Allocation Intention and Upward Indirect Reciprocity: Evidence from Behavior and ERP

Authors: Wang Ting, Zhao Liangfo, Yang Jinpeng, Zhang Dandan, Lei Zhen, Wang Ting

Date: 2024-08-26T00:00:00+00:00

Abstract

Upward indirect reciprocity is widely observed in both real-world and laboratory settings, transcending repetitive, closed reciprocal systems and serving as an important force driving large-scale human cooperation and the expansion of social order. Although this social phenomenon has attracted considerable attention from scholars, existing literature suffers from two major shortcomings: on the one hand, it neglects the fundamental question of what criterion serves as the benchmark for judging benevolence versus non-benevolence in upward indirect reciprocity; on the other hand, most studies on upward indirect reciprocity fail to recognize that "after B receives benevolence or non-benevolence from A, B treats a third party C in the same manner" may be explained by a competing hypothesis known as the "income effect." This study hypothesizes that when individuals receive allocations above the social mean, the amount given to third parties under human allocation conditions will be above the social mean and higher than that under computer allocation conditions; when individuals receive allocations below the social mean, the amount given to third parties under human allocation conditions will be below the social mean but lower than that under computer allocation conditions. The study employs a two-stage dictator game paradigm combined with ERP technology to seek evidence at the level of brain neural activity. Both behavioral and EEG results provide strong support for the hypotheses of this study. Human allocation elicits larger N1 amplitudes than computer allocation; allocations below the social mean elicit larger FRN than allocations above the social mean; under human conditions, receiving allocations below the social mean elicits larger P3 amplitudes than receiving allocations above the social mean, whereas under computer conditions, the P3 amplitudes elicited by receiving allocations below or above the social mean show no significant difference. These results support the upward indirect reciprocity hypothesis based on social allocation means, providing new theoretical foundation and experimental evidence for this research field.

Full Text

Intention and Upstream Indirect Reciprocity: Evidence from Behavior and ERP

WANG Ting[1], **ZHAO Liangfo**[2], **YANG Jinpeng**[2], **ZHANG Dandan**[2,3], **LEI Zhen**[2,4]

[1] Business School, Sichuan University, Chengdu 610065, China

[2] China Center for Behavioral Economics and Finance, Southwestern University of Finance and Economics, Chengdu 611130, China

[3] School of Economics, Southwestern University of Finance and Economics, Chengdu 611130, China

[4] School of Economics, Southwestern University of Finance and Economics, Chengdu 611130, China

Abstract

Upstream indirect reciprocity is a pervasive phenomenon in both real-world and laboratory settings that extends beyond closed, repeated exchange systems, serving as a crucial force for promoting large-scale human cooperation and expanding social order. Despite attracting substantial scholarly attention, existing literature suffers from two major limitations: First, it overlooks the fundamental question of what benchmark people use to evaluate benevolent versus non-benevolent actions in upstream indirect reciprocity. Second, most studies fail to recognize that the pattern "B receives benevolence or non-benevolence from A, then treats third party C similarly" may be explained by an alternative "income effect" hypothesis.

We propose that when individuals receive allocations above the social mean, they will allocate more to third parties in human allocation conditions compared to computer allocation conditions, and these allocations will exceed the social mean. Conversely, when receiving allocations below the social mean, they will allocate less to third parties, with human allocations resulting in even lower amounts than computer allocations. Using a two-stage dictator game paradigm combined with ERP technology, we examine the neural correlates of these processes. Both behavioral and electrophysiological results strongly support our hypotheses. Human allocation elicited larger N1 amplitudes than computer allocation. Below-mean allocations generated larger feedback-related negativity (FRN) than above-mean allocations. Critically, under human allocation conditions, receiving below-mean allocations evoked larger P3 amplitudes than receiving above-mean allocations; under computer conditions, no significant P3 differences emerged between below- and above-mean allocations. These findings support a social-mean-based account of upstream indirect reciprocity, providing new theoretical foundations and experimental evidence for this research domain.

Keywords: upstream indirect reciprocity, social norms, distributional inten-

tion, event-related potential
Classification Codes: B845; R395

1. Introduction

Reciprocity represents a fundamental mechanism promoting human cooperation. Direct reciprocity—returning kindness directly to benefactors and non-kindness to non-benefactors—can explain long-term, repeated cooperative interactions between individuals (Nowak & Sigmund, 2005). However, in many situations, benefactors and recipients cannot establish sustained, direct relationships. When individuals cannot reciprocate directly to the original actor, they may instead transfer kindness or non-kindness to unrelated third parties—a phenomenon termed indirect reciprocity (Liu et al., 2022). Indirect reciprocity transcends the closed system of direct reciprocity and constitutes a vital force for large-scale human cooperation, even forming an essential component of Hayek's concept of spontaneous social order (Hayek, 1980). Consequently, identifying the internal mechanisms of indirect reciprocity holds significant theoretical and practical importance.

Indirect reciprocity manifests in two forms: upstream and downstream. This study focuses on upstream indirect reciprocity because, compared to its downstream counterpart, it excludes strategic motives and higher-order beliefs associated with reputation mechanisms, thereby facilitating a more direct examination of indirect reciprocity's underlying mechanisms. Upstream indirect reciprocity typically refers to situations where B, after receiving kindness (or non-kindness) from A, treats third party C similarly (Nowak & Sigmund, 2005). According to the homo economicus assumption in neoclassical economics, in anonymous one-shot games, B would have no incentive to treat C with (non-)kindness after receiving (non-)kindness from A, since no benefits can be derived from repeated interactions, making indirect reciprocity chains difficult to sustain (Sigmund, 2010). Nevertheless, upstream indirect reciprocity has been robustly documented in numerous behavioral studies and aligns with real-world observations, attracting considerable attention from psychology and behavioral science researchers (Engelmann & Fischbacher, 2009; Horita et al., 2016; Rutte & Taborsky, 2007; Sun et al., 2022).

Existing experimental literature typically employs two-stage dictator games (DG) to characterize upstream indirect reciprocity (Gray et al., 2014; Hu et al., 2018; Liu et al., 2022). In the first-stage DG, subjects A and B are randomly paired, with A as allocator and B as recipient who cannot reject A's proposal. In the second-stage DG, subjects B and C are randomly paired, with B now serving as allocator and C as recipient who cannot reject B's proposal. Meta-analyses of DG experiments reveal that dictators allocate an average of 28% to recipients, with allocations exceeding 50% being extremely rare (Engel, 2011). Surprisingly, however, existing studies (e.g., Gray et al., 2014; Hu et

al., 2018) almost universally treat A's 50% allocation to B as the threshold for B's judgment of A's benevolence, yet provide insufficient justification for this criterion.

We argue that identifying what constitutes benevolence versus non-benevolence is not only a prerequisite for B to transmit kindness but also the logical starting point for analyzing the cognitive mechanisms of indirect reciprocity. Typically, an allocation close to the social mean (28%) does not evoke surprise, whereas substantial deviations above or below this mean may be interpreted as benevolent or non-benevolent, respectively. In other words, individuals' interpretations of others' actions as benevolent or not are likely made relative to the social mean as a reference point. Accordingly, we define benevolent (non-benevolent) allocation as one that gives the anonymous recipient an amount above (below) the social mean. Based on this definition, we transform the upstream indirect reciprocity proposition—"B transmits benevolence (non-benevolence) after receiving it from A"—into Hypothesis 1: When individuals receive allocations above (below) the social mean, they tend to give third parties allocations above (below) the social mean. In our study, with 10 yuan in the first-stage DG, we use 1 and 2 yuan (below the 28% mean) to represent non-benevolent allocations and 4 and 5 yuan (above the mean) as benevolent allocations. We exclude 0 yuan because consumer preferences exhibit discontinuity in the right neighborhood of 0 (Niemand et al., 2019), and omitting 0 yuan ensures equal frequencies of benevolent and non-benevolent allocations.

Furthermore, if Hypothesis 1 holds, the results might also be explained by an alternative, competing hypothesis—the income effect—whereby people simply give more when they have higher incomes. Under this account, changes in allocations to third parties would be triggered not by the benevolence (or non-benevolence) represented by above- or below-mean allocations, but by income effects devoid of intentionality.

Distributional intention refers to an individual's subjective willingness to actively choose options that (dis)advantage the recipient when making allocation decisions (Falk & Fischbacher, 2006). Falk et al. (2008) demonstrated the importance of intention: when allocation values were determined by a human allocator A, B's return amounts increased with the allocation value, but when allocations were randomly generated by a computer, B's return amounts did not increase with allocation value. The difference between these two conditions represents B's response to A's intention. Accordingly, our study must examine whether intention-based indirect reciprocity persists after controlling for income effects.

Following the definition of distributional intention (Falk & Fischbacher, 2006) and previous experimental practices (Blount, 1995; Charnes & Rabin, 2002; Falk et al., 2008; Stanca, 2010; Zhang et al., 2016; Hu et al., 2018), we manipulate allocator role (human vs. computer) to capture intention effects. Compared to below-mean computer allocations, when B receives above-mean computer allocations and allocates more to C, this could be explained by income effects. However, when B receives above-mean allocations from a human A, B may expe-

rience not only higher income but also perceive benevolence from A, potentially leading to even higher allocations to C than in the computer condition—the difference representing the indirect reciprocity we aim to identify. The converse holds for below-mean allocations. This yields Hypothesis 2: When individuals receive allocations above (below) the social mean, they allocate higher (lower) amounts to third parties under human allocation conditions than under computer allocation conditions.

Beyond testing these two hypotheses behaviorally, we aim to open the "black box" of upstream indirect reciprocity decision-making at the neural level. To our knowledge, only one fMRI study (Hu et al., 2018) has explored the neural mechanisms underlying how allocation outcomes and intentions influence upstream indirect reciprocity. That study found significant activation in anterior insula, dorsal anterior cingulate cortex, and bilateral prefrontal cortex when participants received generous (above-equal) or greedy (below-equal) allocations compared to fair (equal) allocations, and greater activation in right temporoparietal junction and inferior parietal lobule when receiving generous human allocations versus generous computer allocations (compared to fair allocations). However, as noted above, using 50% equality as the threshold for judging benevolence is problematic.

Our study employs event-related potential (ERP) technology with high temporal resolution to reveal the time course of brain activity during upstream indirect reciprocity decisions. Based on previous research (Liu et al., 2022; Miraghaie et al., 2022; Moore et al., 2021), we focus on three ERP components: P1/N1, feedback-related negativity (FRN), and P3.

P1 and N1 components, located in lateral occipital regions, share similar psychological representations and are modulated by early visual and selective attention (Herrmann & Knight, 2001; Luck, 2014). A recent ERP study on social decision-making contexts found that real human interaction elicited larger N1 amplitudes than non-social contexts (Moore et al., 2021). Building on these findings, we propose that when the first-stage DG allocator is human, the allocation outcome will capture more selective attention, producing larger P1 or N1 amplitudes. This leads to Prediction 1: Human allocations will elicit larger P1/N1 amplitudes than computer allocations.

FRN is a frontocentral negativity occurring 200–250 ms after outcome feedback (Ma et al., 2015; Hoy et al., 2021), reflecting processing of personal financial losses (Gehring & Willoughby, 2002) and violations of social expectations and norms. The more an allocation violates expectations or social norms, the larger the FRN negativity (Wu et al., 2011; Mayer et al., 2019). DG studies show that unfair allocations elicit larger FRN than fair allocations (Zhong et al., 2019; Li et al., 2020). We therefore propose that below-mean allocations represent expectancy-violating outcomes, leading to Prediction 2: Below-mean allocations will produce larger FRN amplitudes than above-mean allocations.

The P3 component is a parietal positivity occurring 300–600 ms after feed-

back (Boudreau et al., 2009; Ma & Hu, 2015; Gong et al., 2022; Liu et al., 2022). Classic oddball paradigms show that P3 is elicited by low-probability stimuli (Duncan-Johnson & Donchin, 1977; Johnson & Donchin, 1980), with increased P3 amplitude signaling unexpected events (de Bruijin et al., 2007). This yields Prediction 3a: Below-mean allocations will elicit larger P3 amplitudes than above-mean allocations. Additionally, research indicates P3 reflects subjective value evaluation, with higher subjective value producing larger P3 amplitudes (Gu et al., 2011). In our study, human allocation outcomes imply benevolent (or non-benevolent) intentions, whereas computer allocations are random and intention-free. We therefore expect higher subjective value evaluation and greater attentional resources for human allocations, leading to Prediction 3b: Human allocations will elicit larger P3 amplitudes than computer allocations. Moreover, social decision-making research shows P3 is modulated by interactions between allocation outcomes and social context (Qu et al., 2013). Consequently, because below-mean human allocations contain others' non-benevolent intentions and carry stronger subjective value, we derive Prediction 3c: Under human conditions, receiving below-mean allocations will elicit larger P3 amplitudes than receiving above-mean allocations. Under computer conditions, no significant P3 differences should emerge between below- and above-mean allocations, as computer allocations lack intentional content.

Our study employs a two-stage DG framework to investigate upstream indirect reciprocity, innovatively treating above- (below-) mean allocations in the first-stage DG as (non-)benevolent rather than using equal splits. By manipulating allocator role (human vs. computer) to capture intentionality and employing ERP technology, we examine how allocation outcomes and intentions influence both behavioral decisions and neural activity in upstream indirect reciprocity.

2. Methods

2.1 Participants

We recruited 42 undergraduate students (21 male, 21 female) aged 18–22 years ($M = 20.05 \pm 0.14$ years). All participants were right-handed, had normal or corrected-to-normal vision, no psychiatric history, and no prior EEG experiment experience. Participants provided informed consent and received compensation. The study was approved by the Ethics Committee of the China Center for Behavioral Economics and Finance at Southwestern University of Finance and Economics. Based on the minimum effect size reported in related research (^{2} = 0.09; Liu et al., 2022), we conducted a power analysis using G*Power 3.1.9 (Faul et al., 2009) for repeated-measures ANOVA ($f = 0.31$, $\alpha = 0.05$), which indicated that 32 participants would achieve 99% statistical power. Our sample size therefore met requirements.

2.2 Experimental Procedure and Materials

The experiment comprised two parts: a standard DG behavior task and a two-stage DG indirect reciprocity EEG task.

Part 1: Standard DG Behavior Task. Participants acted as allocators distributing 10 yuan between themselves and randomly matched anonymous recipients. This task served three purposes: First, to verify whether our sample's allocation behavior aligned with previous DG literature and assess sample representativeness, while obtaining our sample's mean allocation to validate using 3 yuan as the threshold between benevolent and non-benevolent allocations. Second, having participants experience the allocator role enhanced credibility of the first-stage DG allocations in the indirect reciprocity task. Third, it allowed us to examine differences in indirect reciprocity behavior across participants with varying altruism levels.

Part 2: Indirect Reciprocity Task (Main Experiment). We employed a 2 (allocation outcome: below-mean vs. above-mean) $\times$ 2 (allocation intention: human vs. computer) within-subjects design. Following the classic two-stage DG framework, all participants completed multiple rounds of two-stage DGs with new random matching each round. As shown in Figure 1 [Figure 1: see original paper]A, in the first-stage DG, allocator A and recipient B shared 10 yuan; participants, as recipient B, could not reject A's allocation. In the second-stage DG, participant B became allocator with recipient C, who could not reject B's allocation.

Based on our hypotheses, we first examined whether participants allocated more to C after receiving above-mean versus below-mean allocations in the first-stage DG. Specifically, we tested whether above-mean allocations led to second-stage allocations exceeding the social mean, and whether below-mean allocations led to allocations below the social mean. Even if these results held, they could be explained by income effects. For our research on indirect reciprocity, the critical test concerned differences between human and computer allocation conditions: relative to computer allocations, when B received above- (below-) mean allocations from a human A, B would allocate higher (lower) amounts to C, with the difference representing the indirect reciprocity we aimed to identify.

The main experiment included both human and computer allocation phases, with phase order counterbalanced across participants. Each phase contained 156 trials (312 total). Within each phase, participants received three allocation types: above-mean allocations (4/6 and 5/5 splits, 30 trials each), below-mean allocations (1/9 and 2/8 splits, 30 trials each), and filler trials (0/10, 3/7, 6/4, 7/3, 8/2, 9/1 splits, 6 trials each). The experiment lasted approximately 45 minutes.

Participants completed 10 practice trials before beginning the formal experiment. As shown in Figure 1, each trial began with "Human" or "Computer" displayed on screen, indicating the current allocator type. After a fixation point,

the allocation amount from the first-stage DG allocator appeared for 1500 ms. Following a 500–800 ms blank screen, a photo of recipient C for the second-stage DG was presented. Participants then had 5 seconds to use a mouse to click numerical keys on screen to allocate 0–10 yuan to recipient C, clicking "Confirm" to submit. Participants were informed that five trials would be randomly selected after the experiment and their decisions in those trials would determine their task bonus, incentivizing careful responses.

Participants received a base payment of 50 yuan plus task bonuses, averaging approximately 100 yuan total. To enhance experimental credibility, we selected 312 standardized ID photos (156 male, 156 female) from an existing face database (Xie et al., 2021) as recipient C photos, with gender balanced across conditions. Before the experiment, the experimenter emphasized that in the human allocation task, each first-stage DG proposal came from a different allocator A, and that participants would make allocations to each recipient C in the second-stage DG. The experimenter also photographed each participant, explaining these photos would be used as experimental materials for other participants. Stimuli were presented using E-Prime 3.0 on a black background. First-stage DG allocation amounts appeared as white Times New Roman font (size 100, 2.3° × 3.0° visual angle). Recipient photos were presented in color at the center of the screen.

2.3 EEG Recording and Analysis

We recorded EEG using a 64-channel eegoTM mylab ERP system (ANT Neuro) with an extended 10–20 electrode layout. Online recording referenced CPz; offline analysis re-referenced to averaged mastoids. An electrode below the left eye recorded blink-related ocular artifacts. Impedance was maintained below 10 kΩ for all electrodes, with A/D sampling at 1000 Hz.

Offline analysis used the EEGlab toolbox in MATLAB R2017a. Data were downsampled to 250 Hz and filtered at 0.1–30 Hz. Independent component analysis removed blink artifacts, and trials exceeding $±$100 V were excluded. Epochs spanned from 200 ms pre- to 1500 ms post-allocation outcome presentation, with the 200 ms pre-stimulus interval serving as baseline.

We focused on three ERP components: occipital P1/N1, frontocentral FRN, and parietal P3. As final results showed no significant P1 differences across conditions (effects emerged in the N1 window), we analyzed occipital N1, frontocentral FRN, and parietal P3 mean amplitudes. Electrode locations and time windows were determined a priori based on existing literature. N1 amplitude was measured as the mean of PO5, PO6, PO7, PO8, O1, and O2 electrodes across 150–200 ms (Dong et al., 2010; Comesaña et al., 2013). FRN amplitude was measured as the mean of Fz, F1, F2, F3, F4, FCz, FC1, FC2, FC3, and FC4 electrodes across 200–250 ms (Zhang et al., 2013; Ma et al., 2015; Hoy et al., 2021). P3 amplitude was measured as the mean of Pz, P1, P2, P3, P4, CPz, CP1, CP2, and POz electrodes across 300–600 ms (Ma et al., 2015; Wei

& Zhou, 2020).

Behavioral data and ERP amplitudes were analyzed using SPSS Statistics 23.0 with 2 (allocation outcome: below-mean vs. above-mean) × 2 (allocation intention: human vs. computer) repeated-measures ANOVAs. Significant interactions were followed by simple effects tests. Descriptive statistics are reported as mean ± standard error. Significance was set at $p < 0.05$, with ^{2}p reported as effect size.

3. Results

3.1 Behavioral Results

3.1.1 Standard Dictator Game Allocations In the standard DG, participants allocated on average 25.48 ± 2.73% (2.548 ± 0.273 yuan) to others. As shown in Figure 2 [Figure 2: see original paper]A, 47.62% of participants allocated 0–2 yuan, 14.29% allocated 3 yuan, and 38.1% allocated 4 or 5 yuan; no participant allocated more than 5 yuan. Our sample mean closely approximates the social mean allocation value. A t-test revealed no significant difference between our sample mean (25.48%) and the literature's social mean of 28% ($p = 0.180$), confirming sample representativeness. This result supports our selection of 1 and 2 yuan as non-benevolent allocations and 4 and 5 yuan as benevolent allocations.

3.1.2 Hypothesis 1 Test Hypothesis 1 posits that under human allocation conditions, receiving above- (below-) mean allocations leads to above- (below-) mean allocations to third parties. t-tests revealed that after receiving below-mean allocations, participants allocated significantly less than the social mean of 25.48% (1.23 ± 0.18 yuan), $t = -7.33$, $p < 0.001$, $d = 1.131$. After receiving above-mean allocations, allocations to others increased significantly (2.37 ± 0.24 yuan; $t = -8.57$, $p < 0.001$, $d = 1.322$) but remained statistically indistinguishable from the social mean ($t = -0.72$, $p = 0.473$).

Which participants failed to allocate above the social mean after receiving above-mean allocations? We hypothesized that individuals with low altruism would be less likely to pass on benevolence. We therefore conducted a heterogeneity analysis, classifying participants based on their standard DG allocations. Participants allocating 0–2 yuan were classified as low-altruism ($n = 20$; 11 male; $M = 19.85 \pm 0.15$ years), while those allocating 3–5 yuan were classified as high-altruism ($n = 22$; 10 male; $M = 20.23 \pm 0.24$ years).

As shown in Figure 2 [Figure 2: see original paper]C, low-altruism participants allocated significantly more after receiving above-mean versus below-mean allocations (1.11 ± 0.18 yuan vs. 0.51 ± 0.12 yuan; $t = -4.65$, $p < 0.001$, $d = 1.04$), yet this amount remained substantially below the social mean of 2.55 yuan ($t = -7.95$, $p < 0.001$, $d = 1.778$). In contrast, high-altruism participants not only allocated significantly more after receiving above-mean versus below-mean allocations (3.52 ± 0.24 yuan vs. 1.45 ± 0.18 yuan; $t = -9.74$, $p < 0.001$, $d =$

2.076), but this amount also significantly exceeded the social mean ($t = 4.09$, $p = 0.001$, $d = 0.873$). These results support and refine Hypothesis 1: when receiving below-mean allocations, people allocate below the social mean to third parties; when receiving above-mean allocations, high-altruism individuals allocate above the social mean, while low-altruism individuals do not.

3.1.3 Hypothesis 2 Test Hypothesis 2 predicts that above-mean allocations from humans (vs. computers) lead to higher allocations to third parties, while below-mean allocations from humans lead to lower allocations than computers. Repeated-measures ANOVA revealed a significant main effect of allocation outcome, $F(1,41) = 53.15$, $p < 0.001$, $^2p = 0.565$, with participants allocating more after receiving above-mean (2.29 ± 0.23 yuan) versus below-mean allocations (1.43 ± 0.19 yuan). Critically, the allocation outcome × intention interaction was significant, $F(1,41) = 17.67$, $p < 0.001$, $^2p = 0.301$. As shown in Figure 2 [Figure 2: see original paper]D, after receiving below-mean allocations, participants allocated significantly less in human (1.23 ± 0.18 yuan) versus computer conditions (1.63 ± 0.22 yuan), $F(1,41) = 12.25$, $p = 0.001$, $^2p = 0.230$. Conversely, after receiving above-mean allocations, participants allocated significantly more in human (2.37 ± 0.24 yuan) versus computer conditions (2.20 ± 0.22 yuan), $F(1,41) = 4.89$, $p = 0.033$, $^2p = 0.106$. Hypothesis 2 is thus supported.

3.2 ERP Results

3.2.1 N1 Component As shown in Figure 3 [Figure 3: see original paper], the main effect of allocation outcome was not significant, $F(1,41) < 1$, with no difference between above-mean (-1.88 ± 0.38 V) and below-mean allocations (-1.96 ± 0.40 V). The main effect of intention was significant, $F(1,41) = 5.24$, $p = 0.027$, $^2p = 0.113$, with human allocations eliciting larger N1 amplitudes (-2.26 ± 0.39 V) than computer allocations (-1.59 ± 0.42 V). The outcome × intention interaction was not significant, $F(1,41) < 1$. Prediction 1 is supported.

3.2.2 FRN Component As shown in Figure 4 [Figure 4: see original paper], the main effect of allocation outcome was significant, $F(1,41) = 38.68$, $p < 0.001$, $^2p = 0.485$, with below-mean allocations (1.67 ± 0.76 V) eliciting more negative FRN than above-mean allocations (4.24 ± 0.70 V). No significant difference emerged between human (2.73 ± 0.83 V) and computer conditions (3.19 ± 0.63 V), $F(1,41) = 1.19$, $p = 0.282$, $^2p = 0.028$. The outcome × intention interaction was not significant, $F(1,41) = 1.26$, $p = 0.268$, $^2p = 0.030$. Prediction 2 is supported.

3.2.3 P3 Component As shown in Figure 5 [Figure 5: see original paper], the main effect of allocation outcome was significant, $F(1,41) = 25.14$, $p < 0.001$, $^2p = 0.380$, with below-mean allocations (9.42 ± 0.74 V) eliciting larger P3 than above-mean allocations (7.93 ± 0.81 V). The main effect of intention was also

significant, $F(1,41) = 51.33$, $p < 0.001$, $^2p = 0.556$, with human allocations producing larger P3 amplitudes (10.83 ± 0.83 V) than computer allocations (6.53 ± 0.81 V). Crucially, the outcome × intention interaction was significant, $F(1,41) = 4.54$, $p = 0.039$, $^2p = 0.100$. Simple effects tests revealed that under human conditions, below-mean allocations (11.95 ± 0.83 V) elicited larger P3 than above-mean allocations (9.70 ± 0.88 V), $F(1,41) = 27.93$, $p < 0.001$, $^2p = 0.405$. In contrast, under computer conditions, no significant P3 difference emerged between below-mean (6.89 ± 0.77 V) and above-mean allocations (6.17 ± 0.91 V), $F(1,41) = 2.04$, $p = 0.161$, $^2p = 0.047$. Prediction 3 is supported.

4. Discussion

In the real world, upstream indirect reciprocity—where "B receives (non-)benevolence from A and treats third party C similarly"—is ubiquitous (Nowak & Sigmund, 2005). Even in anonymous, one-shot experimental settings that preclude repeated interactions, upstream indirect reciprocity remains robust, transcending neoclassical economic predictions. More importantly, by extending closed reciprocal relationships to third parties, upstream indirect reciprocity serves as an internal force complementing Hayek's spontaneous social order, promoting large-scale cooperation and human social evolution. Scholars have explored this phenomenon from various perspectives (Sun et al., 2022), yet we argue that existing literature has not sufficiently examined its internal mechanisms.

Our primary innovation lies in adopting a framework that diverges from previous studies using equal splits as the reference point for judging A's benevolence (Gray et al., 2014; Hu et al., 2018). Instead, we treat above-mean allocations as benevolent in upstream indirect reciprocity, manipulate allocator role (human vs. computer) to isolate intention effects, and combine this with ERP technology to examine behavioral and neural impacts of allocation outcomes and intentions. Our results support our hypotheses at both behavioral and neural levels.

Why propose a framework different from the equal-split reference point? First, social behavior is often influenced by social norms (Kimbrough & Vostroknutov, 2016; Pereda et al., 2017; Huang et al., 2023). As a social behavior, indirect reciprocity should similarly be shaped by social norms. We argue that identifying A's benevolence cannot be isolated from social context. Social norms include injunctive norms (behaviors that most people approve or disapprove of) and descriptive norms (behavioral patterns that most people actually follow) (Cialdini et al., 1990). Because injunctive norms involve subjective attitudes that are difficult to observe and measure accurately in experiments, we focus on descriptive norms, which require only behavioral data and facilitate empirical testing. We propose that individuals continuously update their understanding of social behaviors through interactions, forming beliefs about the distribution of such behaviors based on experience. When encountering an allocation close to one's learned social mean, no surprise is elicited; substantial deviations above or below this mean, however, may trigger strong reactions. Individuals' willingness to

follow norms depends on their beliefs about others' norm compliance (McBride & Ridinger, 2021)—a mechanism promoting norm convergence. We hypothesize that, in aggregate, people's beliefs about the social mean are unbiased estimators of the true social mean. Our results support this: participants allocated more to others after receiving above-mean allocations (passing on benevolence) and less after receiving below-mean allocations (passing on non-benevolence).

Existing literature's reliance on equal splits likely reflects an (unconscious) adoption of injunctive norms as the judgment standard. For example, in Gray et al.'s (2014) behavioral experiment, participants were told that allocator A could allocate \$0, \$3, or \$6 of \$6 to recipient B, and B would then decide how much of \$6 to allocate to C. Gray et al. treated the equal split as fair, \$0 as greedy, and \$6 as generous, basing this classification on an imagined allocation experiment (Messick & Schell, 1992). That study only examined participants as observers rather than active participants, was not one-shot, and could not exclude reputation effects from repeated interactions. We consider Gray et al.'s (2014) reliance on these findings as insufficiently rigorous. Similarly, Hu et al.'s (2018) fMRI experiment used equal, above-equal, and below-equal trials in equal proportions in the first-stage DG, and had allocators choose between unequal and equal proposals in the second-stage DG. However, meta-analyses show that allocations exceeding 50% are extremely rare (Engel, 2011), and our standard DG found no participant allocating more than 5 yuan. Hu et al.'s (2018) design, with equal frequencies of above- and below-equal allocations, conflicts with participants' real-world experiences and may fail to capture genuine responses to actual allocation patterns, limiting external validity. Even treating equal splits as injunctive norms requires investigating their moral reasonableness, which Gray et al. (2014) and Hu et al. (2018) did not do, making their designs vulnerable to demand effects.

Our second innovation involves manipulating allocator role to examine intention-based indirect reciprocity after controlling for income effects. Results show that below-mean human allocations led participants to allocate even less to third parties than below-mean computer allocations, demonstrating stronger negative indirect reciprocity. One explanation is that below-mean allocations from human A were perceived as intentionally non-benevolent, triggering negative emotional experiences (e.g., anger; Nowak & Sigmund, 1998). Since participants could not retaliate against the non-benevolent allocator, they may have vented negative emotions by passing non-benevolence to third parties. In contrast, below-mean computer allocations lack intentional content, reducing motivation to pass on non-benevolence. Conversely, above-mean human allocations produced stronger positive indirect reciprocity than above-mean computer allocations. Previous research suggests gratitude may mediate indirect reciprocity (Chang et al., 2012). We speculate that when above-mean allocations are perceived as intentional kindness rather than random events, they elicit gratitude (McCullough et al., 2001), increasing willingness to pass kindness to others. Although Hu et al. (2018) also examined allocation outcomes and intentions, their design—using equal frequencies of above-equal allocations that partici-

pants would rarely encounter in reality—may have made it difficult for participants to believe such allocations were genuine. More importantly, by grouping both below-mean and above-mean allocations under "greedy" allocations, they conflated opposing allocation patterns, potentially obscuring intention-based indirect reciprocity effects and leading to non-significant behavioral differences between human and computer conditions.

At the neural level, we found that early selective attention, indexed by N1, was modulated by perceived intention. Human allocations elicited larger N1 than computer allocations, indicating greater selective attention to human allocators. This aligns with Prediction 1 and reflects humans' evolutionary advantage as social animals who are especially attuned to conspecifics (Lin et al., 2014; Moore et al., 2021). Real human interaction demands greater attentional resources, producing larger N1 amplitudes.

FRN amplitude was modulated by whether allocations fell below the social mean: below-mean allocations (1/9 and 2/8 splits) elicited larger FRN than above-mean allocations (4/6 and 5/5 splits). This finding supports Prediction 2. According to reinforcement learning theory, FRN reflects negative reward prediction errors, with worse-than-expected outcomes producing pronounced FRN (Holroyd & Coles, 2002). Our FRN results demonstrate that the brain continuously monitors deviations from social norms (Montague & Lohrenz, 2007), with greater norm violations eliciting larger FRN. These findings also validate our use of the social mean as the threshold for benevolence.

Consistent with Prediction 3a, P3 amplitude was associated with allocation magnitude, with below-mean allocations eliciting larger P3 than above-mean allocations. We also found that human allocations produced larger P3 than computer allocations, consistent with P3's role in subjective value evaluation (Gu et al., 2011; Yeung & Sanfey, 2004). Human allocations carry higher subjective value and demand more attentional resources, generating larger P3 amplitudes (Prediction 3b). Critically, we found an interaction between allocation outcome and intention on P3 (Prediction 3c), providing direct neural evidence for our behavioral results and suggesting that allocation intention and outcome can independently influence reward system neural activity (Bartholow et al., 2006).

Our core contribution lies in using the social mean as the reference point for judging A's benevolence in upstream indirect reciprocity, precisely characterizing the importance of allocation intention, and empirically testing this social-mean-based transmission mechanism at behavioral and neural levels. While Hypothesis 1 received only partial support, our heterogeneity analysis revealed that high-altruism participants fully supported Hypothesis 1, whereas low-altruism participants did not. This finding indicates that individual differences in altruism influence the transmission of above-mean allocations, suggesting future research should more comprehensively examine how heterogeneity in altruism and beliefs about social means affect upstream indirect reciprocity. Future ERP studies with larger samples of high-altruism participants should also re-examine how this factor influences our ERP findings.

This study innovatively investigates social-mean-based upstream indirect reciprocity, combining behavioral and ERP methods to examine how allocation outcomes and intentions influence behavior. Results demonstrate that both factors significantly affect upstream indirect reciprocity: compared to computer allocations, human above-mean allocations increase giving to third parties, while human below-mean allocations decrease giving. At the neural level, early-stage N1 was modulated by intention, with human allocations eliciting larger N1; FRN was modulated by outcome, with below-mean allocations producing larger FRN; and P3 was modulated by the interaction between outcome and intention. Our findings provide new insights into the conditions under which positive and negative indirect reciprocity can be inhibited or activated.

References

Bartholow, B. D., Bushman, B. J., & Sestir, M. A. (2006). Chronic violent video game exposure and desensitization to violence: Behavioral and event-related brain potential data. *Journal of Experimental Social Psychology*, 42(4), 532–539.

Blount, S. (1995). When social outcomes aren't fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes*, 63(2), 131–144.

Boudreau, C., McCubbins, M. D., & Coulson, S. (2009). Knowing when to trust others: An ERP study of decision making after receiving information from unknown people. *Social Cognitive and Affective Neuroscience*, 4(1), 23–34.

Chang, Y.-P., Lin, Y.-C., & Chen, L. H. (2012). Pay it forward: Gratitude in social networks. *Journal of Happiness Studies*, 13(5), 761–781.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117(3), 817–869.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6), 1015–1026.

Comesaña, M., Soares, A. P., Perea, M., Piñeiro, A. P., Fraga, I., & Pinheiro, A. (2013). ERP correlates of masked affective priming with emoticons. *Computers in Human Behavior*, 29(3), 588–595.

de Bruijn, E. R. A., Schubotz, R. I., & Ullsperger, M. (2007). An event-related potential study on the observation of erroneous everyday actions. *Cognitive, Affective, & Behavioral Neuroscience*, 7(4), 278–285.

Dong, G., Hu, Y., & Zhou, H. (2010). Event-related potential measures of the intending process: Time course and related ERP components. *Behavioral and Brain Functions*, 6(1), 15.

Duncan-Johnson, C. C., & Donchin, E. (1977). On quantifying surprise: The

variation of event-related potentials with subjective probability. *Psychophysiology*, 14(5), 456–467.

Engel, C. (2011). Dictator game: A meta study. *Experimental Economics*, 14, 583–610.

Engelmann, D., & Fischbacher, U. (2009). Indirect reciprocity and strategic reputation building in an experimental helping game. *Games and Economic Behavior*, 67(2), 399–407.

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness—Intentions matter. *Games and Economic Behavior*, 62(1), 287–303.

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2), 293–315.

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., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3), 817–868.

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, 295(5563), 2279–2282.

Gong, Y., Yao, L., Chen, X., Xia, Q., Jiang, J., & Du, X. (2022). Group membership modulates fairness consideration among deaf college students—An event-related potential study. *Frontiers in Psychology*, 13, 847917.

Gray, K., Ward, A. F., & Norton, M. I. (2014). Paying it forward: Generalized reciprocity and the limits of generosity. *Journal of Experimental Psychology: General*, 143(1), 247–254.

Gu, R., Lei, Z., Broster, L., Wu, T., Jiang, Y., & Luo, Y. J. (2011). Beyond valence and magnitude: A flexible evaluative coding system in the brain. *Neuropsychologia*, 49(14), 3891–3897.

Hayek, F. A. (1980). *Individualism and economic order* (Reprinted ed.). Chicago, IL: University of Chicago Press.

Herrmann, C. S., & Knight, R. T. (2001). Mechanisms of human attention: Event-related potentials and oscillations. *Neuroscience and Biobehavioral Reviews*, 25(6), 465–476.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, 109(4), 679–709.

Horita, Y., Takezawa, M., Kinjo, T., Nakawake, Y., & Masuda, N. (2016). Transient nature of cooperation by pay-it-forward reciprocity. *Scientific Reports*, 6, 19471.

Hoy, C. W., Steiner, S. C., & Knight, R. T. (2021). Single-trial modeling separates multiple overlapping prediction errors during reward processing in human EEG. *Communications Biology*, 4, 910.

Hu, Y., He, L., Zhang, L., Wölk, T., Dreher, J., & Weber, B. (2018). Spreading inequality: Neural computations underlying paying-it-forward reciprocity. *Social Cognitive and Affective Neuroscience*, 13(6), 578–589.

Huang, X., Li, J., & Ni, Y. (2023). Social norm modulates the enhancement effect of behavioral visibility on altruistic preference. *Acta Psychologica Sinica*, 55(3), 481–495.

Johnson, R. Jr., & Donchin, E. (1980). P300 and stimulus categorization: Two plus one is not so different from one plus one. *Psychophysiology*, 17(2), 167–178.

Kimbrough, E. O., & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, 14(3), 608–638.

Li, J., Pan, J., & Zhu, C. (2020). Inter-brain synchronization is weakened by the power to reject offers in bilateral bargaining games. *Social Cognitive and Affective Neuroscience*, 17(7), 625–633.

Lin, H. Y., Gao, H. W., You, J., Liang, J. F., Ma, J. P., Yang, N., et al. (2014). Larger N2 and smaller early contingent negative variation during the processing of uncertainty about future emotional events. *International Journal of Psychophysiology*, 94(3), 292–297.

Liu, M., Zhou, J., Liu, Y., & Liu, S. (2022). The impact of social comparison and (un)fairness on upstream indirect reciprocity: Evidence from ERP. *Neuropsychologia*, 177, 108398.

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique* (2nd ed.). Cambridge, MA: MIT Press.

Ma, Q., & Hu, Y. (2015). Beauty matters: Social preferences in a three-person ultimatum game. *PLoS ONE*, 10(5), e0125806.

Ma, Q., Hu, Y., Jiang, S., & Meng, L. (2015). The undermining effect of facial attractiveness on brain responses to fairness in the ultimatum game: An ERP study. *Frontiers in Neuroscience*, 9, 77.

Mayer, S. V., Rauss, K., Pourtois, G., Jusyte, A., & Schönenberg, M. (2019). Behavioral and electrophysiological responses to fairness norm violations in antisocial offenders. *European Archives of Psychiatry and Clinical Neuroscience*, 269, 731–740.

McBride, M., & Ridinger, G. (2021). Beliefs also make social-norm preferences social. *Journal of Economic Behavior & Organization*, 191(3), 765–784.

McCullough, M. E., Kilpatrick, S. D., Emmons, R. A., & Larson, D. B. (2001). Is gratitude a moral affect? *Psychological Bulletin*, 127(2), 249–266.

Messick, D. M., & Schell, T. (1992). Evidence for an equality heuristic in social decision making. *Acta Psychologica*, 80(1-3), 311–323.

Miraghaie, A. M., Pouretemad, H., Villa, A. E. P., Mazaheri, M. A., Khosrowabadi, R., & Lintas, A. (2022). Electrophysiological markers of fairness and selfishness revealed by a combination of dictator and ultimatum games. *Frontiers in Systems Neuroscience*, 16, 765720.

Montague, P. R., & Lohrenz, T. (2007). To detect and correct: Norm violations and their enforcement. *Neuron*, 56(1), 14–18.

Moore, M., Katsumi, Y., Dolcos, S., & Dolcos, F. (2021). Electrophysiological correlates of social decision-making: An EEG investigation of a modified ultimatum game. *Journal of Cognitive Neuroscience*, 34(1), 54–78.

Niemand, T., Mai, R., & Kraus, S. (2019). The zero-price effect in freemium business models: The moderating effects of free mentality and price-quality inference. *Psychology & Marketing*, 36(8), 773–790.

Nowak, M. A., & Sigmund, K. (1998). The dynamics of indirect reciprocity. *Journal of Theoretical Biology*, 194(4), 561–574.

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

Pereda, M., Brañas-Garza, P., Rodríguez-Lara, I., & Sánchez, A. (2017). The emergence of altruism as a social norm. *Scientific Reports*, 7(1), 9684.

Qu, C., Wang, Y., & Huang, Y. (2013). Social exclusion modulates fairness consideration in the ultimatum game: An ERP study. *Frontiers in Human Neuroscience*, 7, 505.

Rutte, C., & Taborsky, M. (2007). Generalized reciprocity in rats. *PLoS Biology*, 5(7), e196.

Sigmund, K. (2010). *The Calculus of Selfishness*. Princeton, NJ: Princeton University Press.

Stanca, L., Bruni, L., & Corazzini, L. (2009). Testing theories of reciprocity: Do motivations matter? *Journal of Economic Behavior & Organization*, 71(2), 233–245.

Sun, Y. X., Zhang, J. H., & Li, J. P. (2022). Research progress on indirect reciprocity. *Economic Perspectives*, (1), 146–160.

Wei, H., & Zhou, R. (2020). High working memory load impairs selective attention: EEG signatures. *Psychophysiology*, 57(11), e13643.

Wu, Y., Leliveld, M. C., & Zhou, X. (2011). Social distance modulates recipient's fairness consideration in the dictator game: An ERP study. *Biological Psychology*, 88(2-3), 253–262.

Xie, H., Hu, X., Mo, L., & Zhang, D. (2021). Forgetting positive social feedback is difficult: ERP evidence in a directed forgetting paradigm. *Psychophysiology*, 58(5), e13790.

Yeung, N., & Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *The Journal of Neuroscience*, 24(28), 6258–6264.

Zhong, X., Wang, R., Huang, S., Chen, J., Chen, H., & Qu, C. (2019). The neural correlate of mid-value offers in ultimatum game. *PLoS ONE*, 14(8), e0220622.

Zhang, D. D., Gu, R., Wu, T., Broster, L. S., Luo, Y., Yang, J., & Luo, Y. (2013). An electrophysiological index of changes in risk decision-making strategies. *Neuropsychologia*, 51(8), 1397–1407.

Zhang, Y., Yu, H., Yin, Y., & Zhou, X. (2016). Intention modulates the effect of punishment threat in norm enforcement via the lateral orbitofrontal cortex. *The Journal of Neuroscience*, 36(35), 9217–9226.

Note: Figure translations are in progress. See original paper for figures.