in this study, participants were randomly assigned to either (1) a Replication condition, where the computer had the same Proposer and Responder behavior as in the first study, or (2) a Selfish condition, where

the computer's behavior was the average behavior of the five most selfish participants from Study 1. Relative to Study 1, two changes were made, for logistical reasons linked to the switch to online data collection: First, the task was shortened to 144 trials, across nine blocks. Second, time-saving incentives were used, rather than monetary payment. Participants were told that their task earnings would determine the length of time they must spend doing a "boring task" after the economic game. The instructions stated that the amount of time they would need to spend could range from between 0 and 15 minutes, but because participants did not genuinely play with other humans, the boring task was never administered. Time-saving incentives have been shown to elicit similar behavior as monetary incentives (Jouxtel, 2019; Noussair & Stoop, 2015). After completing the task, questionnaires were administered, although these are not the focus of the present report (Supplementary Materials 2.1).

3.1.3. Analytic procedure

Study 2 used identical analytic procedures as Study 1, performed separately for the Replication and Selfish conditions. The average participant responded to 93% of trials (10 misses). Although 8% of participants responded to fewer than 75% of trials, this was not an exclusion criterion because participants with fewer responses would simply contribute less to the final multilevel regression results. As with Study 1, Supplementary Materials 2.2 provides results attempting to dissociate *Strongly Reject* versus *Reject* and *Accept* versus *Strongly Accept*, which again showed how the conclusions below apply to understanding the magnitude of negative evaluations.

Unlike Study 1, in this study and Study 4, the success of deception was specifically assessed with a 5-point Likert scale, in which participants were asked about the extent to which they believed they interacted with computers or humans (1 = "Definitely computers"; 5 = "Definitely humans"). For confirmatory purposes, the analyses below were also conducted on the subset of participants who reported believing the deception (score of 3 or higher, $N = 54$). These analyses are reported in Supplementary Materials 2.3, and the results reiterate the conclusions on *E[Proposed]*.

3.2. Results

Each Study 1 finding was replicated. Across the Replication and Selfish conditions, *E[Proposed]* significantly predicted rejection likelihood (Replication: $\beta = 0.51$ [0.10, 0.93], $p = .02$; Selfish: $\beta = 0.72$ [0.38, 1.06]; $p < .001$), whereas *E[Received]* did not (Replication: $\beta = 0.03$ [−0.18, 0.24], $p = .75$; Selfish: $\beta = 0.07$ [−0.11, 0.24]; $p = .47$). Like in Study 1, these patterns also emerged when calculating *E[Proposed]* and *E[Received]* with temporal weighting (Supplementary Materials 2.4). For the pooled data, the across-subject correlation was also replicated: Participants who proposed more selfishly on average across all trials were also less likely to reject selfish offers ($r = -.28$ [−.45, −.08], $p = .003$; Fig. S2).

Likewise, the mediation was replicated: Across both conditions, receiving a selfish offer in trial[n-2] caused participants to propose more selfishly in trial[n-1] (βs > 0.13, ps $< .002$; Fig. 6, x $\rightarrow$ m), and proposing selfishly in trial[n-1] predicted lower rejection likelihood in trial[n] (βs < -0.52; ps $< .004$; Fig. 6, m $\rightarrow$ y). This yielded significant mediations

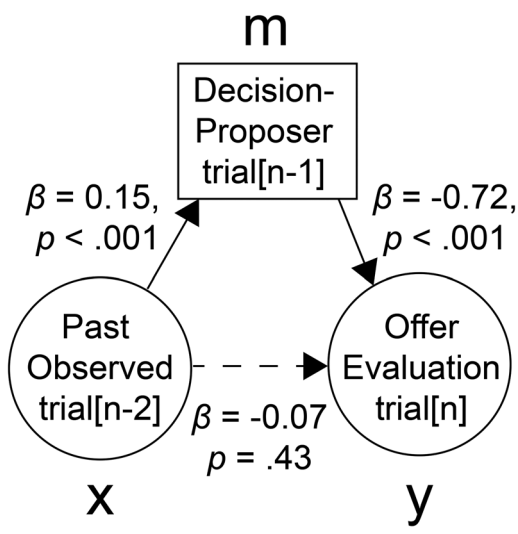

Fig. 6. Replication of the mediation models identified in Study 1. The Study 2 data showed the same mediation between the amount received (trial[n-2]), the amount next proposed (trial[n-1]), and rejection likelihood (trial[n]) originally seen in Study 1. The beta coefficients and *p*-values, here, reflect the pooled data.

(βs < -0.07, ps $< .01$). Notably, even if no participants are excluded, all the Study 2 results remain significant (Supplementary Materials 2.5).

3.3. Discussion

In sum, Study 2 confirms the first study's findings using a larger sample and provides further evidence that participants' own behavior is the primary factor contributing to their expectations. The results confirmed that participants tend to punish behavior that is more selfish than their own, regardless of what they previously received (observed). Additionally, the study reiterates the support for the hypothesized expectation learning pathway whereby receiving (observing) selfish offers prompts participants to propose more selfishly, which in turn makes them more accepting of selfish behavior.

4. Study 3

To further investigate expectations and test the generalizability of the present findings, we reanalyzed a dataset by Herrmann et al. (2008). In their study, participants from different cultures completed the PGG, which is an economic task where four players can cooperate by contributing money to a pot. High contributions benefit the group, although participants would maximize their personal short-term earnings by not contributing. After each round, participants can punish other players. Generally, contributing players use punishment to encourage free-riders to contribute, although Herrmann et al. (2008) showed that antisocial

punishment occurs in some cultures, whereby noncontributors punish contributors seen as excessively generous. We expected that modeling participants' expectations as a function of their own previous contributions would best predict both traditional punishment and antisocial punishment.

4.1. Methods

4.1.1. Participants and task design

Participants ($N = 1120$) were recruited from 16 cities around the globe, including from collectivist cultures and low-income countries (Herrmann et al., 2008). Power analyses based on the Study 2 data show that this sample is highly powered to detect significant effects of *E[Proposed]* (power > 99.9%). Participants played the 4-player PGG for 10 rounds with the same group of players. Each round, participants were given 20 tokens and could contribute tokens to the pot, which would be increased to 160% of their original value and distributed equally (e.g., if every player donates \$20, they all earn \$32). After each round, participants saw how much each other player contributed and could punish specific others by spending tokens. Each token spent would lead to the targeted player losing three tokens. Participants could spend up to 10 tokens for each other player. Participants' earnings, or lack thereof, had no impact on their ability to spend tokens to punish others. Participants were told of punishments directed toward them but were not told who specifically punished them. Participants were also not told of punishments among other pairs of players. Thus, punishment conferred no reputational benefits. Participants also completed a version of the PGG without punishment, although those data are not used here.

4.1.2. Analytic procedure

Each round yielded 12 data points for analysis (i.e., each of the four players' potential punishment toward each of the three other players). *E[Proposed]* was calculated the same as in the UG studies, based on participants' contributions before their choice of whether to exert punishment. Calculating *E[Received]* differed slightly from the UG studies: In the PGG, every player's contribution was shown prior to the punishment phase. Hence, when participant *a* is deciding whether to punish player *b* in round[n], they are aware of player *c*'s and *d*'s contributions in said round. These two players' contributions must be included in *E[Received]*. This differs from the UG calculation of *E[Received]*, which only included received offers before trial[n]. Notably, player *b*'s contribution in round[n] itself should not contribute to *E[Received]* (i.e., a player's contribution does not impact the punisher's expectations for that same contribution).

Overall, this procedure leads to three values of *E[Received]* for each participant in each round. For example, in the first round, suppose player *b* contributed \$6, player *c* contributed \$10, and player *d* contributed \$12. From player *a*'s perspective, their expectation for player *b* ($E[Received]_{1ab}$) is \$11, (\$10 + \$12)/2. Player *a*'s expectation for player *c* ($E[Received]_{1ac}$) is \$9, (\$6 + \$12)/2. Player *a*'s expectation for player *d* ($E[Received]_{1ad}$) is \$8, (\$6 + \$10)/2.

Formally, player *a's* Observational view expectation for a given player *b* in trial *n* is:

$$E[Received]_{nab} = \frac{1}{3n-1}\left(\sum_{i}^{n-1}\sum_{j\neq a}^{J} x_{ij} + \sum_{j\notin[a,b]}^{J} x_{nj}\right)$$

J represents the set of the four players, and x_{ij} represents the amount contributed by player *j* in trial *i*. Note that "*3n – 1"* is used rather than "*3n*" to account for the exclusion of player *b*'s contribution in round[n]. Preliminary analyses considered alternative definitions for *E[Received]*, such as examining the average amount contributed by *c* and *d* in previous trials while excluding player *b* entirely. However, these alternative definitions yielded a worse model fit. Hence, to ensure that *E[Proposed]* is compared to the strongest possible definition of *E[Received]*, we proceeded with the equation above.

Multilevel linear regressions predicted punishment amounts using both *E[Proposed]* and *E[Received]* as predictors. Like in Studies 1 and 2, the regression controlled for two covariates: (1) the contribution amount being evaluated and (2) the average amount of punishment the participant received in the previous round. Given the large sample size, which meant that both *E[Proposed]* and *E[Received]* would likely be significant predictors, follow-up regressions were performed examining model fit. These regressions used the two covariates and either only *E[Proposed]* or only *E[Received]* as predictors. These follow-up regressions also did not include random slopes, which would impede the interpretation of model fit. The mediations reported in Studies 1 and 2 were also tested.

Notably, the PGG introduces the possibility for antisocial punishment, whereby participants punish behavior seen as unexpectedly generous. To test the two perspectives on expectations under this lens, two models of unsigned expectation-violation were tested, *ExV[Received]* and *ExV[Proposed]*, which represent absolute differences relative to the contribution being evaluated.

$$ExV[Received]_{nab} = |E[Received]_{nab} - x_{nb}|$$
$$ExV\left[Proposed\right]_{nab} = |E\left[Proposed\right]_{na} - x_{nb}|$$

ExV[Proposed] and *ExV[Received]* were submitted to multilevel regressions, as above.

Finally, to assess cross-cultural generalizability, further regressions were performed, which each only used one of the four expectation or expectation-violation models as predictors. These regressions were fit separately for each of the 16 cities in the dataset. We then tabulated which model achieved the highest fit across the greatest number of cities.

4.2. Results

Participants' punishment amounts were significantly linked to both *E[Proposed]* ($\beta = .15$ [.11, .19], $p < .001$) and *E[Received]* ($\beta = .13$ [.09, .18], $p < .001$). However, when regressed separately, the *E[Proposed]* regression yielded slightly greater model fit (Table 1). Additional analyses replicated the mediation from Studies 1 and 2, whereby others' contributions in the previous influenced participants' subsequent contributions ($\beta = .43$ [.39, .46], $p < .001$), which in turn influenced participants' subsequent punishments ($\beta = .14$ [.10, .18], $p < .001$; significant mediation: $p < .001$).

Table 1
Model comparison for Study 3

Model	Fit (ΔLL)	Cities best fit
E[Proposed]	131.1	0
E[Received]	124.0	1 (Istanbul)
ExV[Proposed]	645.1	14
ExV[Received]	255.7	1 (Riyadh)

Note. Model fit (Δlog-likelihood [LL]) was calculated as the difference relative to a model without any expectation variable as a predictor. For the two cities where *E[Received]* and *ExV[Received]* achieved the greatest fit, the difference relative to *ExV[Proposed]* was slight (Istanbul: $\Delta LL = 4.2$; Riyadh: $\Delta LL = 1.9$).

Further analyses considered whether both unexpectedly selfish and unexpectedly generous violations beget punishments. Consistent with this, greater punishment was linked to both *ExV[Proposed]* ($\beta = .20$ [.19, .21], $p < .001$) and *ExV[Received]* ($\beta = .03$ [.02, .05], $p < .001$). However, model comparison showed that the *ExV[Proposed]* regression yielded a far higher fit than any other model (Table 1). Furthermore, fitting the regressions separately for each city in the dataset revealed that *ExV[Proposed]* best predicted behavior in the vast majority of cities (Table 1).

4.3. Discussion

Study 3 further supports the idea that participants' expectations primarily depend on their own behavior, demonstrating that they do not apply just to UG rejections but also to PGG punishments and antisocial punishments. Relative to the UG, the PGG has two noteworthy unique features. First, because participants are repeatedly playing in a group, this may create stronger impressions of what constitutes typical behavior. Hence, finding comparisons with one's own behavior to still be more predictive of punishment may be particularly powerful evidence against the idea that expectation-violations are rooted in perceptions of typical behavior (Irwin & Horne, 2013). Second, in the PGG, the amount that participants can punish is independent of the earnings they receive from the other player's contribution (e.g., in the UG, punishment/rejection of \$6:\$4 splits necessarily costs \$4, which is not true for the PGG), which may enhance the clarity of the results' interpretations. Along with generalizing across economic games, the results also show that the expectation pathways laid out here emerge cross-culturally, speaking to the possibilities that they are universal and foundational aspects of social cognition. To further generalize, the next study examined whether these reciprocity-based mechanisms also apply beyond just punishment and to longer-term evaluations.

5. Study 4

Study 4 tested whether catering to participants' expectations over the course of repeated trials (low *ExV[Proposed]*) could generate overall positive evaluations. These overall evaluations were measured using the Trust Game. Specifically, participants were told that they

would play the UG with the same person for the entirety of each block, and with different participants in different blocks. In truth, in each block, participants interacted with a different computer profile representing a unique behavioral strategy, including a *Reciprocity* computer, a *Generous* computer, and a *Control* computer. Building on the results thus far, we expected the Reciprocity profile to elicit high levels of trust, higher than what would be elicited by the Control profile and potentially comparable to the trust shown to the Generous profile.

5.1. Methods

5.1.1. Participants

A total of 227 undergraduate students were recruited from the local university for this online study ($M_{age} = 20.8$ [18, 26], $SD_{age} = 2.5$; 59% female, 40% male, 1% other), although six participants were excluded because they did not give responses in the UG or Trust Game. Post hoc power analyses show that this sample has over 99.9% power ($\alpha = .05$) for the paired *t*-tests of interest, and sensitivity analyses show that it is sufficient to detect effects of at least $d_z = 0.19$ (d_z is the form of Cohen's *d* most appropriate for repeated-measures power analyses; Lakens, 2013). All participants provided informed consent under a protocol approved by the Institutional Review Board and received course credit as compensation.

5.1.2. Task design

This study used trial protocols identical to those of Studies 1 and 2. Compensations procedures were identical to those of Study 2. However, in the present study, participants were told that they would only change UG partners between blocks. In each block, participants played with one of five computer partners, administered in a counterbalanced order. Of these five partners, three are the focus of the present report. First, a *Reciprocity* profile copied participants' Proposer and Responder choices. It performed tit-for-tat with slight changes to obfuscate the copying (detailed in Supplementary Materials 3.1). Second, a *Generous* profile was based on the average behavior of the five most generous Proposer participants and the five most accepting Responder participants in Study 1. This profile largely proposed equitable offers (\$5:\$5 in 63% of trials, \$6:\$4 in 32% of trials, and \$7:\$3 in 5% of trials) and accepted most offers (accepted all offers $\geq$ \$2 and accepted 76% of \$1 offers). Third, a *Control* profile was based on the distribution of behavior seen in Study 1 (Fig. 4 histogram). Participants played with each profile either once (five blocks total) or twice (10 blocks total), although this distinction had no impact on the results.

After the last UG trial of each block, participants played the standard Trust Game as the Investor (Berg, Dickhaut, & McCabe, 1995). Participants could invest \$0, \$5, \$10, \$15, or \$20. Money invested would purportedly be tripled but under the possession of the other player, so greater investment indicates greater trust that the player will return some money. Participants were told that the other player's decision would not be revealed until the experiment ended to prevent influence on investments in later blocks. Trust Game trials where participants failed to respond were excluded from the analysis. After the task, participants completed the Study 2 questionnaires, but these are beyond the present focus.

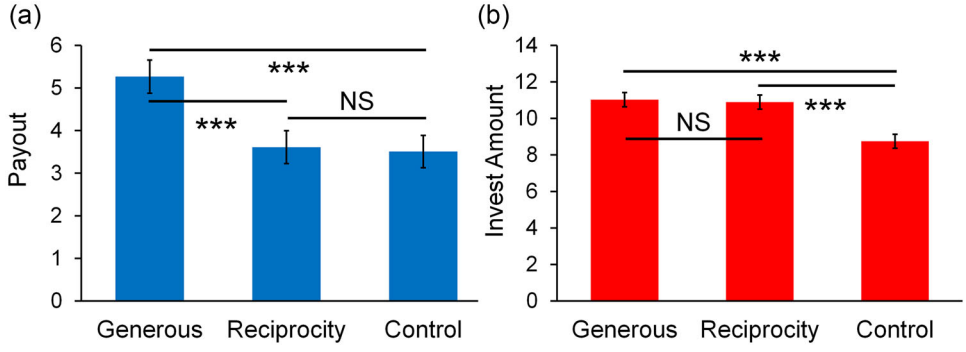

Fig. 7. Comparison of payouts and trust in interactions with different playstyle partners. (a) As planned, the Generous condition led to higher average earnings for the participant than the other two conditions. The Reciprocity and Control conditions, on the other hand, led to moderate payouts. (b) Consistent with the Experiential view, participants tended to highly trust their partner in the Reciprocity condition. Participants trusted the Reciprocity partner as much as the Generous profile and more than the Control profile. ***, $p < .001$; NS, nonsignificant.

As in Study 2, the success of deception was assessed using a 5-point scale in which participants were asked about the extent to which they believed that they interacted with humans or computers (1 = "Definitely computers"; 5 = "Definitely humans"). For confirmatory purposes, the present analyses were also conducted on the subset of participants who reported believing the deception (score of 3 or higher, $N = 125$). These analyses are reported in Supplementary Materials 3.2, and the results showed that each significant *t*-test result below remained significant, and each insignificant *t*-test result remained insignificant.

5.2. Results

Examination of the UG payouts confirmed that the Generous profile yielded much higher earnings for participants than the Reciprocity or Control profiles (d_zs > 1.02, ps $< .001$; Fig. 7A). The latter two yielded similar payouts ($d_z = 0.08$, $p = .74$; Fig. 7A). Hence, any trust-related differences between the Reciprocity and Control partners would not be confounded by UG payouts.

Analysis of the Trust Game showed that the level of trust elicited by the Reciprocity computer was virtually identical to the trust elicited by the Generous partner ($d_z = 0.01$, $p = .92$; Fig. 7B) and substantially greater than the trust elicited by the Control partner ($d_z = 0.40$, $p < .001$; Fig. 7B). Follow-up analyses showed that high investment was also linked to low average *ExV[Proposed]* across the block, whereas average *ExV[Received]* had no effect (Supplementary Materials 3.3).

5.3. Discussion

Study 4 provided further evidence that aligning others' behaviors to match the participants' leads to positive evaluations. Remarkably, observing reciprocity elicited similar trust as experiencing generosity. Hence, the mechanisms underlying the Experiential view are not

just subtle cognitive patterns but major forces that can shape social interactions and subsequent behaviors. Relatedly, the present results suggest that selfish individuals will trust others who are also selfish like them, even at their own expense, paralleling the Study 3 results on antisocial punishment.

6. General discussion

The goal of the present research was to investigate the interplay among observations, social expectations, and decision-making. We aimed to clarify how these processes influence evaluations of fairness and trustworthiness. These aspects were studied with economic games (the UG, PGG, and Trust Game), and investigated in a dataset from an international sample. Study 1 showed the expectation model based on participants' previously proposed selfish/generous offers better predicted participants' later likelihood of rejecting UG offers. Further analyses showed that changes to participants' Proposer behavior (reciprocity) also mediated an indirect effect of previously observing selfish/generous behavior. Study 2 replicated each result from the first study and found that they also apply within overall selfish contexts. Study 3 further replicated and generalized the results and showed that they apply to different economic games, apply to antisocial punishment, and apply cross-culturally. Finally, Study 4 causally tested the effects of interacting with another player who reciprocates behavior, showing that it yields high levels of trust.

We can now provide an answer to our original question of whether a generous person who frequently observes selfish behavior will learn to accept it. Converging evidence across multiple economic games suggests that the generous person would continue punishing selfish behavior (Studies 1–3) and continue distrusting those who are selfish (Study 4) unless the drive to reciprocate compels them to start acting selfishly themselves. Furthermore, in cases where participants are selfish, even if those around them are generous, participants will continue disliking generous others and may even antisocially punish them (Study 3). Study 3 also shows that the link between participants' evaluations and their own generous/selfish decision-making emerges across almost all sampled cultures, suggesting that this pathway may be a foundational piece of social cognition. In this sense, evaluation based on dissimilarity is analogous to reciprocity, which is also seen cross-culturally (Curry et al., 2019). Although the idea that participants like others who behave similarly to them may seem obvious, it remains striking that similarity will yield positive evaluations of even personally harmful behavior and that this occurs cross-culturally. Taken together, this prominent effect and cross-cultural evidence suggest that the link between decision-making and social evaluation is a foundational piece of social cognition.

The results fill gaps in earlier theories on the role of social norms in impacting evaluations. Research arguing for the Observational view (i.e., the "Common is Moral heuristic") posit that atypical behavior will be judged negatively (Goldring & Heiphetz, 2020; Gollwitzer, Martel, Bargh, & Chang, 2020; Lindström et al., 2018; Xiang et al., 2013). The present results suggest that expectation-violations should instead be primarily seen as deviations relative to people's own behavior. As our four studies demonstrated, this clearly manifests in the UG and PGG.

This conclusion also applies beyond economic games. For example, the Observational view has been used to explain why individuals judge harmful and weird acts as worse than typical harmful behavior (e.g., hitting someone with a frozen fish is seen as worse than punching them; Walker, Turpin, Fugelsang, & Białek, 2021). Although previous studies have explained patterns like these in terms of how typical the behavior is, the present results suggest that it is more accurate to reason that they arise because individuals would be unwilling to perform the weird behavior themselves.

The present findings point to a potential mechanistic link between how participants make decisions and how they evaluate others' decisions. This is consistent with neural evidence and conceptual accounts, arguing for how the brain co-opts neural operations for multiple purposes (Lockwood, Apps, & Chang, 2020; Parkinson & Wheatley, 2015). For example, studies on mirror neurons show how the neural populations responsible for performing some action (e.g., waving hello) also encode information when observing another person perform said action (Iacoboni et al., 2005; Oosterhof, Tipper, & Downing, 2013). Another parallel comes from studies on reward processing. Receiving a reward recruits some of the same brain regions that also activate when observing another person receive a reward (e.g., the ventromedial prefrontal cortex; Morelli, Sacchet, & Zaki, 2015). Cognitive psychology also suggests co-opting of algorithmic processes between "non-social" and social cognition. For instance, reinforcement learning models originally designed for decision-making research also apply to understanding how individuals infer another person's goals based on their decisions (Jara-Ettinger, 2019; Jern, Lucas, & Kemp, 2017). Adding to this theoretical position, the present findings may suggest that there is an overlap between the cognitive processes responsible for decision-making and those responsible for social evaluation (Cushman, 2013).

Each study also showed significant effects of reciprocity, whereby participants behaved more selfishly after being treated selfishly and more generously after being treated generously. Consistent with earlier research, we identified both direct reciprocity in repeated dyadic/group interactions (Studies 3 and 4) and indirect reciprocity in interactions with multiple anonymous others (Studies 1 and 2; Kelley & Stahelski, 1970b; Nowak & Sigmund, 2005; Romano, Saral, & Wu, 2022). Our results are also consistent with evidence that individuals often cooperate, conditionally, depending on others' cooperation (Fischbacher et al., 2001; Krueger, DiDonato, & Freestone, 2012). Our main point of novelty in this area was showing that reciprocity effects mediate the impact of previous observations on expectations and later decisions to punish, providing further evidence for the importance of measuring participants' own decision-making when investigating their evaluations.

Our findings on expectations and those on reciprocity can additionally be integrated in terms of social alignment. Regarding participants' expectations, participants punished dissimilarity via rejection and rewarded similarity via acceptance. Punishment and reward often serve as forms of communication (Ho, Cushman, Littman, & Austerweil, 2019; Sarin, Ho, Martin, & Cushman, 2021), and punishing dissimilarity while rewarding similarity compels other people to shift their behavior to be more like the punishers'. Alongside these punishments, participants reciprocated others' behavior. Taken together, these give rise to alignment: Punishment/reward encourages others to be more like oneself, whereas reciprocity involves making oneself more like others.

6.1. Caveats

It should be noted that mediations (Studies 1–3) generally cannot provide evidence supporting a specific causal direction (Valente, Pelham, Smyth, & MacKinnon, 2017), which limits inferences into causal directions. However, correlational research develops key foundations for future studies (Grosz, Rohrer, & Thoemmes, 2020), and Study 4 is an initial step focusing on causality. An additional limitation was the use of deception in Studies 1, 2, and 4, where participants interacted with computers and not genuine human players. The use of deception may undermine a study's validity (Colson, Corrigan, Grebitus, Loureiro, & Rousu, 2016; Krupat & Garonzik, 1994). Nonetheless, confirmatory analyses show how the Study 2 and 4 results remain significant when examining only participants who believed the deception (Supplementary Materials 2.3 and 3.2), and the Study 3 results show how the conclusions remain robust in a design without any deception. Another caveat is that Studies 2 and 4 used time-saving as an incentive rather than monetary payment. Although earlier work has validated the use of time-saving incentives (Jouxtel, 2019; Noussair & Stoop, 2015), time-saving incentives require further research. Finally, questions remain regarding injunctive norms, as the present studies only considered expectations in terms of descriptive norms. Potentially, the patterns found here generalize toward injunctive norms—for example, instructing participants what offers they should expect only impacts their evaluations insofar as participants begin to propose those offers—although this remains a question for future research.

7. Conclusion

Comparing the effects of past observations and personal decision-making in forming expectations shows that the latter has a more primary role, such that participants punish behavior dissimilar from their own more than behavior dissimilar to what is typical. However, previous observations have an indirect effect on evaluations, mediated by changes in participants' own selfish/generous behavior. In a sense, these results invert the Golden Rule, which states that "you should treat others how you wish to be treated." The present results show that you additionally expect others to treat you the same way that you treat others. Studies 1–3 demonstrated this point by modeling expectations and the temporal flow between observations, decision-making, and punishments. Study 4 provided causal evidence in support of this hypothesis, demonstrating that participants will tend to trust others who copy them as much as they trust outright generosity. Overall, our results point to the importance of social alignment not only for decision-making, which is characterized by reciprocity but also for social evaluation, which we show is principally characterized by an expectation of reciprocity. Beyond the laboratory, these results are relevant for understanding social cooperation and have practical implications for teaching acceptance of others' unfamiliar behaviors and improving interaction in less cooperative groups. Specifically, to teach acceptance of another person's behavior, individuals must be convinced to change their decision-making such that they would behave the same way if they were in this other person's shoes.

Author contributions

F.D., S.D., I.K., M.M., and P.C.B. conceived the study; I.K., M.M., and P.C.B collected data; P.C.B., M.M., S.D., and F.D. planned the analytic approach with feedback from S.A.C.; P.C.B. performed the analyses with feedback from F.D., S.D., and S.A.C.; P.C.B. and F.D. wrote the manuscript with feedback from M.M., S.D., and S.A.C., and all authors approved the content of the manuscript.

Acknowledgments

This research was carried out in part at the University of Illinois' Beckman Institute for Advanced Science & Technology, and was supported by funds to the Dolcos Lab from the University of Illinois' Psychology Department and the Beckman Institute. During the preparation of this manuscript, P.C.B. was supported by a Predoctoral Fellowship provided by the Beckman Foundation and a Dissertation Completion Fellowship provided by the University of Illinois; F.D. was supported by an Emanuel Donchin Professorial Scholarship in Psychology from the University of Illinois; M.M. was supported by Pre- and Postdoctoral Fellowships provided by the Beckman Foundation, and I.K. was supported by a Fulbright Foundation Scholarship. The authors wish to thank members of the Dolcos Lab for assisting with data collection, Dr. Scott Huettel for feedback during the design of the study and on an earlier version of the manuscript, and Dr. Aron K. Barbey for feedback on an earlier version of this manuscript.

Data availability statement

Our task materials, analytic code, and collected data were added to the following OSF repository (https://osf.io/jxfpg/), to the extent possible: The task materials for Study 1 (PsychoPy) and for Studies 2 and 4 (PsychoJS/Pavlovia) were uploaded. Study 3 is a reanalysis of a public dataset, and its task materials are detailed by Herrmann et al. (2008). All analytic code (Python & R) was also uploaded. The data we collected for Studies 2 and 4 were uploaded as well. Per consultation with our Institutional Review Board, the Study 1 data could not be uploaded.

Note

1 The data, tasks, and analytic code were uploaded to a public OSF repository (see Data Availability).

References

Aksoy, O., & Weesie, J. (2012). Beliefs about the social orientations of others: A parametric test of the triangle, false consensus, and cone hypotheses. *Journal of Experimental Social Psychology*, *48*(1), 45–54.

Allidina, S., Arbuckle, N. L., & Cunningham, W. A. (2019). Considerations of mutual exchange in prosocial decision-making. *Frontiers in Psychology*, *10*, 1216.

Bardsley, N., & Sausgruber, R. (2005). Conformity and reciprocity in public good provision. *Journal of Economic Psychology*, *26*, (5), 664–681.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv*.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*, (1), 122–142.

Bogaert, S., Boone, C., & Declerck, C. (2008). Social value orientation and cooperation in social dilemmas: A review and conceptual model. *British Journal of Social Psychology*, *47*, (3), 453–480.

Bogdan, P. C., Moore, M., Kuznietsov, I., Frank, J. D., Federmeier, K. D., Dolcos, S., & Dolcos, F. (2022). Direct feedback and social conformity promote behavioral change via mechanisms indexed by centroparietal positivity: Electrophysiological evidence from a role-swapping ultimatum game. *Psychophysiology*, *59*, (4), e13985.

Boudewyn, M. A., Luck, S. J., Farrens, J. L., & Kappenman, E. S. (2018). How many trials does it take to get a significant ERP effect? It depends. *Psychophysiology*, *55*, (6), e13049.

Chang, L. J., & Sanfey, A. G. (2013). Great expectations: Neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience*, *8*, (3), 277–284.

Colson, G., Corrigan, J. R., Grebitus, C., Loureiro, M. L., & Rousu, M. C. (2016). Which deceptive practices, if any, should be allowed in experimental economics research? Results from surveys of applied experimental economists and students. *American Journal of Agricultural Economics*, *98*, (2), 610–621.

Cooper, D. J., & Dutcher, E. G. (2011). The dynamics of responder behavior in ultimatum games: A meta-study. *Experimental Economics*, *14*, (4), 519–546.

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60*, (1), 47–69.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*, (3), 273–292.

De Cremer, D., & Van Lange, P. A. (2001). Why prosocials exhibit greater cooperation than proselfs: The roles of social responsibility and reciprocity. *European Journal of Personality*, *15*(1_suppl), S5–S18.

Denant-Boemont, L., Masclet, D., & Noussair, C. N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory*, *33*, (1), 145–167.

Eriksson, K., Strimling, P., & Coultas, J. C. (2015). Bidirectional associations between descriptive and injunctive norms. *Organizational Behavior and Human Decision Processes*, *129*, 59–69.

Fatfouta, R., Meshi, D., Merkl, A., & Heekeren, H. R. (2018). Accepting unfairness by a significant other is associated with reduced connectivity between medial prefrontal and dorsal anterior cingulate cortex. *Social Neuroscience*, *13*, (1), 61–73.

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

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, (6868), 137–140.

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behavior*, *2*, (7), 458–468.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*, (3), 397–404.

Gächter, S., Herrmann, B., & Thöni, C. (2004). Trust, voluntary cooperation, and socio-economic background: Survey and experimental evidence. *Journal of Economic Behavior & Organization*, *55*, (4), 505–531.

Goldring, M. R., & Heiphetz, L. (2020). Sensitivity to ingroup and outgroup norms in the association between commonality and morality. *Journal of Experimental Social Psychology*, *91*, 104025.

Gollwitzer, A., Martel, C., Bargh, J. A., & Chang, S. W. (2020). Aversion towards simple broken patterns predicts moral judgment. *Personality and Individual Differences*, *160*, 109810.

Grecucci, A., Giorgetta, C., Van't Wout, M., Bonini, N., & Sanfey, A. G. (2013). Reappraising the ultimatum: An fMRI study of emotion regulation and decision making. *Cerebral Cortex*, *23*, (2), 399–410.

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.
Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, *15*, (5), 1243–1255.
Henrich, J. (2017). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.
Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*, (5868), 1362–1367.
Hetu, S., Luo, Y., D'Ardenne, K., Lohrenz, T., & Montague, P. R. (2017). Human substantia nigra and ventral tegmental area involvement in computing social error signals during the ultimatum game. *Social Cognitive and Affective Neuroscience*, *12*, (12), 1972–1982.
Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2019). People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General*, *148*, (3), 520.
Horne, C., & Mollborn, S. (2020). Norms: An integrated framework. *Annual Review of Sociology*, *46*, 467–487.
Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology*, *3*, (3), e79.
Irwin, K., & Horne, C. (2013). A normative explanation of antisocial punishment. *Social Science Research*, *42*, (2), 562–570.
Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110.
Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64.
Jouxtel, J. (2019). Voluntary contributions of time: Time-based incentives in a linear public goods game. *Journal of Economic Psychology*, *75*, 102139.
Jung, H., Seo, E., Han, E., Henderson, M. D., & Patall, E. A. (2020). Prosocial modeling: A meta-analytic review and synthesis. *Psychological Bulletin*, *146*, (8), 635.
Kawamura, Y., & Kusumi, T. (2020). Altruism does not always lead to a good reputation: A normative explanation. *Journal of Experimental Social Psychology*, *90*, 104021.
Kelley, H. H., & Stahelski, A. J. (1970a). Errors in perception of intentions in a mixed-motive game. *Journal of Experimental Social Psychology*, *6*, (4), 379–400.
Kelley, H. H., & Stahelski, A. J. (1970b). Social interaction basis of cooperators' and competitors' beliefs about others. *Journal of Personality and Social Psychology*, *16*, (1), 66.
Krueger, J. I., DiDonato, T. E., & Freestone, D. (2012). Social projection can solve social dilemmas. *Psychological Inquiry*, *23*, (1), 1–27.
Krupat, E., & Garonzik, R. (1994). Subjects' expectations and the search for alternatives to deception in social psychology. *British Journal of Social Psychology*, *33*, (2), 211–222.
Kubota, J. T., Li, J., Bar-David, E., Banaji, M. R., & Phelps, E. A. (2013). The price of racial bias: Intergroup negotiations in the ultimatum game. *Psychological Science*, *24*, (12), 2498–2504.
Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863.
Leimgruber, K. L. (2018). The developmental emergence of direct reciprocity and its influence on prosocial behavior. *Current Opinion in Psychology*, *20*, 122–126.
Lindström, B., Jangard, S., Selbing, I., & Olsson, A. (2018). The role of a "common is moral" heuristic in the stability and change of moral norms. *Journal of Experimental Psychology: General*, *147*, (2), 228.
Lockwood, P. L., Apps, M. A., & Chang, S. W. (2020). Is there a 'social' brain? Implementations and algorithms. *Trends in Cognitive Sciences*, *24*, (10), 802–813.
Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092.
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.

Myslinska Szarek, K., & Tanas, Ł. (2022). I scratched your back; Should you not scratch mine? The expectation of reciprocity in 4- to 6-year-old children following a prosocial investment. *European Journal of Developmental Psychology*, *19*, (3), 383–399.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*, *92*, (1–2), 91–112.

Noussair, C. N., & Stoop, J. (2015). Time as a medium of reward in three social preference experiments. *Experimental Economics*, *18*, (3), 442–456.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*, (7063), 1291–1298.

Oosterbeek, H., Sloof, R., & Van de Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, *7*, (2), 171–188.

Oosterhof, N. N., Tipper, S. P., & Downing, P. E. (2013). Crossmodal and action-specific: Neuroimaging the human mirror neuron system. *Trends in Cognitive Sciences*, *17*, (7), 311–318.

Parkinson, C., & Wheatley, T. (2015). The repurposed social brain. *Trends in Cognitive Sciences*, *19*, (3), 133–141.

R Core Team. (2013). R: A language and environment for statistical computing.

Romano, A., Saral, A. S., & Wu, J. (2022). Direct and indirect reciprocity among individuals and groups. *Current Opinion in Psychology*, *43*, 254–259.

Sarin, A., Ho, M. K., Martin, J. W., & Cushman, F. A. (2021). Punishment is organized around principles of communicative inference. *Cognition*, *208*, 104544.

Selig, J. P., & Preacher, K. J. (2008). Monte Carlo method for assessing mediation: An interactive tool for creating confidence intervals for indirect effects [Computer software].

Valente, M. J., Pelham, III, W. E., Smyth, H., & MacKinnon, D. P. (2017). Confounding in statistical mediation analysis: What it is and how to address it. *Journal of Counseling Psychology*, *64*, (6), 659.

Van Lange, P. A. (1992). Confidence in expectations: A test of the triangle hypothesis. *European Journal of Personality*, *6*, (5), 371–379.

Vavra, P., Chang, L. J., & Sanfey, A. G. (2018). Expectations in the ultimatum game: Distinct effects of mean and variance of expected offers. *Frontiers in Psychology*, *9*, 992.

Walker, A. C., Turpin, M. H., Fugelsang, J. A., & Białek, M. (2021). Better the two devils you know, than the one you don't: Predictability influences moral judgments of immoral actors. *Journal of Experimental Social Psychology*, *97*, 104220.

Wedekind, C., & Milinski, M. (1996). Human cooperation in the simultaneous and the alternating Prisoner's Dilemma: Pavlov versus Generous Tit-for-Tat. *Proceedings of the National Academy of Sciences*, *93*, (7), 2686–2689.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience*, *33*, (3), 1099–1108.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information
Supporting Information
Supporting Information