The hidden cost of receiving favors:

A theory of indebtedness

Xiaoxue Gao[1,2 *], *Eshin Jolly* [3], *Hongbo Yu* [4], *Huiying Liu* [5],

Xiaolin Zhou [1,2,6 *], *Luke J. Chang* [3 *]

[1] Shanghai Key Laboratory of Mental Health and Psychological Crisis Intervention, School of Psychology and Cognitive Science, East China Normal University, Shanghai, China, 200062

[2] School of Psychological and Cognitive Sciences, Peking University, Beijing 100871, China

[3] Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755, USA

[4] Department of Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara, CA 93106-9660, USA

[5] Mental Health Education Center, Zhengzhou University, Zhengzhou 450001, Henan, China

[6] PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China

*Correspondence to:

Xiaoxue Gao (gxx114455@gmail.com), Xiaolin Zhou (xz104@pku.edu.cn) and Luke J. Chang (luke.j.chang@dartmouth.edu)

Abstract

Receiving help or a favor from another person can sometimes have a hidden cost for the beneficiary. In this study, we explore these hidden costs by developing and validating a conceptual model of indebtedness across three studies that combine a large-scale online questionnaire, an interpersonal game, computational modeling, and neuroimaging. Our model captures how individuals perceive the altruistic and strategic intentions of the benefactor. These inferences produce distinct feelings of guilt and obligation that together comprise indebtedness and motivate reciprocity. Perceived altruistic intentions convey care and concern and are associated with activity in insula, ventromedial prefrontal cortex and dorsolateral prefrontal cortex, while inferred strategic intentions convey expectations of future reciprocity and are associated with activation in temporal parietal junction and dorsomedial prefrontal cortex. We further develop a neural utility model of indebtedness using multivariate patterns of brain activity that captures the tradeoff between these feelings and reliably predicts reciprocity behavior.

Key words: indebtedness; guilt; obligation; reciprocity; intention; gratitude

Introduction

Giving gifts and exchanging favors are ubiquitous behaviors that provide a concrete expression of a relationship between individuals or groups [1,2]. Altruistic favors convey concern for a partner's well-being and signal a communal relationship such as a friendship, romance, or familial tie [3-5]. These altruistic favors are widely known to foster the beneficiary's positive feeling of gratitude, which can motivate reciprocity behaviors that reinforce the communal relationship [6-9]. Yet in daily life, favors and gifts can also be strategic and imply an expectation of reciprocal exchanges, particularly in more transactive relationships [2,4,5,10-12]. Accepting these favors can have a hidden cost, in which the beneficiary may feel indebted to the favor-doer and motivated to reciprocate the favor at some future point in time [13-21]. These types of behaviors are widespread and can be found in most domains of social interaction. For example, a physician may preferentially prescribe medications from a pharmaceutical company that treated them to an expensive meal [22,23], or a politician might vote favorably on policies that benefit an organization, which provided generous campaign contributions [24]. However, very little is known about the psychological, computational and neural mechanisms underlying this hidden cost of *indebtedness* and how it ultimately impacts the beneficiary.

Immediately upon receipt of an unsolicited gift or favor, the beneficiary is likely to engage in a mentalizing process to infer the benefactor's intentions [25-27]. Does this person care about me? Or do they expect something in return? According to appraisal theory [28-33], these types of cognitive evaluations can evoke different types of feelings, which will ultimately impact how the beneficiary responds. Psychological Game Theory (PGT) [34-36] provides tools for modeling these higher order beliefs about intentions, expectations, and fairness in the context of reciprocity decisions [26,27,37,38]. Actions that are inferred to be motivated by altruistic intentions are more likely to be rewarded, while those thought to be motivated by strategic or self-interested intentions

are more likely to be punished [26,27,37,38]. These intention inferences can produce different emotions in the beneficiary [39]. For example, if the benefactor's actions are perceived to be altruistic, the beneficiary may feel gratitude for receiving help, but this could also be accompanied by the feeling of guilt for personally burdening the benefactor [40-43]. Both feelings motivate reciprocity out of concern for the benefactor, which we refer to as "communal concern" throughout the paper [44,45]. In contrast, if the benefactors' intentions are perceived to be strategic or even duplicitous, then the beneficiary is more likely to feel a sense of obligation [13,14,21,46,47]. Obligation can also motivate the beneficiary to reciprocate [13,14,21,46,47], but unlike communal concern, it arises from external pressures, such as social expectations and reputational costs [48,49] and has been linked to feelings of pressure, burden, anxiety, and resentment [49-51]. Indebtedness has often been considered a unitary construct, defined singularly as either the feeling of guilt for personally burdening the benefactor [40-43] or as a sense of obligation to repay [13,14,21,46,47]. However, in everyday life, inferences about a benefactor's intentions are often mixed and we argue that indebtedness is a superordinate emotion that includes feelings of guilt for burdening the benefactor and social obligation to repay the favor.

We propose a conceptual model to capture how feelings of indebtedness arise and influence reciprocal behaviors (Fig. 1). Specifically, we posit that there are two distinct components of indebtedness - guilt and the sense of obligation, which are derived from appraisals about the benefactor's altruistic and strategic intentions respectively. The guilt component of indebtedness, along with gratitude, arises from appraisals of the benefactor's altruistic intentions (i.e., perceived care from the help) and reflects communal concern. In contrast, the obligation component of indebtedness results from appraisals of the benefactor's strategic intentions (e.g., second-order belief of the benefactor's expectation for repayment). Both feelings of communal concern and obligation motivate the beneficiary's reciprocal behaviors. We examine support for this

conceptual model in Study 1, in which participants describe memories of past emotional experiences in a large-scale online questionnaire using regression analysis and topic modeling.

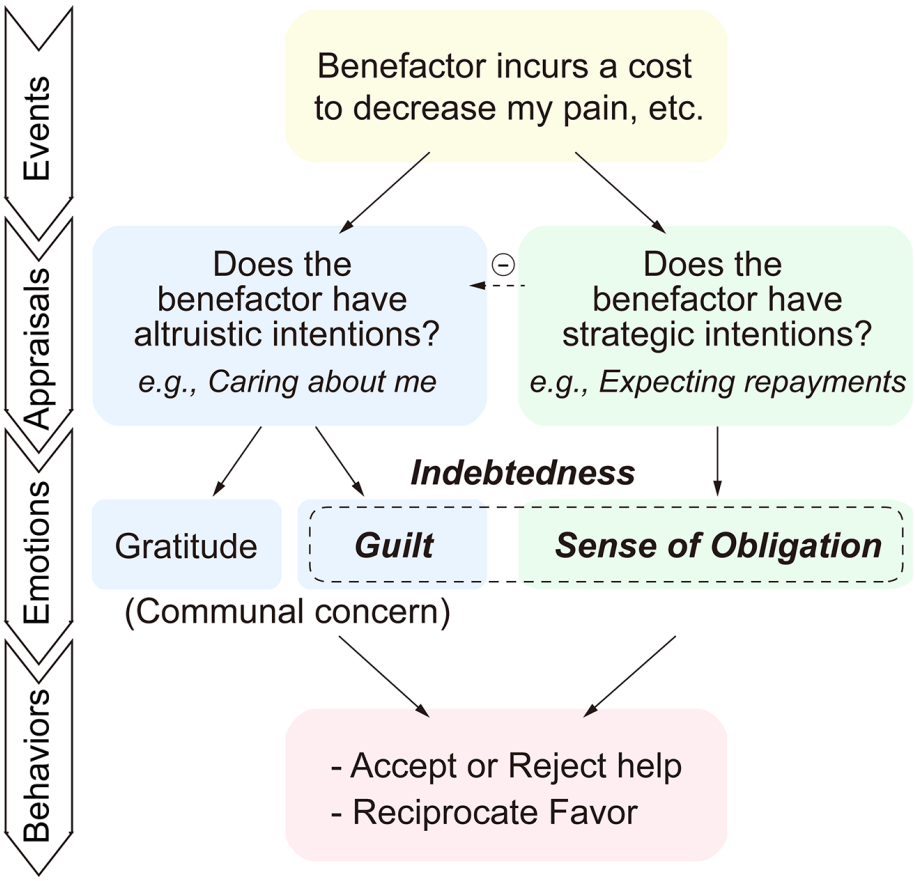

Fig. 1 Conceptual model of indebtedness. *We propose that there are two distinct components of indebtedness, guilt and the sense of obligation, which are derived from appraisals about the benefactor's altruistic and strategic intentions and can differentially impact the beneficiary's reciprocity behaviors. Following an event in which a benefactor provides help to the beneficiary (Yellow), the beneficiary is likely to appraise the benefactor's intentions. The higher the perception of the benefactor's strategic intention, the lower the perception of the benefactor's altruistic intention. The guilt component of indebtedness, along with gratitude, arises from appraisals of the benefactor's altruistic intentions (i.e., perceived care from the help) and reflects communal concern (Blue). In contrast, the obligation component of indebtedness results from appraisals of the benefactor's strategic intentions (e.g., second-order belief of the benefactor's expectation for repayment; Green). Both feelings of communal concern and obligation motivate the beneficiary's reciprocal behaviors (e.g., accept or reject the help and reciprocity after receiving help; Pink).*

In Study 2, we move beyond self-report and focus specifically on how the guilt and obligation components of indebtedness arise and influence behaviors in the context of an interpersonal game. In this study, participants receive electrical shocks and anonymous benefactors (co-players) can choose to provide aid to the participants by spending money to reduce the duration of their pain experience. The participants, in turn, have the opportunity to accept or reject this help and also to reciprocate the benefactor's help by sharing some of their own money back. We experimentally manipulate the participants' beliefs about the benefactors' intentions by providing information about whether or not the co-players are aware that the participants have the opportunity to repay after receiving help. We test the hypothesis that appraisals of altruistic intentions produce guilt as well as gratitude (i.e., communal concern) while appraisals of strategic intentions lead to obligation. Building on previous models of other-regarding preferences [37,38,52], we develop computational models to predict reciprocity and help-acceptance decisions respectively in this interpersonal task by quantifying the tradeoff between the latent motivations of self-interest, communal concern (consisting of guilt & gratitude), and obligation based on appraisals induced by the task (Eq. 1).

In Study 3, we provide further validations of the conceptual model by examining the brain processes associated with the two components of indebtedness by scanning an additional cohort of participants playing the interpersonal game while undergoing functional magnetic resonance imaging (fMRI). Finally, we construct a neural utility model of indebtedness by applying our computational model directly to multivariate brain patterns to demonstrate that neural signals reflect the tradeoff between these feelings and can be used to predict participants' trial-to-trial reciprocity behavior.

Results

Indebtedness is a mixed feeling comprised of guilt and obligation

In Study 1, we explore support for our conceptual model in self-reported experiences of Chinese participants collected via an online questionnaire. First, participants (N = 1,619) described specific events, in which they either accepted or rejected help from another individual and rated their subjective experiences of these events. A regression analysis revealed that both self-reported guilt and obligation ratings independently and significantly contributed to increased indebtedness ratings ($\beta_{guilt} = 0.70 \pm 0.02$, $t = 40.08$, $p < 0.001$; $\beta_{obligation} = 0.40 \pm 0.02$, $t = 2.31$, $p = 0.021$; Fig. 2A-I; Table S1). Second, participants were asked to attribute sources of indebtedness in their daily lives. While 91.9% participants stated that their feelings of indebtedness arose from feeling guilt for burdening the benefactor, 39.2% participants reported feeling obligation based on the perceived ulterior motives of the benefactor (Fig. 2A-II, Fig. S1A). Third, participants were asked to describe their own personal definitions of indebtedness. We applied Latent Dirichlet Allocation (LDA) based topic modeling [53] to the emotion-related words extracted from the 100 words with the highest weight/frequency in the definitions of indebtedness based on annotations from an independent sample of raters (N = 80). We demonstrate that indebtedness is comprised of two latent topics (Fig. S1, B-C). Topic 1 accounted for 77% of the variance of emotional words, including communal-concern-related words such as "guilt," "feel," "feel sorry," "feel indebted," and "gratitude". In contrast, Topic 2 accounted for 23% of the emotional word variance, including words pertaining to burden and negative bodily states, such as "uncomfortable," "uneasy," "trouble," "pressure," and "burden" (Fig. 2A-III). These results support the relationship between indebtedness and the feelings of guilt and obligation posited by our conceptual model.

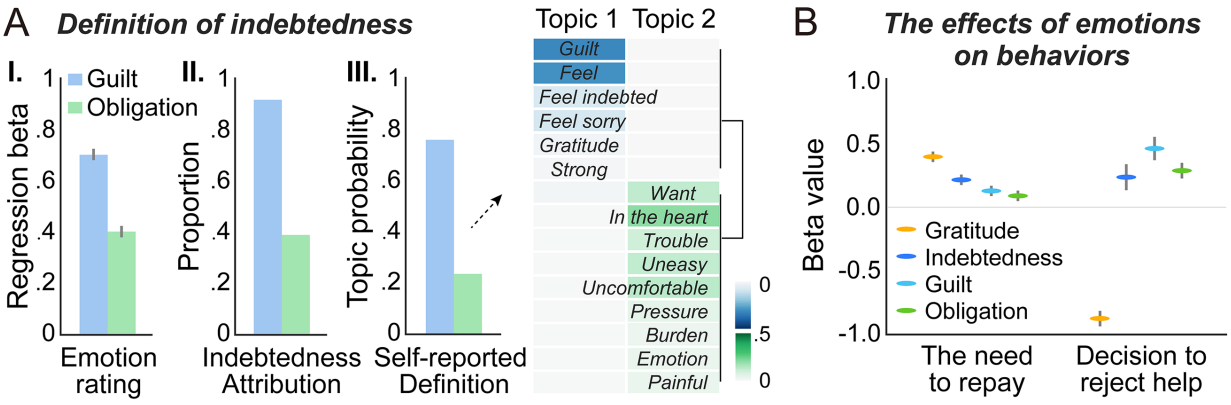

Fig. 2 Subjective experiences of indebtedness in Study 1. (A) *Contributions of guilt and obligation to indebtedness in Study 1 in (I) the emotion ratings in the daily event recalling, (II) attribution of guilt and obligation as source of indebtedness, and (III) topic modeling of the emotional words in self-reported definition of indebtedness. The background color underlying each word represents the probability of this word in the current topic.* ***(B)*** *The influence of emotions on the self-reported need to reciprocate after receiving help and decisions to reject help. Error bars represent ± 1 SE.*

Next, we examined how these distinct feelings relate to participants' self-reported responses to the help (Fig. 2B). Participants described events in which they chose to accept help and reported their experienced emotions. We found that self-reported indebtedness, $\beta = 0.20 \pm 0.04$, $t = 5.60$, $p < 0.001$, guilt ($\beta = 0.12 \pm 0.04$, $t = 2.98$, $p = 0.002$), obligation ($\beta = 0.09 \pm 0.04$, $t = 2.27$, $p = 0.023$), and gratitude ($\beta = 0.38 \pm 0.04$, $t = 9.86$, $p < 0.001$) all independently contributed to participants' reported need to repay after receiving help. Participants also described events, in which they chose to reject help and reported their anticipated counterfactual emotions had they instead accepted the benefactor's help [54]. Decisions to reject help were negatively associated with gratitude ($\beta = -0.87 \pm 0.06$, $t = -13.65$, $p < 0.001$), but positively associated with indebtedness ($\beta = 0.23 \pm 0.10$, $t = 2.40$, $p = 0.017$), guilt ($\beta = 0.46 \pm 0.09$, $t = 5.06$, $p < 0.001$), and obligation ($\beta = 0.28 \pm 0.06$, $t = 4.70$, $p < 0.001$). These results based on subjective experiences indicate that while gratitude, guilt, and obligation all contribute to reciprocating favors, only gratitude appears to be associated with increasing the

likelihood of accepting help. The guilt and obligation components of indebtedness instead appear to be associated with increasing the likelihood of rejecting help.

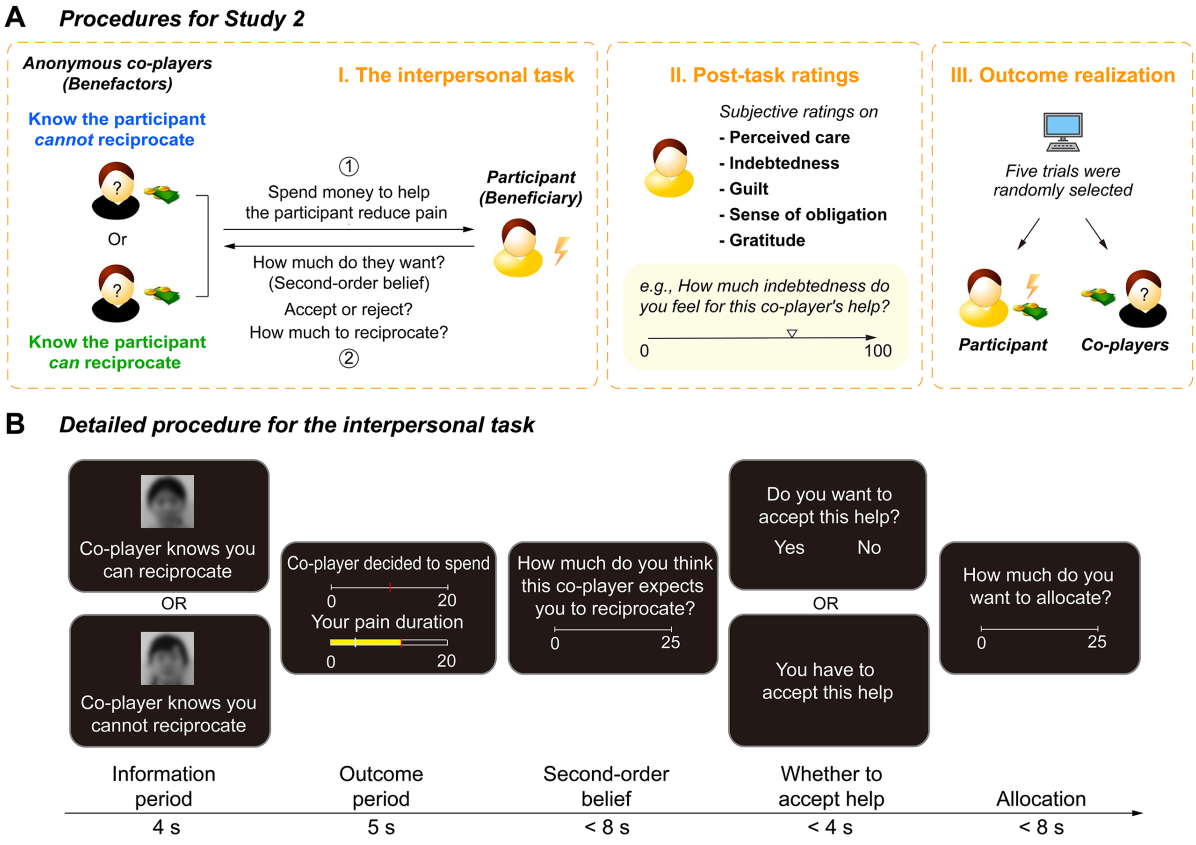

Fig. 3 Experimental procedures for Study 2. ***(A)*** *General procedures. In the interpersonal task (I), the participant was instructed that anonymous co-players (benefactors) made single-shot decisions to help reduce the duration of the participant's pain, and the participant, in turn, (1) reported how much they believed benefactors expected them to reciprocate for their help (i.e., second-order belief), (2) decided whether to accept help, and (3) decided how much money to return to the benefactor. After the interpersonal task, participants recalled how much they believed the benefactor cared for them (i.e., perceived care), as well as their feelings of indebtedness, obligation, guilt, and gratitude in response to the help they received for each trial (II. Post-task ratings). At the end of the experiment, five trials in the interpersonal task were randomly selected to be realized to determine the participant's final amount of pain and payoff, and the selected benefactor's final payoffs (III. Outcome realization).* ***(B)*** *Detailed procedure for the interpersonal task. In each round, a different anonymous same-gender benefactor decided how much of their endowment to spend (i.e., benefactor's cost) to reduce the participant's pain duration. The more the benefactor spent, the more the duration of the participant's pain decreased. Participants indicated how much they thought the benefactor expected them to reciprocate (i.e., second-order belief). In half of the trials, the participant had to passively accept the benefactor's help; in the other half, the participant could freely decide whether to accept or reject the benefactor's help. Finally, at the end of each trial, the participant decided how much of their own endowment he/she wanted to*

allocate to the benefactor as reciprocity for their help. Unbeknownst to participants, benefactors' decisions (i.e., benefactor's cost) were pre-determined by the computer program (Table S2). We manipulated the perception of the benefactor's intentions by providing additional information about whether the benefactor knew the participant could (i.e., Repayment possible condition), or could not (i.e., Repayment impossible condition) reciprocate after receiving help. In fact, participants could reciprocate in both conditions during the task.

Benefactor's intentions cause diverging components of indebtedness

We next sought to more specifically examine how indebtedness impacts behavior in the context of a laboratory-based task involving interactions between participants in Study 2a (N=51, Fig. 3). In this task, participants were randomly paired with a different anonymous same-gender co-player (benefactor) in each trial and were instructed that they would receive 20 seconds of pain stimulation in the form of a burst of medium intensity electrical shocks. The participant was informed that each benefactor had been endowed with 20 yuan (~ $3.1 USD) and made a decision about how much to spend from this endowment to reduce the duration of pain experienced by the participant (i.e., ***benefactor's cost***) during a separate lab visit. Unbeknownst to participants, each benefactor's cost was pre-determined by a computer program (Table S2). After seeing how much money the benefactor chose to spend, the participant reported how much they believed this benefactor expected them to reciprocate (i.e., second-order belief of the benefactor's expectation for repayment). In half of the trials, the participant had to passively accept the benefactor's help; in the other half, the participant could freely decide whether to accept or reject the benefactor's help. Finally, at the end of each trial, the participant decided how much of their own 25 yuan endowment (~ $3.8 USD) he/she wanted to allocate to the benefactor as reciprocity for their help. We experimentally manipulated the participant's beliefs about the benefactor's intentions by providing additional information regarding the benefactor's expectations of reciprocation. Each participant was instructed that before making decisions, some benefactors knew that the participant would be endowed with 25 yuan and could decide whether to allocate some endowments to them as reciprocity (i.e., ***Repayment possible***

condition), whereas the other benefactors were informed that the participant had no chance to reciprocate after receiving help (i.e., ***Repayment impossible condition***). In fact, participants could reciprocate in both conditions during the task. After the task, all trials were displayed again in a random order and participants recalled how much they believed the benefactor cared for them (i.e., perceived care), as well as their feelings of indebtedness, obligation, guilt, and gratitude in response to the help they received for each trial. To ensure incentive compatibility, five trials were randomly selected to be enacted and participants received the average number of shocks and money based on their decisions at the end of the experiment. We ran an additional version of this experiment (Study 2b, N = 57), in which we systematically varied the exchange rate of how much it cost the benefactor to reduce the participant's duration of pain (i.e., help efficiency). However, we did not observe any significant interaction effect between efficiency and any of other experimental variables in Study 2b and chose to combine these two studies for all Study 2 analyses (N=108, Table S3-1; see *SI Results* and Tables S3-2 to S3-4 for separate results).

Our experimental manipulation successfully impacted participants' appraisals of the benefactors' hidden intentions behind their help. Participants reported increased second-order beliefs of the benefactor's expectations for repayment (β = 0.53±0.03, t = 15.71, $p < 0.001$) and decreased perceived care (β = -0.31±0.02, t = -13.89, $p < 0.001$) (Fig. 4A) when the participant believed the benefactor knew they could reciprocate (Repayment possible) compared to when they could not reciprocate (Repayment impossible). Both of these effects were amplified as the benefactor spent more money to reduce the participant's duration of pain (Fig. 4, B-C; second-order belief: β = 0.22±0.02, t = 13.13, $p < 0.001$; perceived care: β = -0.08±0.01, t = -6.64, $p < 0.001$). In addition, perceived care was negatively associated with second-order beliefs (β = -0.44±0.04, t = -11.29, $p < 0.001$) controlling for the effects of experimental variables (i.e., extra information about benefactor's intention, cost, and efficiency).

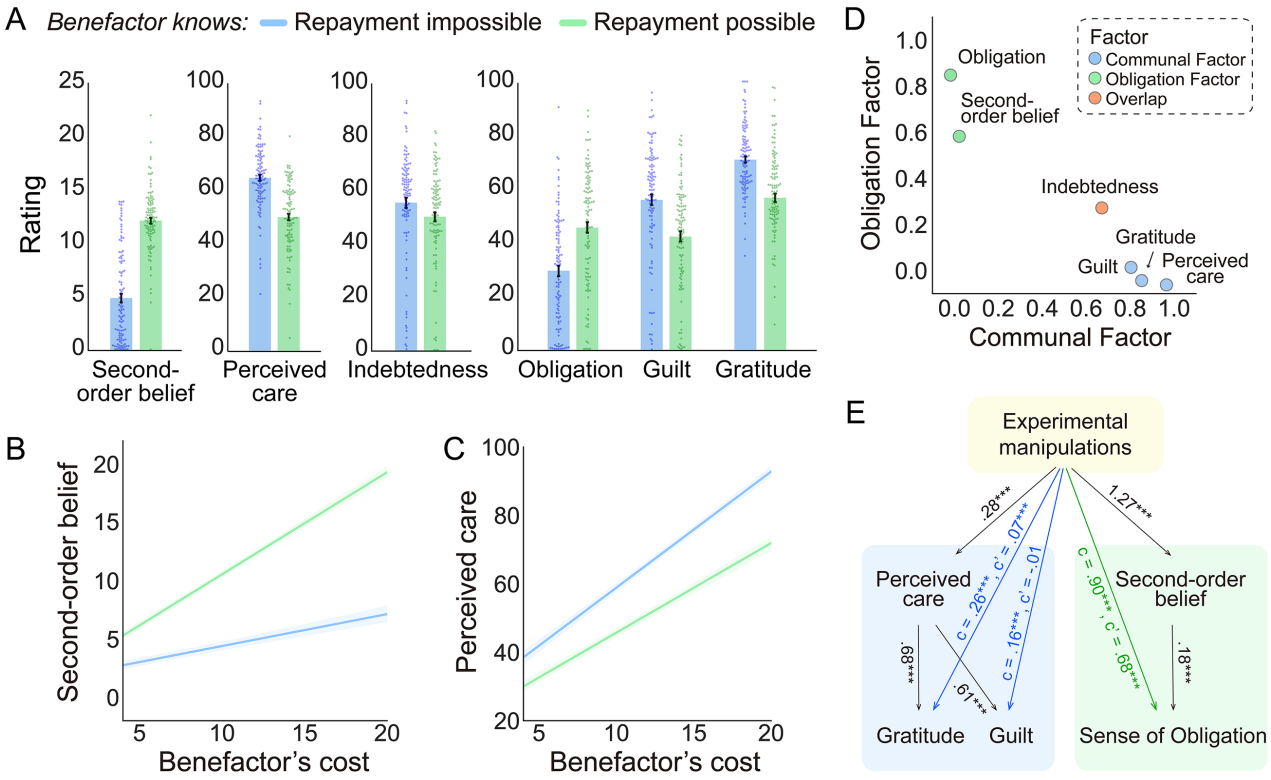

Fig. 4 Appraisals and emotional responses to benefactor's help with different intentions. ***(A)*** *Participant's appraisals (i.e., second-order belief of how much the benefactor expected for repayment and perceived care) and emotion ratings (indebtedness, obligation, gratitude, and guilt) in Repayment impossible and Repayment possible conditions. Each dot represents the average rating in the corresponding condition for each participant.* ***(B and C)*** *Participant's second-order beliefs of how much the benefactor expected for repayment and perceived care plotted as functions of extra information about benefactor's intention (Repayment impossible vs. Repayment possible) and benefactor's cost.* ***(D)*** *Factor analysis showed that participants' appraisals and emotions could be explained by two independent factors, which appeared to reflect two distinct subjective experiences. The Communal Factor reflects participants' perception that the benefactor cared about their welfare and resulted in emotions of gratitude and guilt, while the Obligation Factor reflects participants' second-order beliefs about the benefactor's expectation for repayment and the sense of obligation.* ***(E)*** *Simplified schematic representation of mediation analysis. See full model in Fig. S3C. Results showed that second-order beliefs and perceived care appraisals differentially mediated the effects of the experimental manipulations on emotional responses. Second-order belief mediated the effects of the experimental manipulations on the sense of obligation, while perceived care mediated the effects of experimental manipulations on gratitude and guilt. Error bars represent ± 1 SE.*

The belief manipulation not only impacted the participants' appraisals, but also their feelings. Our conceptual model predicts that participants will feel indebted to benefactors who spent money to reduce their pain, but for different reasons depending on the perceived intentions of the benefactors. Consistent with this prediction, participants reported feeling indebted in both conditions, but slightly more in the Repayment impossible compared to the Repayment possible condition (Fig. 4A, Fig. S2A, $\beta = 0.09 \pm 0.03$, $t = 2.98$, $p = 0.003$). Moreover, participants reported feeling greater obligation (Fig. 4A, Fig. S2B, $\beta = 0.30 \pm 0.03$, $t = 9.28$, $p < 0.001$), but less guilt ($\beta = -0.25 \pm 0.02$, $t = -10.30$, $p < 0.001$), and gratitude ($\beta = -0.27 \pm 0.02$, $t = -13.18$, $p < 0.001$) in the Repayment possible condition relative to the Repayment impossible condition (Fig. 4A, Fig. S2, C-D). Similar to the appraisal results, these effects were magnified as the benefactor's cost increased (Fig. S2, B-D; obligation: $\beta = 0.11 \pm 0.01$, $t = 8.85$, $p < 0.001$; guilt: $\beta = -0.05 \pm 0.01$, $t = -4.28$, $p < 0.001$; gratitude: $\beta = -0.06 \pm 0.01$, $t = -4.20$, $p < 0.001$).

We conducted two separate types of multivariate analyses to characterize the relationships between appraisals and emotions. First, exploratory factor analysis (EFA) on the subjective appraisals and emotion ratings in Study 2 revealed that 66% of the variance in ratings could be explained by two factors (Fig. 4D, and Fig. S2E; Fig. S3, A-B). The Communal Factor reflected participants' perception that the benefactor cared about their welfare and resulted in emotions of guilt and gratitude, while the Obligation Factor reflected participants' second-order beliefs about the benefactor's expectation for repayment and the sense of obligation. Interestingly, indebtedness moderately loaded on both factors supporting its mixed relationship to guilt and obligation. Second, a mediation analysis revealed that, second-order beliefs and perceived care appraisals differentially mediated the effects of the experimental manipulations on emotional responses (total indirect effect = 0.59 ± 0.04, $Z = 14.49$, $p < 0.001$; Fig. 4E and Fig. S3C). Second-order beliefs mediated the effects of the experimental manipulations on

obligation (Indirect effect = 0.22±0.03, Z = 7.18, p < 0.001), while perceived care mediated the effects of the experimental manipulations on guilt (Indirect effect = 0.17±0.01, Z = 13.23, p < 0.001) and gratitude (Indirect effect = 0.19±0.01, Z = 13.72, p < 0.001). Together, these results provide further support for the predictions of our conceptual model that indebtedness is comprised of two distinct feelings. The guilt component arises from the belief that the benefactor acts from altruistic intentions (i.e., perceived care), while the obligation component arises when the benefactor's intentions are perceived to be strategic (e.g., expecting repayment).

Beneficiary's behaviors are influenced by benefactor's intentions

Next, we examined participants' behaviors in response to receiving help from a benefactor. Specifically, we were interested in how much participants would reciprocate after receiving the favor and also whether they might outright reject the benefactor's help given the opportunity. We found that participants reciprocated more money as a function of the amount of help received from the benefactor, β = 0.63±0.02, t = 25.60, p < 0.001 (Fig. 5A). This effect was slightly enhanced in the Repayment impossible condition relative to the Repayment possible condition, β = 0.03±0.01, t = 2.99, p = 0.003. A logistic regression revealed that when given the chance, participants were more likely to reject help in the Repayment possible condition when they reported more obligation (rejection rate = 0.37±0.10), compared to the Repayment impossible condition (rejection rate = 0.30±0.03), β = 0.27±0.08, z = 3.64, p < 0.001 (Fig. 5B). Moreover, as the benefactor's cost increased, participants were less likely to reject the help (β = -0.65±0.13, z = -5.16, p < 0.001). No significant interaction effect between condition and benefactor's cost was observed (β = 0.07±0.07, z = 1.08, p = 0.279).

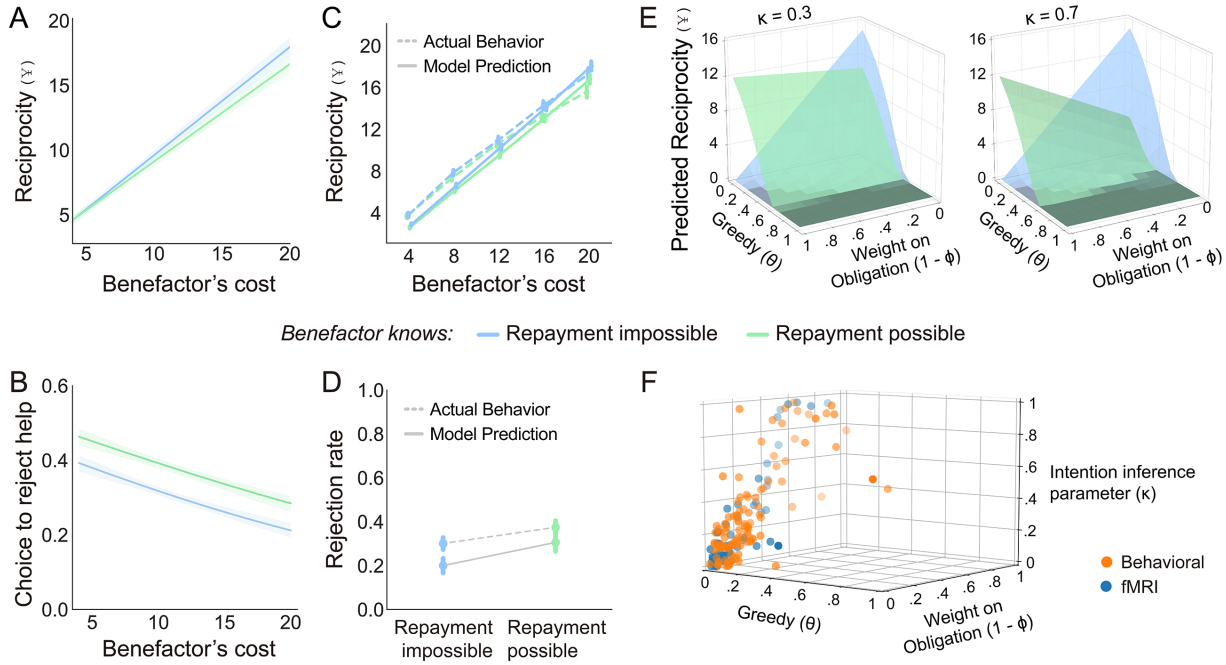

Fig. 5 Computational models of indebtedness. ***(A)*** *Participants' reciprocity behavior in each trial plotted as function of extra information about benefactor's intention (Repayment impossible vs. Repayment possible) and benefactor's cost.* ***(B)*** *Participants' decisions of accept or reject help in each trial plotted as a logistic function of extra information about benefactor's intention and benefactor's cost.* ***(C)*** *The observed amounts of reciprocity after receiving help and predictions generated by the reciprocity model at each level of the benefactor's cost in Repayment impossible and Repayment possible conditions.* ***(D)*** *The observed rates of rejecting help and predictions generated by the help-acceptance model in Repayment impossible and Repayment possible conditions.* ***(E)*** *Model simulations for predicted reciprocity behavior in Repayment impossible and Repayment possible conditions at different parameterizations. The y axis shows the average values of the predicted amount of reciprocity across all levels of benefactor's cost.* ***(F)*** *Best fitting parameter estimates of the computational model for reciprocity decisions for each participant. Error bars represent the standard error of means.*

Computational models of how indebtedness impacts behaviors

Building on our conceptual model of indebtedness, we developed two computational models using a Psychological Game Theoretic framework [34-36] to predict reciprocity and help-acceptance decisions that maximize the beneficiary's expected utility based on the competing latent motivations of self-interest, communal concern (i.e., guilt and gratitude), and obligation (Eq. 1).

$$U(D_B) = \theta_B * \pi_B + (1 - \theta_B) * (\phi_B * U_{Communal} + (1 - \phi_B) * U_{Obligation}) \quad \textbf{Eq.1}$$

The central idea is that upon receiving a favor D_A from a benefactor A, the beneficiary B chooses an action D_B that maximizes his/her overall utility U. This utility is comprised of a mixture of values arising from self-interest π weighted by a greed parameter θ, and feelings of communal concern $U_{Communal}$ and obligation $U_{Obligation}$, which are weighted by the parameter ϕ. Larger ϕ values reflect the beneficiary's higher sensitivity to feelings of communal concern relative to obligation. $U_{Communal}$ reflects a linear combination of guilt and gratitude components (see *Methods*).

The *reciprocity model* (Model 1.1) predicts the participant's amount of money reciprocated to the benefactor, while the *help-acceptance model* (Model 2.1) predicts binary decisions to accept or reject help. Though the two models are conceptually similar, the values of $U_{Communal}$ and $U_{Obligation}$ are computed slightly differently due to differences in the types of data (i.e., continuous vs. binary decisions) and how appraisals are inferred. It is important to note that in the reciprocity model, we are unable to distinguish between the separate motivations of guilt and gratitude because both positively contribute to reciprocity. In contrast, based on the findings from Study1, we divide up the parameter space for the help-acceptance model such that $\phi > 0$ indicates a preference for gratitude and motives accepting the help, while $\phi < 0$ indicates a preference for guilt and motives rejecting the help.

For both models, we define $U_{Obligation}$ as the appraisal of the amount of money that B believes A expect them to return (i.e., B's second-order beliefs E_B'') normalized by B's endowment size γ_B.

$$U_{Obligation} = \begin{cases} -\left(\frac{E_B'' - D_B}{\gamma_B}\right)^2 & \text{Reciprocity model} \\ -\frac{E_B''}{\gamma_B} & \text{Help-acceptance model} \end{cases} \quad \textbf{Eq. 2}$$

where E_B'' is operationalized as D_A in the Repayment possible condition and zero in the Repayment impossible condition.

$$E_B'' = \begin{cases} 0 & \text{Repayment impossible condition} \\ D_A & \text{Repayment possible condition} \end{cases} \quad \textbf{Eq. 3}$$

In contrast, we define $U_{Communal}$ in terms of the appraisal of how much B believes A cares about their welfare (i.e., perceived care ω_B).

$$U_{Communal} = \begin{cases} -(\frac{\omega_B * \gamma_B - D_B}{\gamma_B})^2 & \text{Reciprocity model} \\ \omega_B & \text{Help-acceptance model} \end{cases} \quad \textbf{Eq. 4}$$

We assume that B infers perceived care ω_B proportional to how much A spent D_A from his/her endowment γ_A and that this effect might be mitigated by the amount of money B believes A expects them to return (i.e., second-order belief E_B'').

$$\omega_B = \frac{D_A - \kappa_B * E_B''}{\gamma_A} \quad \textbf{Eq. 5}$$

where κ reflects the degree to which the perceived strategic intention E_B'' reduces the perceived altruistic intention ω_B. See details for the models in *Methods*.

The reciprocity model

We performed a rigorous validation of our reciprocity model (Model 1.1) across a variety of different types of evaluations. First, we were interested in how well our computational model for reciprocity captured trial-to-trial reciprocity decisions. Model parameters were estimated by minimizing the sum of squared error between the model predicted behaviors and participants' reciprocity decisions separately for every participant. Our computational model was able to successfully predict participants'

continuous reciprocity decisions after receiving help ($r^2 = 0.81$, $p < 0.001$; Fig. 5C; Fig. S6; Table S11) and significantly outperformed other plausible models, such as: (a) a model with linear formulations of utilities for self-interest, communal concern, and obligation (Model 1.2), (b) models that solely included the term for communal concern (Model 1.3) or obligation (Model 1.4) besides the self-interest term, (c) models with separate parameters for self-interest, communal concern, and obligation (Model 1.5 and Model 1.6), (d) a model that assumes participants reciprocate purely based on the benefactors helping behavior (i.e., tit-for-tat) [37,38] (Model 1.7), and (e) a model that assumes that participants are motivated to minimize inequity in payments [52,55] (Model 1.8) (Table S7). Parameter recovery tests indicated that the parameters of the reciprocity model were highly identifiable (correlation between true and recovered parameters: reciprocity $r = 0.94 \pm 0.07$, $p < 0.001$; Table S9).

A simulation of the reciprocity model across varying combinations of the θ, ϕ and κ parameters revealed diverging predictions of the beneficiaries' response to favors in Repayment impossible and Repayment possible conditions (Fig. 5E). Not surprisingly, greedier individuals (higher θ) are less likely to reciprocate others' favors. However, reciprocity changes as a function of the tradeoff between communal and obligation feelings based on ϕ and interacts with the intention inference parameter κ. Increased emphasis on obligation corresponds to increased reciprocity to favors in the Repayment possible condition, but decreased reciprocity in the Repayment impossible condition; this effect is amplified as κ increases. We found that most participants had low θ values (i.e., greed), but showed a wide range of individual differences in κ and ϕ parameters (Fig. 5F). Interestingly, the degree to which the perceived strategic intention reduced the perceived altruistic intention during intention inference κ, was positively associated with the relative weight on obligation ($1 - \phi$) during reciprocity ($r = 0.79$, $p < 0.001$). This suggests that the participants who cared more about the benefactor's strategic

intentions also tended to be motivated by obligation when deciding how much money to reciprocate.

Beyond just simply predicting behaviors, we conducted additional validations to assess how well our reciprocity model's predictions of second-order belief (E_B''; Eq. 3) and perceived care (ω_B; Eq. 5) were able to capture trial-to-trial variations in participants' self-reported ratings of appraisals and feelings. A regression analysis showed that the reciprocity model's representations of E_B'' and ω_B were associated with trial-to-trial variations in self-reported values of second-order belief of the benefactor's expectation for repayment ($\beta = 0.68 \pm 0.03$, $t = 21.48$, $p < 0.001$; Fig. S5, A-B) and perceived care ($\beta = 0.72 \pm 0.03$, $t = 26.76$, $p < 0.001$; Fig. S5, C-D), respectively. Moreover, κ appeared to successfully capture individual differences as participants who reported an overall higher level of perceived care were also observed to have a higher overall level of ω_B ($r = 0.27$, $p = 0.004$).

We further assessed if the reciprocity model's representations of perceived care (ω_B) and second-order belief (E_B'') appraisals corresponded to self-reported communal and obligation feelings. Supporting our predictions, the reciprocity model's predictions of ω_B significantly predicted self-reported guilt ratings ($\beta = 0.47 \pm 0.03$, $t = 17.21$, $p < 0.001$) as well as the Communal Factor scores obtained from EFA in Fig. 4D ($\beta = 0.81 \pm 0.03$, $t = 25.81$, $p < 0.001$), while the model predictions of E_B'' significantly predicted self-reported obligation ratings ($\beta = 0.38 \pm 0.03$, $t = 12.67$, $p < 0.001$) and the Obligation Factor scores ($\beta = 0.64 \pm 0.06$, $t = 15.97$, $p < 0.001$).

The help-acceptance model

Next, we evaluated how well the help-acceptance model (Model 2.1) was able to capture participants' trial-to-trial decisions of whether or not to accept the benefactor's help. We estimated the parameters by maximizing the log-likelihood of the predicted

probability of the chosen option (accept or reject) separately for each participant. Overall, we found that our model was able to predict participants' decisions to accept or reject help (accuracy = 80.37%; Fig. 5D; Fig. S6; Table S12). The help-acceptance model outperformed models with separate parameters for self-interest, communal concern, and obligation (Model 2.4 and 2.5), but did not significantly outperform models that solely included terms for communal concern (Model 2.2) or obligation (Model 2.3) (Table S8). This likely stems from a slight instability in the parameterization of the model (see *Methods* and *Discussion*), which is confirmed by the moderate level of identifiability indicated by the parameter recovery tests ($r = 0.43 \pm 0.40$, $p < 0.001$; and Table S10).

Communal concern and obligation involve distinct neural processes

Next, in Study 3 (N = 53), we explored the neural basis of indebtedness and examined whether the processing of communal concern and obligation involve differential brain processes as suggested by our conceptual model. Participants completed the same task as Study 2 while undergoing fMRI scanning, except that they were unable to reject help. First, we successfully replicated all of the behavioral results observed in Study 2 (see detailed statistics in Tables S1 and S4, and Figs. S7 and S8). In addition, we found that the two-factor EFA model we estimated using the self-report data in Study 2 generalized well to the independent sample in Study 3 using confirmatory factor analysis (CFA; Fig. S7G), with comparative fit indices exceeding the > 0.9 acceptable threshold (CFI = 0.986, TLI = 0.970) and the root mean square error of approximation and the standardized root mean squared residual were within the reasonable fit range of < 0.08 (RMSEA = 0.079, SRMR = 0.019) [56-58].

Second, we performed univariate analyses to identify brain processes during the Outcome period (Fig. 3B), where participants learned about the benefactor's decision to help. Using a model-based fMRI analytic approach [59], we fit three separate general

linear models (GLMs) to each voxel's timeseries to identify brain regions that tracked different components of the computational model. These included trial-by-trial values for: (1) the amount of reciprocity, (2) communal concern, which depended on the perceived care from the help (ω_B), and (3) obligation, which depended on the second-order belief of the benefactor's expectation for repayment (E_B'') (see *Methods*). We found that trial-by-trial reciprocity behavior correlated with activity in bilateral dorsal lateral prefrontal cortex (dlPFC), bilateral inferior parietal lobule (IPL), precuneus, and bilateral inferior temporal gyrus (ITG) (Fig. 6A, Table S13). Trial-by-trial communal feelings tracked with activity in the anterior insula, ventromedial prefrontal cortex (vmPFC), precuneus, bilateral dlPFC, and bilateral ITG (Fig. 6B; Table S13). The processing of obligation was associated with activations in dorsomedial prefrontal cortex (dmPFC) and left temporo-parietal junction (TPJ) (Fig. 6C, Table S13).

To aid in interpreting these results, we performed meta-analytic decoding [60] using Neurosynth [61]. Reciprocity-related activity was primarily associated with "Attention," "Calculation," and "Memory" terms. Communal feelings related activity was similar to the reciprocity results, but was additionally associated with "Default mode" term. Obligation activity was highly associated with terms related to "Social," "Theory of mind (ToM)," and "Memory" (Fig. 6D). Together, these neuroimaging results reveal differential neural correlates of feelings of communal concern and obligation and support the role of intention inference in the generation of these feelings proposed by our conceptual model. The processing of communal feelings was associated with activity in vmPFC, an area in default mode network that has been linked to gratitude [62-64], positive social value and kind intention, [65,66] as well as the insula, which has been previously related to guilt [54,67,68]. In contrast, the processing of obligation was associated with activations of theory of mind network, including dmPFC and TPJ, which is commonly observed when representing other peoples' intentions or strategies [66,69,70].

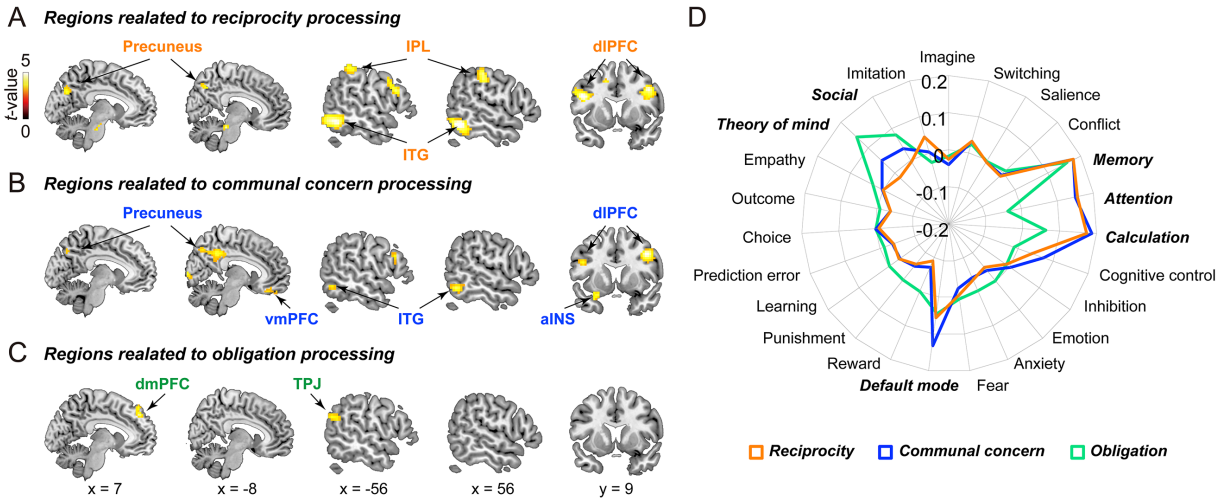

Fig. 6 Neural processes associated with reciprocity, communal concern and obligation. ***(A)*** *Brain regions responding parametrically to trial-by-trial amounts of reciprocity.* ***(B)*** *Brain regions responding parametrically to trial-by-trial communal concern, which depended on the perceived care from the help (ω_B).* ***(C)*** *Brain regions identified in the parametric contrast for obligation (E_B''), the responses of which monotonically increased in the Repayment possible condition relative to the Repayment impossible condition.* ***(D)*** *Meta-analytical decoding for the neural correlates of reciprocity, communal concern and obligation, respectively. All brain maps thresholded using cluster correction FWE $p < 0.05$ with a cluster-forming threshold of $p < 0.001$* [71].

Neural utility model of indebtedness predicts reciprocity behavior

Finally, we sought to test whether we could use signals directly from the brain to construct a utility function and predict reciprocity decisions (Fig. 7A). Using brain activity during the Outcome period of the task (Fig. 3B), we trained two whole-brain models using principal components regression with 5-fold cross-validation [72-74] to predict the appraisals associated with communal concern (ω_B) and obligation (E_B'') separately for each participant. These whole-brain patterns were able to successfully predict the model representations of these feelings for each participant on new trials, though with modest effect sizes (communal concern pattern: average $r = 0.21 \pm 0.03$, $p < 0.001$; obligation pattern: average $r = 0.10 \pm 0.03$, $p = 0.004$; Fig. 7A). Moreover, these patterns appear to be capturing distinct information as they were not spatially correlated, $r = 0.03$, $p = 0.585$. These results did not simply reflect differences between

the Repayment possible and Repayment impossible conditions as the results were still significant after controlling for this experimental manipulation (communal concern: average $r = 0.18 \pm 0.02$, $p < 0.001$; obligation: average $r = 0.04 \pm 0.02$, $p < 0.024$). Furthermore, we were unable to successfully discriminate between these two conditions using a whole brain classifier (*accuracy* = 55.0 ± 1.25%, permutation $p = 0.746$).

Next, we assessed the degree to which our brain models could account for reciprocity behavior. We used cross-validated neural predictions of communal concern (ω_B) and obligation (E_B") feelings as inputs to our computational model of reciprocity behavior instead of the original terms (Eq. 6):

$$U(D_B) = \theta_B * \pi_B + (1 - \theta_B) * (\phi_B * \vec{\beta}_{map} \cdot \vec{Communal}_{map} + (1 - \phi_B) * \vec{\beta}_{map} \cdot \vec{Obligation}_{map}) \quad \textbf{Eq. 6}$$

where $\vec{\beta}_{map}$ refers to the pattern of brain activity during the Outcome period (Fig. 3B) of a single trial and $\vec{Communal}_{map}$ and $\vec{Obligation}_{map}$ refer to the multivariate brain models predictive of each participant's communal concern and obligation utilities respectively. We were able to reliably predict reciprocity behavior with our computational model informed only by predictions of communal and obligation feelings derived purely from brain responses (average $r = 0.19 \pm 0.02$, $p < 0.001$, AIC = 317.70 ± 5.00; Fig. 7B). As a benchmark, this model numerically outperformed a whole-brain model trained to directly predict reciprocity (average $r = 0.18 \pm 0.03$, $p < 0.001$, AIC = 317.54 ± 5.00; Fig. 7A), but this difference only approached statistical significance, $t_{52} = 1.64$, $p = 0.108$.

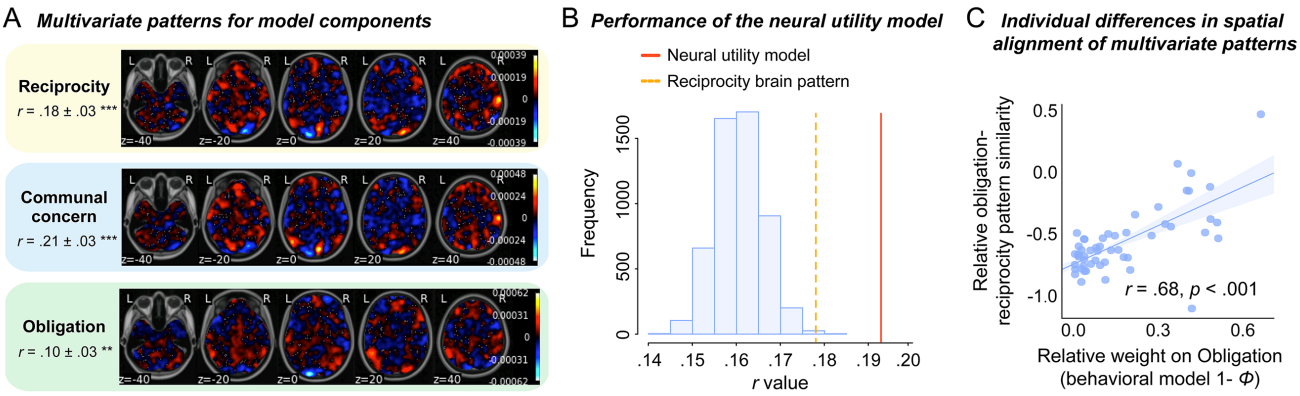

Fig. 7 Neural utility model of indebtedness. (A) *Unthresholded multivariate patterns used to predict the amounts of reciprocity, trial-by-trial communal concern (ω_B) and obligation (E_B'') separately.* ***(B)*** *We assessed the importance of the participant-specific model parameters estimated from the neural utility model (i.e., ϕ) by generating a null distribution of predictions after permuting the estimated ϕ parameter across participants 5,000 times. The red line indicates the performance of our neural utility model (r value of prediction), and the yellow line indicates the performance of the whole-brain model trained to directly predict reciprocity. The subject-specific weightings were important in predicting behavior as our neural utility model significantly outperformed a null model using parameters estimated for a different participant.* ***(C)*** *The relationship between the relative weight on obligation ($1 - \phi$) derived from behavior and a neurally derived metric of how much obligation vs. communal feelings influenced reciprocity behavior.*

We performed several additional validations of the neural utility model to demonstrate its overall performance. First, we compared the parameter ϕ, which reflects the tradeoff between communal concern and obligation estimated from the neural utility model and found that it strongly correlated with the same parameter estimated from the behavioral computational model across participants, $r = 0.88$, $p < 0.001$. Second, we assessed the individual specificity of ϕ derived from the neural utility model, to test how uniquely sensitive individuals are to communal concern versus obligation. To do so, we generated a null distribution of predictions after permuting the estimated ϕ parameter across participants 5,000 times. We found that the participant-specific weightings were highly important in predicting behavior as our neural utility model significantly outperformed null models using randomly shuffled ϕ parameters, $p < 0.001$ (Fig. 7B).

Third, we tested how well our neural-utility model reflected the trade-off between an individual's feelings of communal concern or obligation estimated from the behavioral model. We hypothesized that the relative influence of a particular feeling on behavior should be reflected in the spatial alignment of their corresponding brain patterns [75] (see *Methods*). Our results support this hypothesis. Participants who cared more about obligation relative to communal concern (higher behavioral $1 - \phi$) also exhibited greater spatial alignment between their obligation and reciprocity brain patterns relative to communal concern and reciprocity patterns, $r = 0.68$, $p < 0.001$ (Fig. 7C). These results provide evidence at the neural level indicating that individuals appear to trade-off between feelings of communal concern and obligation when deciding how much to reciprocate after receiving help from a benefactor.

Discussion

Gift-giving, favor-exchanges, and providing assistance are behavioral expressions of relationships between individuals or groups. While favors from friends and family often engender reciprocity and gratitude, they can also elicit guilt in a beneficiary who may feel that they have burdened a benefactor. Favors in more transactive relationships, however, can evoke a sense of obligation in the beneficiary to repay the favor. In this study, we sought to develop a conceptual model of indebtedness that outlines how appraisals about the intentions behind a favor are critical to the generation of these distinct feelings, which in turn motivates how willing individuals are to accept or reject help and ultimately reciprocate the favor.

We provide a systematic validation of this conceptual model of indebtedness across three separate experiments by combining a large-scale online questionnaire, behavioral measurements in an interpersonal game, computational modeling, and neuroimaging. First, we used an open-ended survey to capture lay intuitions about indebtedness based on regression analysis of past emotional experiences and topic modeling based-text

analysis of self-reported definitions. Overall, we find strong support that the feeling of indebtedness can be further separated into two distinct components - guilt from burdening the favor-doer and obligation to repay the favor. Using topic modeling on lay definitions of indebtedness, we find that guilt and gratitude appear to load on the same topic, while feeling words pertaining to burden and negative body states load on a separate topic. Second, we used a laboratory task designed to elicit indebtedness in the context of an interpersonal interaction and specifically manipulated information intended to shift the benefactor's perceptions of the beneficiary's intentions underlying their decisions. Although our manipulation was subtle, we find that it was able to successfully change participants' appraisals about how much the beneficiary cared about them and their beliefs about how much money the benefactor expected in return. Consistent with appraisal theory [28-33], these shifts in appraisals influenced participants' subjective feelings and ultimately their behaviors. Intentions perceived to be altruistic led to increased guilt and gratitude, while intentions viewed as more strategic increased feelings of obligation. While all three feelings increased reciprocity decisions, the guilt and obligation components of indebtedness increased the probability of rejecting help when that option was available to the participant.

One contribution of this work is the use of computational modeling to predict reciprocity and help-acceptance decisions based on our conceptual model of indebtedness. The majority of empirical research on indebtedness [21,46,47,76] and other emotions [77,78] has relied on participants' self-reported feelings in response to explicit questions regarding social emotions, which has significant limitations, such as its dependence on participants' ability to introspect [79,80]. Formalizing emotions using computational models is critical to advancing theory, characterizing their impact on behavior, and identifying associated neural and physiological substrates [39,81,82]. However, the application of computational modeling to the study of social emotions is a relatively new enterprise [39,54,83,84]. Previous research has had success modeling belief-

dependent utility using Psychological Game Theory [36,37] in interactive social contexts. Building on these works, we model participants' appraisals and emotions [28-33] based on the state of the game to predict two different types of decisions [39]. The current work contributes to a growing family of game theoretic models of social emotions such as guilt [34,54], gratitude [85], and anger [86,87], and can be used to infer feelings in the absence of self-report, providing new avenues for investigating other social emotions.

We provide a rigorous validation of our computational models using behaviors in the interpersonal game, self-reported subjective experiences, and neuroimaging. First, we can accurately predict participants' reciprocity and help-acceptance decisions. Second, we observed that the model predictions of second-order belief and perceived care in the reciprocity model accurately captured participant's trial-to-trial self-reported appraisal and feeling ratings. Third, our brain imaging analyses demonstrate that each feeling reflects a distinct psychological process, and that intention inference plays a key role during this process. Consistent with previous work on guilt [54,67,68,88] and gratitude [62-64], our model representation of communal concern correlated with increased activity in the insula, dlPFC, and default mode network including the vmPFC and precuneus. Obligation, in contrast, captured participants' second order beliefs about expectations of repayment and correlated with increased activation in regions routinely observed in mentalizing including the dmPFC and TPJ [66,69,70].

We provide an even stronger test of our ability to characterize the neural processes associated with indebtedness by deriving a "neural utility" model. Previous work has demonstrated that it is possible to build brain models of preferences that can predict behaviors [89,90] and the hidden motives behind the behaviors [91]. Here, we trained multivoxel patterns of brain activity to predict participants' communal and obligation utility. We then used these brain-derived representations of communal concern and obligation to predict how much money participants ultimately reciprocated to the

beneficiary. Remarkably, we found that this neural utility model of indebtedness was able to predict individual decisions entirely from brain activity and numerically outperformed (but not significantly) a control model that provided a theoretical upper bound of how well reciprocity behavior can be predicted directly from brain activity. Importantly, the neural utility model was able to accurately capture each participant's preference for communal concern relative to obligation. We observed a significant drop in our ability to predict behavior when we randomly shuffled the weighting parameter across participants. In addition, we find that the more the pattern of brain activity predicting reciprocity behavior resembled brain patterns predictive of communal concern or obligation, the more our behavioral computational model weighted this feeling in predicting behavior, demonstrating that these distinct appraisals/feelings are involved in motivating reciprocity decisions.

This work advances our theoretical understanding of social emotions. First, we highlight the complex relationship between gratitude and indebtedness. We propose that feeling cared for by a benefactor, which we call communal concern [44,45], is comprised of both guilt and gratitude. Each emotion diverges in valence, with gratitude being positive [6-9], and guilt being negative [40-42,44,54], but both promote reciprocity behavior. When faced with the offer of help, anticipated gratitude should motivate the beneficiary to accept help in order to establish or promote a relationship [6,7], whereas anticipated guilt should motivate the beneficiary to reject help out of concern to protect the benefactor from incurring a cost [44,54,92]. Although we observed support for this prediction, our interpersonal task was not designed to explicitly differentiate guilt from gratitude, which limited the ability of our reciprocity model to capture the specific contributions of guilt and gratitude to communal concern and likely impacted identifiability of the parameters of the help-acceptance model. Future work might continue to refine the relationship between these two aspects of communal concern both in terms of behaviors in experiments and computations in models [54,62-64,67,68,88].

Second, our conceptual model provides a framework to better understand the role of relationships and contexts in generating feelings of indebtedness within a single individual. Different types of relationships (see Clark and Mills's theory of communal and exchange relationships [4,5], and Alan Fiske's Relational Models Theory [93]) have been theorized to emphasize different goals and social norms which can impact social emotions [94,95]. For example, communal relationships prioritize the greater good of the community and are more conducive to altruistic sharing, which can be signaled by altruistic favors [3-5]. In contrast, exchange relationships are more transactional in nature [2,4,5,10-12] and emphasize maintaining equity in the relationship, which can be signaled by strategic favors [93]. Our conceptual model proposes that perceptions of the benefactor's intentions directly impact the feelings experienced by the beneficiary (e.g., guilt & obligation). Although we deliberately attempted to minimize aspects of the relationship between the benefactor and beneficiary by making players anonymous to control for reputational effects, future work might experimentally manipulate these relationships to directly test the hypothesis that relationship types differentially moderate the responses of gratitude and subcomponents of indebtedness.

Third, we present new evidence exploring the relationship between indebtedness and guilt. Guilt and indebtedness are interesting emotions in that they are both negatively valenced, yet promote prosocial behaviors. In previous work, we have operationalized guilt as arising from disappointing a relationship partner's expectations [39,54,55,96], which is conceptually related to the feeling of obligation in this paper. This feeling results from disappointing a relationship partner or violating a norm of reciprocity and is a motivational sentiment evoked by social expectations reflecting a "sense of should" that is associated with other negative affective responses such as feelings of pressure, burden, anxiety, and even resentment [49-51]. In other work, we have investigated how guilt can arise from causing unintended harm to a relationship partner [68,97]. This is

conceptually more similar to how we frame guilt here, which arises from the feeling that one has unnecessarily burdened a relationship partner even though the help was never explicitly requested by the benefactor. We believe that continuing efforts to refine mathematical models of emotions across a range of contexts, will eventually allow the field to move beyond relying on the restrictive and imprecise semantics of linguistic labels to define emotion categories (e.g., guilt, gratitude, indebtedness, obligation, feeling, motivation, etc.).

Our study has several potential limitations, which are important to acknowledge. First, although we directly and conceptually replicate our key findings across multiple samples, all of our experiments recruit experimental samples from a Chinese population. It is possible that there are cultural differences in the experience of indebtedness, which may not generalize to other parts of the world. For example, compared with Westerners who commonly express gratitude when receiving benevolent help, Japanese participants (East Asian population) often respond with "Thank you" or "I am sorry", indicating their higher experience of guilt after receiving favors [40,41]. Cultural differences may perhaps reflect how the two components of indebtedness are weighted, with guilt being potentially more prominent in East Asian compared to Western populations, reflecting broader cultural differences in collectivism and individualism. Second, our computational models may oversimplify the appraisal and emotion generating processes. These models operationalize the appraisals of perceived care and second-order belief using information available to each participant in the task (i.e., benefactor's helping behavior and manipulation about the participants' ability to reciprocate), which may not generalize to other experimental contexts without modification. Although our computational models performed well in capturing participants' decision behaviors in this task, we emphasize the importance of continued refinement. Third, future research is needed to extend our conceptual model by differentiating different types of help-receiving events (e.g., help when moving to a

new apartment vs. help during a period of sickness) and manipulating other related contexts, such as gift-receiving [23] and help-seeking [17].

In summary, in this study we develop a comprehensive and systematic conceptual model of indebtedness and validate it across three studies combining a large-scale online questionnaire, an interpersonal game, computational modeling and neuroimaging. A key aspect to this work is the emphasis on the role of appraisals about the intentions behind a favor in generating distinct feelings of guilt and obligation, which in turn motivates how willing beneficiaries are to accept or reject help and ultimately reciprocate the favor. Together these findings highlight the psychological, computational, and neural mechanisms underlying the hidden costs of receiving favors [22-24].

Methods

Study 1 - Online Questionnaire

Participants. Participants (1,808 graduate and undergraduate students) were recruited from Zhengzhou University, China to complete an online questionnaire. None of the participants reported any history of psychiatric, neurological, or cognitive disorders. Participants were excluded if they filled in information irrelevant (e.g., this question is boring, or I don't want to answer this question) to the question or experiment in the essay question (189 participants), leaving 1,619 participants (self-reported gender: 812 females, 18.9 ± 2.0 (SD) years). While 98.7% participants reported the events of receiving help, 24.4% participants reported the events of rejecting help within the past one year, resulting in 1,991 effective daily events. To extract the words related to emotions and feelings in the definition of indebtedness, 80 additional graduate and undergraduate students (45 females, 22.6 ± 2.58 years) were recruited from different universities in Beijing to complete the word classification task. This experiment was carried out in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the School of Psychological and Cognitive Sciences, Peking University. Informed written consent was obtained from each participant prior to participating.

Experimental Procedures. Participants reported their responses on the Questionnaire Star platform (https://www.wjx.cn/) using their mobile phones. The questionnaire consisted of two parts (see Appendix S1 for full questionnaire). Each participant was asked to recall a daily event in which they **received help** (part 1) or **rejected help** (part 2) from others, and to answer the questions regarding their appraisals, emotions, and details of this event. Events were required to be clearly recalled and to have occurred within the past year. Appraisal questions included: "To what extent do you think the benefactor expected you to repay? (i.e., second-order belief; Q. 15)", and "To what extent do you think the benefactor cared about your welfare when helping you? (i.e.,

perceived care; Q. 21)". Emotion ratings included: gratitude (Q. 9), indebtedness (Q. 10), guilt (Q. 11), and obligation (Q. 12). The questions for guilt [40-43] and obligation [13,14,21,46,47] were designed according to the operational definitions used by previous research. For events in which participants accepted help (Part 1), questions for behaviors included: "To what extent did you think you needed to reciprocate? (Q. 14)" The questions for other related factors (control variables) included: " How was your relationship with this benefactor before receiving help? (i.e., relationship with the benefactor; Q. 3)", "6. How helpful was the help? (i.e., the participant's benefit; Q. 6)", and "7. How much was the benefactor's cost in this help? (i.e., the benefactor's cost; Q. 7)". Questions were the same for Part 2 (i.e., events in which participants rejected help), except that participants were asked to *imagine* how they would feel or behave if they accepted this help.

To explore how participants defined indebtedness, participants answered the following two questions about the definition of indebtedness after recalling the event: (1) In the context of helping and receiving help, what is your definition of indebtedness? (Fill-in-the-blanks test) (2) In daily life, what do you think is/are the source(s) of indebtedness? Multiple-choice question with four options "Negative feeling for harming the benefactor", "Negative feeling for the pressure to repay caused by other's ulterior intentions", "Both" and "Neither" (see details in Appendices S1).

Study 2 - Interpersonal Task

Participants. *For Study 2a (behavioral study)*, 58 graduate and undergraduate Chinese Han students were recruited from Zhengzhou University, China, and 7 participants were excluded due to equipment malfunction, leaving 51 participants (self-reported gender: 33 females, 19.9 ± 1.6 years) for data analysis. *For Study 2b (behavioral study)*, 60 graduate and undergraduate Chinese Han students were recruited from Zhengzhou University, China, and 3 participants were excluded due to failing to respond in more

than 10 trials, leaving 57 participants (45 females, 20.1 ± 1.8 years) for data analyses. None of the participants reported any history of psychiatric, neurological, or cognitive disorders. This experiment was carried out in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the School of Psychological and Cognitive Sciences, Peking University. Informed written consent was obtained from each participant prior to participating.

Experimental Procedure. In Study 2a and Study 2b, seven participants came to the experiment room together. An intra-epidermal needle electrode was attached to the left wrist of each participant for cutaneous electrical stimulation [98]. The first pain stimulation was set as 8 repeated pulses, each of which was 0.2 mA and lasted for 0.5 ms. A 10-ms interval was inserted between pulses. Then we gradually increased the intensity of each single pulse until the participant reported 6 on an 8-point pain scale (1 = not painful, 8 = intolerable). Participants reported that they would only experience the whole pulse train as a single stimulation, rather than as separate shocks. The final intensity of pain stimulation was calibrated to a subjective pain rating of "6", which was a moderate punishment for the participants.

Both Study 2a and Study 2b consisted of two sessions. All stimuli were presented using PsychToolBox 3.0.14 (www.psychtoolbox.org) in Matlab 2016a (Mathworks, Natick, MA, USA). Participants were instructed as following:

"In this experiment, you will play an interpersonal game, which is composed of two roles: the Decider and the Receiver. The Receiver will be in some trouble and the Decider can decide whether to help the Receiver at the cost of his/her own interests. Several previous participants have come to our lab during Stage 1 of our study and made decisions as the Deciders. Now this experiment belongs to Stage 2 of this study. In the two sessions of the experiment, you will perform as the Receiver, facing the

decisions made by each anonymous previous Decider in Stage 1 and make your own decisions."

During Session 1 (the main task), each participant played multiple single-shot rounds of the interpersonal game (Fig. 3) as a Receiver. In each trial, the participant was paired with an anonymous same-gender Decider (co-player), and was informed that the co-player in each trial was distinct from the ones in any other trials and only interacted with the participant once during the experiment. In each round, the participant had to receive a 20-second pain stimulation with the intensity of 6. The participant was instructed that each co-player: (a) had come to the lab before the participant, (b) had been endowed with 20 yuan (~ $3.1 USD), and (c) had decided whether and how much to spend from this endowment to help the participant reduce the duration of pain (i.e., ***benefactor's cost***, D_A). The more the benefactor spent, the shorter the duration of the participant's pain experience. The maximum pain reduction was 16 seconds to ensure that participants had some amount of pain on each trial. Unbeknownst to the participant, the benefactor's costs were uniformly sampled from the available choices from an unpublished pilot study on helping behaviors (Table S2).

Each trial began by informing the participant which Decider from Stage 1 was randomly selected as the co-player for the current trial (Information period, 4 sec), with a blurred picture and the participant ID of the co-player, and extra information regarding the co-player's intention to help (see below). The co-player's decision on how much they chose to spend to help the participant was presented (Outcome period, 5 sec). Next, the participant indicated how much he/she thought this co-player expected him/her to reciprocate (i.e., Second-order belief of the co-player's expectation for repayment; continuous rating scale from 0 to 25 using mouse, step of 0.1 yuan, < 8 sec). In half of the trials, the participant had to passively accept the co-player's help (force-accept situation). The sentence "You have to accept this help" was presented on the

screen (4 sec). In the other half, the participant could decide whether or not to accept the co-player's help (free-choice situation, < 4 sec). The order of options was counterbalanced across trials. If the participant accepted the help, the co-player's cost and the participant's pain reduction in this trial would be realized according to the co-player's decision; if the participant did not accept the help, the co-player would spend no money and the duration of participant's pain stimulation would be the full 20 seconds. At the end of each trial, the participant was endowed with 25 yuan (~ 3.8 USD) and decided how much they wanted to allocate to the co-player as reciprocity in this trial from this endowment (D_B, continuous choice from 0 to 25 using mouse, step of 0.1 yuan, < 8 sec). We focused on two types of behaviors in this help-receiving context: (1) the participant's amounts of allocation when they passively accept the co-player's help (i.e., reciprocity decisions), and (2) the participant's decisions of whether to accept or reject help in free-choice trials (i.e., help-acceptance decisions).

We manipulated the perceived intention of the co-player (i.e., the benefactor) by providing participants with extra information regarding the co-player's expectation of reciprocity (i.e., **extra information about benefactor's intention**) below the co-player's subject id at the beginning of each trial. Each participant was instructed that before making decisions, some co-players were informed that the participant would be endowed with 25 yuan and could decide whether to allocate some endowments to them as reciprocity (i.e., ***Repayment possible condition***). The other co-players were informed that the participant had no chance to reciprocate after receiving help (i.e., ***Repayment impossible condition***). In fact, participants could reciprocate in both conditions during the task. The endowment of the co-player (γ_A) was always 20 yuan, and the endowment of the participant (γ_B) in each trial was always 25 yuan. The endowment of the participant was always larger than the endowment of the co-player to make the participant believe that the co-player expected repayments in Repayment possible condition.

In Study 2a, we manipulated the participant's beliefs about the benefactor's intentions by providing additional information about benefactor's intention (condition: Repayment possible vs. Repayment impossible) and benefactor's cost (12 levels of 5, 7, 8, 9, 10, 11, 12, 14, 15, 16, 18, 20). We included one trial for each condition-benefactor's cost combination for free-choice and force-accept situations, totaling 24 trials in each situation. As a result, there were a total of 48 trials with different anonymous co-players in Study 2a.

Study 2b also included information about benefactor's intention and benefactor's cost. In addition, to disentangle the effect of the benefactor's cost and participant's benefit (i.e., pain reduction), we manipulated the exchange rate between the co-player's cost and participant's pain reduction (i.e., efficiency, 0.5, 1.0, and 1.5; the efficiency was always 1.0 in Study 2a). Thus, the participant's pain reduction was calculated as follows: Pain reduction = co-player's cost / co-player's endowment × efficiency × maximum pain reduction (16s). In Study 2b, we only included 5 levels for the benefactor's cost (i.e., 4, 8, 12, 16, 20). Participants were informed that the co-player could only choose discrete amounts of money to spend in this experiment. Furthermore, because the pain duration when benefactor's cost = 20 and efficiency = 1.5 exceeds the maximum pain reduction, this combination was eliminated, leaving 14 benefactor's cost-efficiency combinations. We included one trial for each condition-benefactor's cost-efficiency combination for free-choice and force-accept situations, totaling 28 trials in each situation. As a result, there were 56 trials with different anonymous co-players in Study 2b. See Table S2 for more details about experimental settings.

Post-task ratings. During Session 2 of the interpersonal game, all of the trials from session 1 were displayed again in a random order. The participant was shown the co-player's information (Information period with a blurred picture, co-player's ID and

extra information about benefactor's intention, i.e., the benefactor knew *Repayment possible* or *Repayment impossible*; 4 sec) and their decision (Outcome period with benefactor's cost and the duration of pain reduction for the participant; 5 sec). Then the participant was asked to recall how much they believed the benefactor cared about them, as well as their feelings of indebtedness, obligation, guilt, and gratitude at the time point when they had received the help of the co-player, but had not indicated decisions of accept/reject help and the amount of reciprocity. Ratings were conducted on a scale from 0 to 100, with 0 represented "not at all" and 100 represented "extremely intense". The order of these ratings was counter-balanced across trials. The questions for self-reported ratings on guilt [40-43] and obligation [13,14,21,46,47] were based on previous research. Notably, the participant's second-order belief of how much the benefactor expected, decisions of whether to accept or reject help, and the amount of reciprocity were not shown in this session to minimize the influence of their prior behaviors on their reported feelings.

- "How much gratitude do you feel for this co-player's decision?" (Gratitude)
- "How much indebtedness do you feel for this co-player's decision?" (Indebtedness)
- "How much do you think this decider cares about you?" (Perceived care)
- "How much pressure did you feel for the decider's expectation for repayment?" (Obligation)
- "How much guilt do you feel for this co-player's decision?"(Guilt)

At the end of the experiment, five trials in Session 1 were randomly selected to be realized. The participant received the average pain stimulation in these five trials. The participant's final payoff was the average amount of endowment the participant left for him/herself across the chosen trials. The participant was instructed that the final payoff of each co-player was the amount of endowment the co-player left plus the amount of endowment the participant allocated to him/her. Participants were informed of this

arrangement before the experiment began. After the experiment, participants were further debriefed that the co-players' decisions they were faced with during the experiment were actually pre-selected from participants' decisions in a previous experiment by the experimenters, and the co-players' decisions did not necessarily reflect the natural distributions of others' helping behaviors.

Study 3 - FMRI Study

Participants. For Study 3, 57 right-handed healthy graduate and undergraduate Chinese Han students from Beijing, China took part in the fMRI scanning. Four participants with excessive head movements (>2mm) were excluded, leaving 53 participants (self-reported gender: 29 females, 20.9 ± 2.3 years) for data analysis. None of the participants reported any history of psychiatric, neurological, or cognitive disorders. This experiment was carried out in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the School of Psychological and Cognitive Sciences, Peking University. Informed written consent was obtained from each participant prior to participating.

Experimental Procedure. Each participant came to the scanning room individually. The pain-rating procedure and the two sessions of the task in the fMRI study were identical to Study 2a, except that participants always had to accept their co-player's help. Session 1 (the main task; Fig. 3B) was conducted in the fMRI scanner, while Sessions 2 was conducted after participants exited the scanner. The scanning session consisted of three runs (in total 54 trials) and lasted for approximately 39 min. Each run lasted for 13 min and consisted of 18 trials, including 9 levels of the benefactor's cost (4, 6, 8, 10, 12, 14, 16, 18, and 20) in Repayment possible condition and Repayment impossible conditions respectively. Trial order was pseudorandomized. See Table S2 for additional details about the experimental design.

Each trial began with a 4 sec Information period, which showed the randomly selected co-player's subject id, blurred picture, and information of whether this co-player knew that the participant could or could not repay. This was followed by the 5 sec Outcome period, which included the co-player's decision on how much they spent to help the participant. Participants then had up to 8 sec to report how much he/she thought this co-player expected him/her to reciprocate (i.e., second-order belief of the co-player's expectation for repayment; rating scale from 0 to 25 using left and right buttons to move the cursor, step of 1 yuan). Next, participants had 8 sec to decide how much of their 25 yuan endowment (~ 3.8 USD) to reciprocate to the co-player (from 0 to 25 using left and right buttons to move the cursor, step of 1 yuan). Before and after each period, a fixation cross was presented for a variable interval ranging from 2 to 6 s, which was for the purpose of fMRI signal deconvolution.

Data Analyses in Study 1 (Online Questionnaire)

Validating Conceptual Model with Emotion Ratings. We first attempted to validate the conceptual model using the emotional ratings for daily-life events of receiving and rejecting help obtained from online-questionnaire in Study 1. We conducted between-participant linear regressions predicting indebtedness from guilt and obligation ratings. All variables were normalized before regression analysis. We additionally examined the degree of multicollinearity between guilt and obligation ratings using the variance inflation factor (VIF). The VIF reflects the degree that any regressor can be predicted by a linear combination of the other regressors (VIF = 5 serves as informal cutoff for multicollinearity – lower numbers indicate less collinearity). Results demonstrated an acceptable level of multicollinearity between guilt and obligation ratings (Table S1). To rule out the possibility that these emotion ratings might covary with other related factors in Study 1 (e.g., benefactor's cost, the participant's benefit and the social distance between the participant and the benefactor), we re-estimated the above models with

these factors as control variables, which did not appreciably change the results (Table S1).

Validating Conceptual Model with Self-Reported Appraisals. Next, we summarized participants' self-reported sources of their feelings of indebtedness. We calculated the frequency that participants selected each of the four options in the question "In daily life, what do you think is/ are the source(s) of indebtedness?" in Study 1 (Fig, S1A), as well as how often that participants attributed "Negative feeling for harming the benefactor" and "Negative feeling for the pressure to repay caused by other's ulterior intentions" as the sources of indebtedness (i.e., the frequency of choosing each single option plus the frequency of choosing "Both of the above").

Validating Conceptual Model with Topic Modeling. We also attempted to validate the conceptual model by applying topic modeling to participant's open-ended responses describing their own definition of indebtedness in Study 1. We used the "Jieba" (https://github.com/fxsjy/jieba) package to process the text and excluded Chinese stopwords using the stopwords-json dataset (https://github.com/6/stopwords-json). Because Chinese retains its own characters of various structures, we also combined synonyms of the same word as an additional preprocessing step [99]. Next, we computed a bag of words for each participant, which entailed counting the frequency that each participant used each word and transformed these frequencies using Term Frequency-Inverse Document Frequency (TF-IDF) [100,101]. This method calculates the importance of a word in the whole corpus based on the frequency of its occurrence in the text and the frequency of its occurrence in the whole corpus. The advantage of this method is that it can filter out some common but irrelevant words, while retaining important words that affect the whole text. Using this method, the 100 words with the highest weight/frequency in the definitions of indebtedness were extracted (Appendices S2). Words beyond these 100 had TF-IDF weights < 0.01 (Fig. S1B), indicating that the

words included in the current analysis explained vast majority of variance in the definition of indebtedness. These 100 words were then classified by an independent sample of participants (N = 80) into levels of appraisal, emotion, behavior, person and other (see *SI Methods*). We conducted Latent Dirichlet Allocation (LDA) based topic modeling on the emotional words of indebtedness using collapsed Gibbs sampling implemented in "lda" package (https://lda.readthedocs.io/en/latest/) [102]. LDA is a generative probabilistic model for collections of discrete data such as text corpora, which is widely used to discover the topics that are present in a corpus [53]. It finds latent factors of semantic concepts based on the co-occurrence of words in participant's verbal descriptions without constraining participants' responses using rating scales, which currently dominates emotion research [103]. We selected the best number of topics by comparing the models with topic numbers ranging from 2 to 15 using 5-fold cross validation. Model goodness of fit was assessed using perplexity [104], which is a commonly used measurement in information theory to evaluate how well a statistical model describes a dataset, with lower perplexity denoting a better probabilistic model. We found that the two-topic solution performed the best (Fig. S1C).

Validating Conceptual Model with Self-Reported Behaviors. We next sought to test the predictions of the conceptual model using the self-reported behaviors from Study 1. First, we used data from Part 1 of the questionnaire and used linear regression to predict self-reported need to reciprocate from self-reported feelings of indebtedness, guilt, obligation and gratitude, separately. Second, we combined the data of the events associated with receiving (Part 1) and rejecting help (Part 2) and used logistic regression to classify reject from accept behavior using self-reported counterfactual ratings of indebtedness, guilt, and obligation, and gratitude.

Data analyses in Study 2 (Interpersonal Task)

Validating Conceptual Model with Emotion Ratings. Similar to Study 1, we tested whether guilt and obligation contribute to indebtedness using the trial-by-trial emotional ratings in Study 2. We fit mixed effects regressions using lme4 predicting indebtedness ratings from guilt and obligation ratings with random intercepts and slopes for participants and experiments (e.g., 2a, 2b). All variables were normalized before regression analysis. Hypothesis tests were conducted using the lmerTest package [105] in R. We additionally examined the degree of multicollinearity between guilt and obligation ratings using VIF (Table S1). To rule out the possibility that these emotion ratings might covary with other related factors, i.e., the experimental variables in Studies 2 (e.g., benefactor's cost, extra information about the benefactor's intention and efficiency), we fit additional models controlling for these factors. Results of Study 2 replicated those in Study 1, and did not change after controlling for these variables (Table S1).

The Effects of Experimental Manipulations on Participants' Appraisal, Emotional and Behavioral Responses. To test the effects of experimental manipulations on beneficiary's appraisals (i.e., second-order belief and perceived care), emotions (i.e., gratitude, indebtedness, guilt, and obligation) and behaviors (reciprocity and help-acceptance decisions), in Study 2a we conducted LMM analyses for each dependent variable separately with the benefactor's cost, extra information about benefactor's intention (Repayment possible vs. Repayment impossible) and their interaction effect as fixed effects and the participant ID as a random intercept and slope [106] (Table S3). All variables were normalized before regression analysis. Visualizations of the regression analyses (Fig. 4, B - C) were created using the lmplot function of seaborn 0.9.0 (https://seaborn.pydata.org/index.html) in IPython/Jupyter Notebook (Python 3.6.8) [107] and using the data of all trials for all participants. We note that the confidence

intervals were created via bootstrapping that respected the repeated measurements within a participant.

Relationships between Appraisals and Emotions. To reveal the relationships between appraisals (i.e., second-order belief and perceived care) and emotions (i.e., indebtedness, guilt, obligation, and gratitude), we estimated the correlations between these variables at both within-participant and between-participant levels. For within-participant analysis, for each pair of these six variables, we estimated the pearson correlation for each participant, transformed the data using a fisher *r* to *z* transformation, and then conducted a one-sample test using *z* values of all participants to evaluate whether the two variables were significantly correlated at the group level. This analysis captured the variability of appraisals and emotions across trials within participants (Fig. S3A). For between-participant analysis, for each of the six variables, we computed the average value of the variable across all trials for each participant. We then estimated the correlations between each pair of variables based on variability across participants (Fig. S3B).

Given the strong correlations between appraisals and emotions (Fig. S3, A-B, Tables S5 and S6), we conducted a factor analysis to examine the relationship between appraisals and emotions [108]. The Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy [109] and Bartlett's test of sphericity [110] showed that the current data sets in Studies 2 and 3 were adequately sampled and met the criteria for factor analysis (Study 2: KMO value = 0.76, Bartlett's test $\chi^2 = 8801.85$, $df = 15$, $p < 0.001$; Study 3: KMO value = 0.77, Bartlett's test $\chi^2 = 2970.53$, $df = 15$, $p < 0.001$). All the variables were centered within participant to exclude the influences of individual differences in the range of ratings. We first applied exploratory factor analysis (EFA) in Study 2 to identify the number of common factors and the relationships between appraisals and emotions. To determine the number of components to retain, the correlation matrix

between the 6 variables was submitted to a parallel analysis using the "psych" package [111] for R. Parallel analysis performs a principal factor decomposition of the data matrix and compares it to a principal factor decomposition of a randomized data matrix. This analysis yields components whose eigenvalues (magnitudes) are greater in the observed data relative to the randomized data. The nScree function was used to determine the number of factors to retain. The result pointed to a two-factor solution (Fig. S2E). Factors were then estimated and extracted by combining ML factor analysis with oblique rotation using the "GPArotation " package for R [108]. Next, we conducted confirmatory factor analysis (CFA) using the data of Study 3 to test the two-factor model built by Study 2 in an independent sample. CFA was conducted using "lavaan" package [112] for R. Results remained the same after controlling for the experimental variables.

To test whether the two appraisals mediated the observed effects of experimental variables on emotional responses, we conducted a multivariate mediation analysis using structural equation modeling using the 'lavaan' package in R [112]. In this analysis, experimental variables (extra information about benefactor's intention, benefactor's cost, information-cost interaction, and efficiency) were taken as independent variables, ratings of second-order belief and perceived care were taken as mediators, and ratings of guilt, gratitude and the sense of obligation were taken as dependent variables. First, we built a full model that included all pathways between variables. Then, non-significant pathways in the full model were excluded from the full model to improve the fitness of the model. In the final model, experimental variables included extra information about the benefactor's intentions, the benefactor's cost, and their interaction; efficiency was excluded due to the non-significant effects. Moreover, in the final model, second-order beliefs mediated the effects of the experimental variables on obligation, whereas perceived care mediated the effects of experimental variables on guilt and gratitude. This model performed well (RSMEA = 0.023, SRMR = 0.004, CFI

= 1.000, TLI = 0.997, BIC = 27496.99) and explained participants' responses better than the full model (RSMEA = 0.046, SRMR = 0.004, CFI = 1.000, TLI = 0.986, BIC = 27543.52).

Using Communal and Obligation Factors as Predictors for Behaviors. To investigate how participants' appraisals and emotions influenced their behavioral responses, we conducted two separate LMMs to predict participants' reciprocity decisions and help-acceptance decisions, respectively. Each model included the scores for Communal and Obligation Factors estimated from the factor analysis (Fig. 4D) as fixed effects and random intercepts and slopes for participants. All variables were normalized before regression analysis. See detailed results in the *SI Results*.

Computational Modeling. We built separate computational models predicting participant's reciprocity decisions and help-acceptance decisions based on the conceptual model of indebtedness (see Tables S14 for all model object definitions). The utility of each behavior $U(D_B)$ was modeled based on the competing latent motivations of self-interest, communal concern (guilt and gratitude), and obligation using Eq. 1.

For reciprocity decisions, self-interest π_B was defined as the percentage of money kept by the participant out of their endowment γ_B. For help-acceptance decisions, self-interest π_B for accepting help was defined as the percentage of pain reduction from the maximum amount possible, which depended on how much the benefactor spent to help D_A and the exchange rate between the benefactor's cost and the participant's benefit μ (i.e., help efficiency).

$$\pi_B = \begin{cases} \frac{\gamma_B - D_B}{\gamma_B} & \text{Reciprocity model} \\ \frac{D_A * \mu}{\max(D_A * \mu)} & \text{Help-acceptance model} \end{cases} \quad \textbf{Eq. 7}$$

As stated in the *Results* section, we separately modeled the appraisals of second-order beliefs E_B'' of the benefactor's expectation for repayment (Eq. 3) and perceived care ω_B (Eq. 5) as intermediate representations of the feelings of obligation and communal concern (guilt and gratitude), respectively (Eq. 2 and Eq. 4).

For each trial, we modeled the participant's second-order belief E_B'' of how much they believed the benefactor expected them to reciprocate based on the amount of help offered by the benefactor D_A and whether the benefactor knew repayment was possible (Eq. 3). In the Repayment impossible condition, participants knew that the benefactor did not expect them to reciprocate, so we set E_B'' to zero. However, in the Repayment possible condition, the benefactor knew that the participant had money that they could spend to repay the favor. In this condition, we modeled the E_B'' as proportional to the amount of money the benefactor spent to help the participant.

The appraisals of perceived care ω_B (Eq. 5) were defined as a function of the benefactor's cost D_A and second-order belief E_B''. Specifically, we assumed that the perceived care from help increased as a linear function of how much the benefactor spent D_A from his/her endowment γ_A. In other words, the more the benefactor spent, the more care the participant would perceive from the help. However, we assume that this effect is mitigated by the second-order belief of the benefactor's expectation for repayment E_B''. That is, when faced with a specific amount of benefactor's cost, if the participant thought this benefactor expected more repayment, the less care the participant would perceive from the help. Here, the parameter κ ranges from [0, 1] and represents the degree to which the perceived strategic intention E_B'' reduces the perceived altruistic intention ω_B. This creates a nonlinear relationship between ω_B and E_B'' such that the relationship is negative when κ is close to one, positive when κ is close to zero, and uncorrelated in the current dataset with $\kappa = 0.32 \pm 0.01$, $\beta = -0.03 \pm 0.03$, $t = -1.23$, $p = 0.222$ (Fig. S4).

Furthermore, our conceptual model proposed that the feelings of obligation and communal concern (guilt and gratitude) stem from the appraisals of benefactor's strategic intention (i.e., second-order belief) and altruistic intention (i.e., perceived care from the help), respectively. This hypothesis was supported by the results of the mediation analysis (Fig. 4E). Thus, we modeled the utilities of obligation and communal concern (i.e., $U_{Obligation}$ and $U_{Communal}$) as the functions of E_B'' and ω_B, respectively (Eq. 2 and Eq. 4). Though the two decision models are conceptually similar, the values of $U_{Communal}$ and $U_{Obligation}$ are computed slightly differently due to differences in the types of data (i.e., continuous vs binary decisions) and how appraisals are inferred (e.g., there is no money to return D_B in help acceptance decisions).

Predicting reciprocity decisions

To predict continuous reciprocity decisions, we assume that participants were motivated to meet the expectation of the benefactor due to the sense of obligation, and thus maximized $U_{Obligation}$ by minimizing the difference between the amount they reciprocated D_B and their second-order belief of how much they believed the benefactor expected them to return E_B'', scaled by the participant's endowment size γ_B (Eq. 2). We note that our mathematical operationalization of obligation here is more akin to how we have previously modeled guilt from disappointing others in previous work [34,39,54,55] (see also *Discussion*). We also assumed that participants were motivated to reciprocate in response to the benefactor's perceived care due to guilt and gratitude (i.e., communal concern), and thus maximized $U_{Communal}$ by minimizing the difference between the benefactor's reciprocity D_B and their perception of how much they believed the benefactor cared about them ω_B, scaled by the participant's endowment size γ_B (Eq. 4). These non-linear formulations are a standard way to model social preferences in utility functions [55,113] by adding convexity to the utility function via exponential error signals resulting from failing to meet perceived social standards (the benefactor's expectation

or the benefactor's perceived care). In terms of model performance, this non-linear model for reciprocity outperformed a model with linear formulations of utilities for self-interest, communal concern, and obligation (Model 1.2; Tables S7 and S9; see details in *Methods* below and *SI Methods*).

Based on our conceptual model (Fig. 1), we defined $U_{Communal}$ as a mixture of feelings of gratitude $U_{Gratitude}$ and guilt U_{Guilt}, in which the parameter δ_B ranged from [0,1] and reflected how much gratitude contributed to communal concern in comparison to guilt.

$$U_{Communal} = \delta_B * U_{Gratitude} + (1 - \delta_B) * U_{Guilt} \quad \textbf{Eq. 8}$$

We note that both guilt and gratitude positively contribute to reciprocity and our interpersonal task was not designed to explicitly differentiate the effects of guilt and gratitude, which precluded our ability to estimate the specific value of δ_B for predicting reciprocity decisions due to a lack of identifiability. Thus, in this paper we can only make inferences about the broader $U_{Communal}$, which may reflect guilt and/or gratitude. However, our help-acceptance model does attempt to differentiate the contributions of guilt and gratitude to decisions of whether or not to accept help as discussed below.

We modeled the utility U associated with the participants' reciprocity decisions D_B after receiving help in Eq. 1, where ϕ is a free parameter constrained between [0, 1] that captures the trade-off between feelings of communal concern and obligation. The reciprocity model (Model 1.1) selects the participant's decision D_B associated with the highest utility.

$$U(D_B) = \theta_B * \frac{\gamma_B - D_B}{\gamma_B} - (1 - \theta_B) * (\phi_B * (\frac{\omega_B * \gamma_B - D_B}{\gamma_B})^2 + (1 - \phi_B) * (\frac{E''_B - D_B}{\gamma_B})^2)$$

Eq. 9 (Model 1.1)

We estimated the model parameters for Model 1.1 by minimizing the sum of squared errors of the percentages that the model's behavioral predictions deviated from actual behaviors across all the trials that participants had to passively accept help [55]. Specifically, for each parameterization, we select the maximum $U(D_B)$ across the range of D_B (i.e., $max(U(D_B))$) to predict each participant's reciprocity decision for each trial. We calculate the difference between the participant's actual choice and the model predicted choice, yielding SSE, or the residual sum of squares. We used Matlab's fmincon routine to identify parameters that minimized the sum of SSE of all trials separately for each participant (Eq. 10).

$$SSE = \sum_{t=1}^{n} \left(\frac{D_B(t) - max(U(D_B(t)))}{\gamma_B} * 100 \right)^2 \qquad \textbf{Eq. 10}$$

with t indicating trial number. To avoid ending the fitting procedure at a local minimum, the model-fitting algorithm was initialized at 1000 random points in the three-dimensional theta-phi-kappa parameter space for each participant.

Predicting help-acceptance decisions

We created a separate model (Model 2.1) to predict help-acceptance decisions. $U_{Obligation}$ was defined as a linear function of E_B'' (Eq. 2). $U_{Communal}$ was defined as a linear function of ω_B (Eq. 4). We modeled the utility of accepting and rejecting help as:

$$\begin{cases} U(Accept) = \theta_B * \pi_B + (1 - \theta_B) * (\phi_B * U_{Communal} - (1 - |\phi_B|) * U_{Obligation}) \\ \qquad\qquad = \theta_B * \dfrac{D_A * \mu}{\max(D_A * \mu)} + (1 - \theta_B) * (\phi_B * \omega_B - (1 - |\phi_B|) * \dfrac{E_B''}{\gamma_B}) \\ U(Reject) = 0 \end{cases}$$

Eq. 11 (Model 2.1)

In this model, *U(Reject)* was set to zero, because the participant's emotional responses would not change if the participant did not accept help. Increased obligation reduces the likelihood of accepting help to avoid being in the benefactor's debt [13,14,114]. In contrast, $U_{Communal}$ has a more nuanced influence on behavior, with guilt decreasing the likelihood of accepting help to avoid burdening a benefactor [34,54], and gratitude motivating accepting help to build a communal relationship [6,7]. However, because $U_{Communal} = U_{Guilt} = U_{Gratitude} = \omega_B$ in this formulation, there is no variability in the design for the model to be able to disentangle the effect of gratitude from that of guilt. To address this complexity, we constrain ϕ to be within the interval of [-1, 1], and explicitly divide up the parameter space such that $\phi > 0$ indicates a preference for gratitude and motives the participants to accept the help, while $\phi < 0$ indicates a preference for guilt and motives the participants to reject the help.

$$\begin{cases} \phi_B > 0 & Gratitude \\ \phi_B < 0 & Guilt \end{cases} \qquad \textbf{Eq. 12}$$

Regardless of whether the participant is motivated primarily by guilt or gratitude, participants can still have a mixture of obligation captured by $1 - |\phi|$, which ranges from [0,1]. Unfortunately, if participants are equally sensitive to gratitude and guilt, ϕ will reduce to zero and the weight on obligation increases, which decreases the model fit and leads to some instability in the parameters.

We computed the probability of the decision of whether to accept or reject help using a softmax specification with inverse temperature parameter λ, which ranges from [0,1]. In each trial, the probability of the participant choosing to accept help is given by

$$P(Accept) = \frac{e^{U(Accept)/\lambda}}{e^{U(Accept)/\lambda} + e^{U(Reject)/\lambda}} \qquad \textbf{Eq. 13}$$

We then conducted maximum likelihood estimation at the individual level by minimizing the negative log likelihood of the decision that the participant made (D_B = Accept or Reject) over each trial t with 1000 different starting values across the four dimensional theta-phi-kappa-lambda parameter space using the fmincon routine implemented in Matlab:

$$LLE = -\sum_{t=1}^{n} log(P(D_B(t))) \qquad \textbf{Eq. 14}$$

Validations of Computational Modeling.

Model comparison for reciprocity decisions

We compared Model 1.1 with other plausible models, including: (a) a model with linear formulations of utilities for self-interest, communal concern and obligation (Model 1.2), (b) models that solely included the term for communal concern (Model 1.3) or obligation (Model 1.4) besides the self-interest term, (c) models with separate parameters for self-interest, communal concern and obligation with separate parameters (Model 1.5 and Model 1.6), (d) a model that assumes participants reciprocate purely based on the benefactors helping behavior (i.e., tit-for-tat) [37,38] (Model 1.7), and (e) a model that assumes that participants are motivated to minimize inequity in payments [52,55] (Model 1.8). Model performance was measured and compared using the AIC [115], which rewards model fit and penalizes model complexity (number of free parameters). See *SI Methods*.

Model comparison for help-acceptance decisions

We compared Model 2.1 with other plausible models, including: (a) models that solely included the term for communal concern (Model 2.2) or obligation (Model 2.3) besides the self-interest term, (c) models with separate parameters for self-interest, communal

concern and obligation with separate parameters (Model 2.4 and Model 2.5). Model performance was measured and compared using the AIC [115]. See *SI Methods*.

Parameter recovery.

Covariance between model terms implies that there might be multiple configurations of parameters that can produce the same predicted behavior. This means that, in practice, the more that these constructs covary, the less identifiable our parameters will become. We conducted parameter recovery analyses to ensure that our models were robustly identifiable [116]. To this end, we simulated data for each participant using our model parameters and the data from each trial of the experiment and compared how well we were able to recover these parameters by fitting the model to the simulated data. We refit the model using 1000 random start locations to minimize the possibility of the algorithm getting stuck in a local minimum. We then assessed the degree to which the parameters could be recovered by calculating the similarity between all the parameters estimated from the observed behavioral data and all the parameters estimated from the simulated data using a Pearson correlation.

Associations between model predictions and actual responses. To validate the model representations of appraisals/feelings, we predicted participants self-reported appraisals, emotions and the two factors extracted from EFA separately using the trial-to-trial model representations of second-order beliefs E_B'' (Eq. 3) and perceived care ω_B (Eq. 5) in the reciprocity model by conducting LMMs that included random intercepts and slopes for each participant. All variables were normalized before regression analysis.

Data analyses in Study 3 (FMRI Study)

FMRI Data Acquisition and Preprocessing. Images were acquired using a 3T Prisma Siemens scanner (Siemens AG, Erlangen, Germany) with a 64-channel head coil at

Peking University (Beijing, China). T2-weighted echoplanar images (EPI) were obtained with blood oxygenation level-dependent (BOLD) contrast. Sixty-two transverse slices of 2.3 mm thickness that covered the whole brain were acquired using multiband EPI sequence in an interleaved order (repetition time = 2000 ms, echo time = 30 ms, field of view = 224×224 mm^2, flip angle = 90°). The fMRI data preprocessing and univariate analyses were conducted using Statistical Parametric Mapping software SPM12 (Wellcome Trust Department of Cognitive Neurology, London). Images were slice-time corrected, motion corrected, resampled to 3 mm × 3 mm × 3 mm isotropic voxels, and normalized to MNI space using the EPInorm approach in which functional images are aligned to an EPI template, which is then nonlinearly warped to stereotactic space [117]. Images were then spatially smoothed with an 8 mm FWHM Gaussian filter, and temporally filtered using a high-pass filter with a cutoff frequency of 1/128 Hz.

Univariate fMRI Analyses. We used a model-based fMRI analytic approach [59] to identify brain regions that parametrically tracked different components of the computational model for reciprocity during the Outcome period of the task (5s; Fig. 3B), where participants learned about the benefactor's decision to help. To ensure that each hypothesis tested had maximum variance, we chose to separately test each hypothesis using a separate model to minimize issues with multicollinearity. GLM 1 identified reciprocity related brain responses based on the parametric modulator of participant's reciprocity behavior D_B. GLM 2 identified brain responses related to communal concern based on the parametric modulator of the participant's appraisal of perceived care ω_B. GLM 3 identified brain responses related to obligation, which we modeled as a linear contrast of the participant's second-order belief of the benefactor's expectation for repayment E_B''. We chose to use the appraisals rather than the $U_{Communal}$ and the $U_{Obligation}$ terms, as those terms create costs based on the squared deviation from reciprocity behavior, which results in a large proportion of trials where the deviations are near zero as a result of participant's decisions, making them inefficient for

parametric analysis to capture how successfully participants behaved in accordance with their feelings. Instead, ω_B and E_B'' better captured the inferences that comprised participants' feelings and were more suitable for testing our hypotheses about brain responses.

Regressors for GLM1 and GLM 2 included: (a) Outcome period (onset of the presentation of the benefactor's decision, 5s) with the corresponding parametric modulator, (b) Information period (onset of the presentation of the benefactor's picture and extra information regarding intention, 4s), (c) Second-order belief rating period (starting from the time the rating screen presented and spanning to the time that the participant made choice), (d) Allocation period (starting from the time the rating screen presented and spanning to the time that the participant made choice), (e) Missed responses (the missing decision period for second-order belief or allocation, 8s), and (f) six head motion realignment parameters. Contrasts were defined as the positive effect of the parametric modulator of interest.

For GLM3, because our computational model's representation of second order beliefs E_B'' had a non-normal distribution (zero in Repayment impossible condition and linear increase in Repayment possible condition, Eq. 3), we constructed a piecewise linear contrast, instead of linear parametric analysis. This entailed creating four separate regressors modeling different parts of the function during the Outcome period: (1) Repayment impossible, (2) Repayment possible and low benefactor's cost (i.e., 4, 6, or 8), (3) Repayment possible and medium benefactor's cost (i.e., 10, 12, or 14), (4) Repayment possible and high benefactor's cost (i.e., 16, 18, or 20). Subsequently, for each participant, we constructed a contrast vector of c = [-6, 1, 2, 3]. This piecewise linear contrast ensures that brain responses to the Repayment impossible trials are lower than all of the Repayment possible trials. We have successfully used this approach in

previous work modeling guilt using similar Psychological Game Theoretic utility models [54].

For all GLMs, events in each regressor were convolved with a double gamma canonical hemodynamic response function. Second-level models were constructed as one-sample *t* tests using contrast images from the first-level models. For whole brain analyses, all results were corrected for multiple comparisons using cluster correction $p < 0.05$ with a cluster-forming threshold of $p < 0.001$, which attempts to control for family wise error (FWE) using Gaussian Random Field Theory. This approach attempts to estimate the number of independent spatial resels or resolution elements in the data necessary to control for FWE. This calculation requires defining an initial threshold to determine the Euler Characteristic of the data. It has been demonstrated that an initial threshold of $p < 0.001$ does a reasonable job of controlling for false positives at 5% using this approach [71].

Meta-analytical Decoding. To reveal the psychological components associated with the processing of reciprocity, communal concern and obligation, we conducted meta-analytic decoding using the Neurosynth Image Decoder [61] (http://neurosynth.org). This allowed us to quantitatively evaluate the spatial similarity [60] between any Nifti-format brain image and selected meta-analytical images generated by the Neurosynth database. Using this online platform, we compared the unthresholded contrast maps of reciprocity, communal concern and obligation against the reverse inference meta-analytical maps for 23 terms generated from this database, related to basic cognition (i.e., Imagine, Switching, Salience, Conflict, Memory, Attention, Cognitive control, Inhibition, Emotion, Anxiety, Fear, and Default mode) [118], social cognition (Empathy, Theory of mind, Social, and Imitation) [119] and decision-making (Reward, Punishment, Learning, Prediction error, Choice, and Outcome) [66].

Neural Utility Model of Indebtedness. We constructed a neural utility model by combining our computational model for reciprocity with multivariate pattern analysis (MVPA) [120]. First, using principal components regression with 5-fold cross-validation, we trained two separate multivariate whole-brain models predictive of communal concern (ω_B) and obligation (E_B'') terms in our behavioral model separately for each participant [72-74]. This analysis was carried out in Python 3.6.8 using the NLTools package [121] version 0.3.14 (https://nltools.org/). This entailed first performing temporal data reduction by estimating single-trial beta maps of the Outcome period for each participant. Then for each participant, we separately predicted ω_B and E_B'' from a vectorized representation of the single trial beta maps. Because these models have considerably more voxel features (~328k) then trial observations, we performed a principal components analysis to reduce the feature space and used the principal components to predict the model appraisal representations (e.g., ω_B and E_B''). We then back-projected the estimated beta components from the regression back into the full voxel feature space, and then back to 3-D space. We have previously demonstrated that this approach is effective in reliably mapping the independent contribution of each voxel in the brain to a psychological state to identify the neural representations of affective states [73,122,123]. For each whole-brain model, we extracted the cross-validated prediction accuracy (r value) for each participant, conducted r-to-z transformation, and then conducted a one-sample sign permutation test to evaluate whether each model was able to significantly predict the corresponding term.

We used the cross-validated models to generate predictions for each trial for each participant and then input the brain-predicted communal concern and second-order beliefs into our neural utility model (Eq. 6). We estimated the θ values (i.e., weight on greed) and ϕ weighting parameters (i.e., relative trade-off between on communal concern and obligation) using the same procedure described in the behavioral

computational modeling section by fitting the neural utility model directly to participant's reciprocity behavior by minimizing the SSE (Eq. 10).

As a benchmark for our neural utility model, we were interested in determining how well we could predict participant's reciprocity behavior directly from brain activity. We used the same training procedure described above, but predicted trial-to-trial reciprocity behavior using principal components regression separately for each participant. In theory, this should provide a theoretical upper bound of the best we should be able to predict reciprocity behavior using brain activity. If our neural utility model is close, then it means that we are able to predict reciprocity behavior using brain representations of communal concern and obligation as well as the optimal linear weighting of brain weights that can predict trial-to-trial reciprocity behavior. To determine the importance of the participant-specific model parameters, we ran a permutation test to determine how well we could predict reciprocity behavior for each participant using parameters from a randomly selected different participant. We ran 5,000 permutations to generate a null distribution of average prediction accuracy after randomly shuffling the participant weights. The empirical *p*-value is the proportion of permutations that exceed our average observed correlation.

Finally, we were interested in evaluating how well we could estimate how much each participant had a relative preference for communal concern or obligation by computing the relative spatial alignment of their communal and obligation predictive spatial maps with their reciprocity predictive spatial map. We operationalized this relative pattern similarity as:

$$relative\ pattern\ similarity = corr(\vec{Obligation}_{map}, \vec{Reciprocity}_{map}) - corr(\vec{Communal}_{map}, \vec{Reciprocity}_{map})$$

Eq. 15

The intuition for this analysis is that if the optimal brain map for predicting a participant's decision is relatively more similar to their communal concern or obligation

map, then we would expect that the participant cared more about that particular component of indebtedness during behavioral decision-making. For example, if a participant weights obligation more than communal concern during reciprocity (higher $1 - \phi$ estimated from the behavioral model), then the spatial similarity between their obligation brain pattern and the pattern that directly predicts their reciprocity behavior (reciprocity brain pattern) should be relatively higher compared to the spatial similarity between their communal concern pattern and reciprocity brain pattern. We tested the correlation between this relative pattern similarity and the (1- ϕ) parameters estimated by fitting the computational model (Eq. 1) directly to the participants' behaviors.

Software

Behavioral data analyses were carried out in RStudio Version 1.1.383 [124] and IPython/Jupyter Notebook (Python 3.6.8) [107], and was plotted using matplotlib [125], and seaborn 0.9.0 (https://seaborn.pydata.org/index.html). The fMRI data preprocessing and univariate analyses were conducted using Statistical Parametric Mapping software SPM12 (Wellcome Trust Department of Cognitive Neurology, London). Unless otherwise noted, all of fMRI multivariate analyses were performed with our open source Python NLTools package [121] version 0.3.14 (https://nltools.org/).

Data availability

Behavioral data from all the three studies is available on github (https://github.com/xiaoxuepsy/Indebtedness_Gao2021). First and second level maps from the fMRI study is available on OSF (https://osf.io/k8rxh/). Raw imaging data is available from the corresponding author upon reasonable request.

Code availability

The codes used in the current study are available on github (https://github.com/xiaoxuepsy/Indebtedness_Gao2021).

Author contributions

X.G., E.J., H.Y., X.Z. and L.J.C. designed the experiments. X.G. and H. L. implemented the study design and collected the data. X.G. and L.J.C. carried out the analyses. X.G., E.J., H.Y., X.Z. and L.J.C. wrote the paper. X.Z. and L.J.C. supervised the work. All authors provided critical revisions and approved the final paper for submission.

Competing interests

The authors declare no competing interests.

Acknowledgements

We thank Dr. Matthew Rushworth, Dr. Christian C. Ruff, Dr. Rebecca Saxe and three anonymous reviewers for their comments and suggestions on this article. In addition, we thank Ms. Wan Wang, Mr. Shuaiqi Li and Mr. Sensen Song for their assistances in data collection, Ms. Yunyan Duan's for her advice in topic modeling, and Ms. Zhewen He for the preparation of the manuscript. Dr. Gao thanks Dr. Can Tang, Dr. Li Zhang, Ms. Jiyan Lyu, and Ms. Kewalin Boonsatta for their support during the revision process. This work was supported by National Natural Science Foundation of China (71942001, 31900798, 31630034), Young Elite Scientists Sponsorship Program by China Association for Science and Technology (YESS20210176, 2021QNRC001), China Postdoctoral Science Foundation (2019M650008), the National Science Foundation of USA (CAREER 1848370), the National Institute of Health (R01MH116026), the Research Project of Shanghai Science and Technology Commission (20dz2260300) and the Fundamental Research Funds for the Central Universities. We also acknowledge support from the Graduate School of Peking University to fund Dr. Gao's training at Dartmouth College.

Reference

1 Sherry Jr, J. F. Gift giving in anthropological perspective. *J. Consum. Res.* **10**, 157-168 (1983).

2 Carmichael, H. L. & MacLeod, W. B. Gift giving and the evolution of cooperation. *Int. Econ. Rev.*, 485-509 (1997).

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

4 Clark, M. S. & Mills, J. The difference between communal and exchange relationships: What it is and is not. *Pers. Soc. Psychol. Bull.* **19**, 684-691 (1993).

5 Clark, M. S. & Mills, J. R. in *Handbook of theories of social psychology, Vol. 2* 232-250 (Sage Publications Ltd, 2012).

6 Algoe, S. B. Find, remind, and bind: The functions of gratitude in everyday relationships. *Soc. Pers. Psychol. Compass* **6**, 455-469 (2012).

7 Algoe, S. B., Haidt, J. & Gable, S. L. Beyond reciprocity: gratitude and relationships in everyday life. *Emotion* **8**, 425 (2008).

8 Elfers, J. & Hlava, P. *The Spectrum of Gratitude Experience*. (Springer, 2016).

9 McCullough, M. E., Kilpatrick, S. D., Emmons, R. A. & Larson, D. B. Is gratitude a moral affect? *Psychol. Bull.* **127**, 249 (2001).

10 Trivers, R. L. The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35-57 (1971).

11 Neilson, W. S. The economics of favors. *J. Econ. Behav. Organ.* **39**, 387-397 (1999).

12 Akerlof, G. A. Labor contracts as partial gift exchange. *Q. J. Econ.* **97**, 543-569 (1982).

13 Greenberg, M. S. in *Social exchange* 3-26 (Springer, 1980).

14 Greenberg, M. S. & Westcott, D. R. Indebtedness as a mediator of reactions to aid. *New directions in helping* **1**, 85-112 (1983).

15 Regan, D. T. Effects of a favor and liking on compliance. *J. Exp. Soc. Psychol.*

7, 627-639 (1971).

16 Kolm, S.-C. *Reciprocity: An economics of social relations*. (Cambridge University Press, 2008).

17 Nadler, A. in *The Oxford handbook of prosocial behavior. Oxford library of psychology.* 307-328 (Oxford University Press, 2015).

18 Fisher, J. D., Nadler, A. & Whitcher-Alagna, S. Recipient reactions to aid. *Psychol. Bull.* **91**, 27-54 (1982).

19 Fisher, J. *New Directions in Helping: Recipient reactions to aid*. Vol. 1 (Elsevier, 1983).

20 Nadler, A., Mayseless, O., Peri, N. & Chemerinski, A. Effects of opportunity to reciprocate and self-esteem on help-seeking behavior. *J. Pers.* **53**, 23-35 (1985).

21 Watkins, P. C., Scheer, J., Ovnicek, M. & Kolts, R. The debt of gratitude: Dissociating gratitude and indebtedness. *Cognition Emotion* **20**, 217-241 (2006).

22 Bal, A. Doctors and drug companies. *N. Engl. J. Med.* **352**, 733-734 (2005).

23 Malmendier, U. & Schmidt, K. You owe me. (National Bureau of Economic Research, 2012).

24 Fehr, E. & Gächter, S. Fairness and retaliation: The economics of reciprocity. *J. Econ. Perspect.* **14**, 159-181 (2000).

25 Gonzalez, B. & Chang, L. J. Computational models of mentalizing. (2019).

26 Falk, A., Fehr, E. & Fischbacher, U. On the nature of fair behavior. *Econ. Inq.* **41**, 20-26 (2003).

27 Sul, S., Guroglu, B., Crone, E. A. & Chang, L. J. Medial prefrontal cortical thinning mediates shifts in other-regarding preferences during adolescence. *Sci. Rep.* **7**, 8510 (2017).

28 Ellsworth, P. C. & Scherer, K. R. Appraisal processes in emotion. *Handbook of affective sciences* **572**, V595 (2003).

29 Frijda, N. H. The Place of Appraisal in Emotion. *Cognition Emotion* **7**, 357-387 (1993).

30 Frijda, N. H., Kuipers, P. & Ter Schure, E. Relations among emotion, appraisal, and emotional action readiness. *J. Pers. Soc. Psychol.* **57**, 212 (1989).

31 Lazarus, R. S. & Smith, C. A. Knowledge and appraisal in the cognition—emotion relationship. *Cognition Emotion* **2**, 281-300 (1988).

32 Scherer, K. R. Appraisal theory. (1999).

33 Smith, C. A. & Ellsworth, P. C. Patterns of cognitive appraisal in emotion. *J. Pers. Soc. Psychol.* **48**, 813 (1985).

34 Battigalli, P. & Dufwenberg, M. Dynamic psychological games. *J. Econ. Theory.* **144**, 1-35 (2009).

35 Battigalli, P., Corrao, R. & Dufwenberg, M. Incorporating belief-dependent motivation in games. *J. Econ. Behav. Organ.* **167**, 185-218 (2019).

36 Geanakoplos, J., Pearce, D. & Stacchetti, E. Psychological games and sequential rationality. *Game. Econ. Behav.* **1**, 60-79 (1989).

37 Dufwenberg, M. & Kirchsteiger, G. A theory of sequential reciprocity. *Game. Econ. Behav.* **47**, 268-298 (2004).

38 Rabin, M. Incorporating fairness into game theory and economics. *Am. Econ. Rev.*, 1281-1302 (1993).

39 Chang, L. J. & Smith, A. Social emotions and psychological games. *Curr. Opin. Behav. Sci.* **5**, 133-140 (2015).

40 Benedict, R. Chrysanthemum and the Sword. Patterns of Japanese Culture, Cleveland, New York (The World Publishing Company) 1946. (1946).

41 Kotani, M. Expressing gratitude and indebtedness: Japanese speakers' use of "I'm sorry" in English conversation. *Res. Lang. Soc. Interac.* **35**, 39-72 (2002).

42 Naito, T. & Washizu, N. Note on cultural universals and variations of gratitude from an East Asian point of view. *J. Behav. Sci.* **10**, 1-8 (2015).

43 Washizu, N. & Naito, T. The emotions sumanai, gratitude, and indebtedness, and their relations to interpersonal orientation and psychological well-being among Japanese university students. *International Perspectives in Psychology:*

Research, Practice, Consultation **4**, 209 (2015).

44 Baumeister, R. F., Stillwell, A. M. & Heatherton, T. F. Guilt: an interpersonal approach. *Psychol. Bull.* **115**, 243-267 (1994).

45 Le, B. M., Impett, E. A., Lemay Jr, E. P., Muise, A. & Tskhay, K. O. Communal motivation and well-being in interpersonal relationships: An integrative review and meta-analysis. *Psychol. Bull.* **144**, 1-25 (2018).

46 Naito, T. & Sakata, Y. Gratitude, Indebtedness, and Regret on Receiving a Friend's Favor in Japan. *Psychologia* **53**, 179-194 (2010).

47 Tsang, J. A. The effects of helper intention on gratitude and indebtedness. *Motiv. Emotion* **30**, 199-205 (2006).

48 Rotella, A., Sparks, A. M. & Barclay, P. Feelings of obligation are valuations of signaling-mediated social payoffs. *Behav. Brain Sci.* **43**, e85 (2020).

49 Tomasello, M. The Moral Psychology of Obligation. *Behav. Brain Sci.*, 1-33 (2019).

50 Beeler-Duden, S., Yucel, M. & Vaish, A. The role of affect in feelings of obligation. *Behav. Brain Sci.* **43**, e60 (2020).

51 Theriault, J. E., Young, L. & Barrett, L. F. The sense of should: A biologically-based framework for modeling social pressure. *Phys. Life Rev.* **36**, 100-136 (2021).

52 Fehr, E. & Schmidt, K. M. A theory of fairness, competition, and cooperation. *Q. J. Econ.* **114**, 817-868 (1999).

53 Blei, D. M. & Lafferty, J. D. in *Proceedings of the 23rd international conference on Machine learning.* 113-120 (ACM).

54 Chang, L. J., Smith, A., Dufwenberg, M. & Sanfey, A. G. Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron* **70**, 560-572 (2011).

55 van Baar, J. M., Chang, L. J. & Sanfey, A. G. The computational and neural substrates of moral strategies in social decision-making. *Nat. Commun.* **10**, 1-

14 (2019).

56 Browne, M. W. & Cudeck, R. Alternative Ways of Assessing Model Fit. *Sociological Methods & Research* **21**, 230-258 (1992).

57 Hu, L. Evaluating model fit. *Structural equation modelling : concepts, issues and applications*, 76-99 (1995).

58 West, S. G., Taylor, A. B. & Wu, W. in *Handbook of structural equation modeling.* 209-231 (The Guilford Press, 2012).

59 O'doherty, J. P., Hampton, A. & Kim, H. Model‐based fMRI and its application to reward learning and decision making. *Ann. N. Y. Acad. Sci.* **1104**, 35-53 (2007).

60 Chang, L. J., Yarkoni, T., Khaw, M. W. & Sanfey, A. G. Decoding the role of the insula in human cognition: functional parcellation and large-scale reverse inference. *Cereb. Cortex* **23**, 739-749 (2013).

61 Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods* **8**, 665 (2011).

62 Fox, G. R., Kaplan, J., Damasio, H. & Damasio, A. Neural correlates of gratitude. *Front. psychol.* **6** (2015).

63 Yu, H., Cai, Q., Shen, B., Gao, X. & Zhou, X. Neural substrates and social consequences of interpersonal gratitude: Intention matters. *Emotion* **17**, 589-601 (2017).

64 Yu, H., Gao, X., Zhou, Y. & Zhou, X. Decomposing gratitude: representation and integration of cognitive antecedents of gratitude in the brain. *J. Neurosci.*, 2944-2917 (2018).

65 Cooper, J. C., Kreps, T. A., Wiebe, T., Pirkl, T. & Knutson, B. When giving is good: ventromedial prefrontal cortex activation for others' intentions. *Neuron* **67**, 511-521 (2010).

66 Ruff, C. C. & Fehr, E. The neurobiology of rewards and values in social decision

making. *Nat. Rev. Neurosci.* **15**, 549 (2014).

67 Koban, L., Corradi-Dell'Acqua, C. & Vuilleumier, P. Integration of error agency and representation of others' pain in the anterior insula. *J. Cogn. Neurosci.* **25**, 258-272 (2013).

68 Yu, H., Hu, J., Hu, L. & Zhou, X. The voice of conscience: neural bases of interpersonal guilt and compensation. *Soc. Cogn. Affect. Neurosci.* **9**, 1150-1158 (2014).

69 Hampton, A. N., Bossaerts, P. & O'Doherty, J. P. Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 6741-6746 (2008).

70 Van Overwalle, F. & Baetens, K. Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *Neuroimage* **48**, 564-584 (2009).

71 Woo, C.-W., Krishnan, A. & Wager, T. D. Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *Neuroimage* **91**, 412-419 (2014).

72 Woo, C. W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365-377 (2017).

73 Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A. & Wager, T. D. A sensitive and specific neural signature for picture-induced negative affect. *PLoS Biol.* **13**, e1002180 (2015).

74 Wager, T. D. *et al.* An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* **368**, 1388-1397 (2013).

75 Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).

76 Mathews, M. A. & Green, J. D. Looking at me, appreciating you: Self-focused attention distinguishes between gratitude and indebtedness. *Cognition Emotion*

24, 710-718 (2010).

77 Lench, H. C., Flores, S. A. & Bench, S. W. Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: A meta-analysis of experimental emotion elicitations. *Psychol. Bull.* **137**, 834-855 (2011).

78 Lindquist, K. A., Siegel, E. H., Quigley, K. S. & Barrett, L. F. The hundred-year emotion war: are emotions natural kinds or psychological constructions? Comment on Lench, Flores, and Bench (2011). *Psychol. Bull.* **139**, 255-263 (2013).

79 Larsen, R. J. & Fredrickson, B. L. in *Well-being: The foundations of hedonic psychology.* 40-60 (Russell Sage Foundation, 1999).

80 Nisbett, R. E. & Wilson, T. D. Telling more than we can know: Verbal reports on mental processes. *Psychol. Rev.* **84**, 231-259 (1977).

81 Jolly, E. & Chang, L. J. The Flatland Fallacy: Moving Beyond Low-Dimensional Thinking. *Top. Cogn. Sci.* **11**, 433-454 (2019).

82 Chang, L. J. & Jolly, E. Emotions as computational signals of goal error. *The nature of emotion: Fundamental questions*, 343-348 (2018).

83 Xiang, T., Lohrenz, T. & Montague, P. R. Computational substrates of norms and their violations during social exchange. *J. Neurosci.* **33**, 1099-1108a (2013).

84 Gao, X. *et al.* Distinguishing neural correlates of context-dependent advantageous- and disadvantageous-inequity aversion. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E7680-E7689 (2018).

85 Khalmetski, K., Ockenfels, A. & Werner, P. Surprising gifts: Theory and laboratory evidence. *J. Econ. Theory.* **159**, 163-208 (2015).

86 Battigalli, P., Dufwenberg, M. & Smith, A. Frustration and Anger in Games. (2015).

87 Chang, L. J. & Sanfey, A. G. Great expectations: neural computations underlying the use of social norms in decision-making. *Soc. Cogn. Affect. Neurosci.* **8**, 277-284 (2013).

88 Krajbich, I., Adolphs, R., Tranel, D., Denburg, N. L. & Camerer, C. F. Economic games quantify diminished sense of guilt in patients with damage to the prefrontal cortex. *J. Neurosci.* **29**, 2188-2192 (2009).

89 Smith, A., Bernheim, B. D., Camerer, C. & Rangel, A. Neural Activity Reveals Preferences Without Choices. *Nber Working Papers* **6**, 1-36 (2014).

90 Knutson, B., Rick, S., Wimmer, G. E., Prelec, D. & Loewenstein, G. Neural Predictors of Purchases. *Neuron* **53**, 147-156 (2007).

91 Hein, G., Morishima, Y., Leiberg, S., Sul, S. & Fehr, E. The brain's functional network architecture reveals human motives. *Science* **351**, 1074-1078 (2016).

92 Haidt, J. The moral emotions. *Handbook of affective sciences* **11**, 852-870 (2003).

93 Fiske, A. P. The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychol. Rev.* **99**, 689-723 (1992).

94 Rai, T. S. & Fiske, A. P. Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychol. Rev.* **118**, 57-75 (2011).

95 Fiske, A. P. Socio-moral emotions motivate action to sustain relationships. *Self and Identity* **1**, 169-175 (2002).

96 van Baar, J. M., Klaassen, F. H., Ricci, F., Chang, L. J. & Sanfey, A. G. Stable distribution of reciprocity motives in a population. *Sci. Rep.* **10**, 18164 (2020).

97 Yu, H. *et al.* A Generalizable Multivariate Brain Pattern for Interpersonal Guilt. *Cereb. Cortex* **30**, 3558-3572 (2020).

98 Inui, K., Tran, T. D., Hoshiyama, M. & Kakigi, R. Preferential stimulation of Adelta fibers by intra-epidermal needle electrode in humans. *Pain* **96**, 247-252 (2002).

99 Liu, Q. in *International Conference on Computer Science & Network Technology.* (2016).

100 Neto, J. L., Santos, A. D., Kaestner, C. A., Alexandre, N. & Santos, D.

Document clustering and text summarization. (2000).

101 Salton, G. & Buckley, C. Term-weighting approaches in automatic text retrieval. *Inform. Process. Manag.* **24**, 513-523 (1988).

102 Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993-1022 (2003).

103 Cowen, A. S. & Keltner, D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences* **114**, E7900 (2017).

104 Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).

105 Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823* (2014).

106 Blair, R. J. The neurobiology of psychopathic traits in youths. *Nat Rev Neurosci* **14**, 786-799 (2013).

107 Pérez, F. & Granger, B. E. IPython: a system for interactive scientific computing. *Comput. Sci. Eng.* **9** (2007).

108 Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Methods* **4**, 272-299 (1999).

109 Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. & Tatham, R. *Multivariate data analysis*. (Uppersaddle River, 2006).

110 Tobias, S. & Carlson, J. E. Brief report: Bartlett's test of sphericity and chance findings in factor analysis. *Multivar. Behav. Res.* **4**, 375-377 (1969).

111 Revelle, W. An overview of the psych package. *Department of Psychology Northwestern University. Accessed on March* **3**, 2012 (2011).

112 Rosseel, Y. lavaan: An R package for structural equation modeling. *J. Stat. Softw.* **48**, 1-36 (2012).

113 Bolton, G. E. & Ockenfels, A. ERC: A theory of equity, reciprocity, and

competition. *Am. Econ. Rev.* **90**, 166-193 (2000).

114 Greenberg, M. S. & Shapiro, S. P. Indebtedness: An adverse aspect of asking for and receiving help. *Sociometry*, 290-301 (1971).

115 Akaike, H. A new look at the statistical model identification. *IEEE transactions on automatic control* **19**, 716-723 (1974).

116 Fareri, D. S., Chang, L. J. & Delgado, M. R. Computational substrates of social value in interpersonal collaboration. *J. Neurosci.* **35**, 8170-8180 (2015).

117 Calhoun, V. D. *et al.* The impact of T1 versus EPI spatial normalization templates for fMRI data analyses. *Hum. Brain Mapp.* **38**, 5331-5342 (2017).

118 Barrett, L. F. & Satpute, A. B. Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain. *Curr. Opin. Neurobiol.* **23**, 361-372 (2013).

119 Adolphs, R. The social brain: neural basis of social knowledge. *Annu. Rev. Psychol.* **60**, 693-716 (2009).

120 Haynes, J.-D. & Rees, G. Neuroimaging: decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* **7**, 523 (2006).

121 cosanlab/nltools: 0.3.11 v. 0.3.11 (Zenodo, 2018).

122 Chang, L. J. *et al.* Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience. *Science Advances* **7**, eabf7129 (2021).

123 Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365-377 (2017).

124 Racine, J. S. RStudio: A Platform‐Independent IDE for R and Sweave. *J. Appl. Economet.* **27**, 167-172 (2012).

125 Hunter & John, D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90-95 (2007).