The hidden cost of receiving favors:

A theory of indebtedness

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

Xiaolin Zhou[1,2,6,7 *], *Luke J. Chang*[3 *]

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

[2] Beijing Key Laboratory of Behavior and Mental Health, 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] School of Business and Management, Shanghai International Studies University,
Shanghai 200083, China

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

*Correspondence to:

Luke J. Chang (luke.j.chang@dartmouth.edu) and Xiaolin Zhou (xz104@pku.edu.cn)

Abstract

Receiving help or a favor from another person can sometimes have a hidden cost. In this study, we explore these hidden costs by developing and validating a theoretical model of indebtedness across three studies that combine large-scale experience sampling, interpersonal games, computational modeling, and neuroimaging. Our model captures how individuals infer the altruistic and strategic intentions of the benefactor. These inferences produce distinct feelings of guilt and obligation that together comprise indebtedness and motivate reciprocity. Altruistic intentions convey care and concern and are associated with activity in the insula, dorsolateral prefrontal cortex and ventromedial prefrontal cortex, while strategic intentions convey expectations of future reciprocity and are associated with activation in the 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

Introduction

Giving gifts and exchanging favors are ubiquitous behaviors that provide a concrete expression of a relationship between individuals or groups (Carmichael and MacLeod, 1997; Sherry Jr, 1983). Altruistic favors convey concern for a partner's well-being and signal a communal relationship such as a friendship, romance, or familial tie (Clark and Mills, 1993; Clark and Mills, 2012; Nowak and Sigmund, 2005). 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 (Algoe, 2012; Algoe et al., 2008; Elfers and Hlava, 2016; McCullough et al., 2001). Yet in daily life, favors and gifts can also be strategic and imply an expectation of reciprocal exchanges, particularly in more transactive relationships (Akerlof, 1982; Carmichael and MacLeod, 1997; Clark and Mills, 1993; Clark and Mills, 2012; Neilson, 1999; Trivers, 1971). 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 (Greenberg, 1980; Greenberg and Westcott, 1983; Kolm, 2008; Regan, 1971). 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 (Bal, 2005; Malmendier and Schmidt, 2012), or a politician might vote favorably on policies that benefit an organization, which provided generous campaign contributions (Fehr and Gächter, 2000). However, very little is known about the psychological 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 (Falk et al., 2003; Gonzalez and Chang, 2019; Sul et al., 2017). Does this person care about me? Or do they expect something in return? According to appraisal theory (Ellsworth and

Scherer, 2003; Frijda, 1993; Frijda et al., 1989; Lazarus and Smith, 1988; Scherer, 1999; Smith and Ellsworth, 1985), these types of cognitive appraisals are critical in determining what types of emotions are experienced and how the beneficiary will ultimately respond. Psychological Game Theory (PGT) (Battigalli et al., 2019; Battigalli and Dufwenberg, 2009; Geanakoplos et al., 1989) has provided tools for modeling these higher order beliefs about intentions, expectations, and fairness in the context of reciprocity decisions (Dufwenberg and Kirchsteiger, 2004; Falk et al., 2003; Rabin, 1993; Sul et al., 2017). 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 (Dufwenberg and Kirchsteiger, 2004; Falk et al., 2003; Rabin, 1993; Sul et al., 2017). These intention inferences can produce different emotions in the beneficiary (Chang and Smith, 2015). For example, if the benefactor's actions are believed to be altruistic and convey concern for the beneficiary's outcome, the beneficiary is likely to experience gratitude, but may also feel personally responsible for burdening the benefactor and experience the negative feeling of guilt (Baumeister et al., 1994; Benedict, 1946; Chang et al., 2011; Kotani, 2002; Naito and Washizu, 2015). Both of these feelings motivate reciprocity out of concern for the benefactor, i.e., communal concern (Baumeister et al., 1994; Le et al., 2018). In contrast, if the benefactors' intentions are perceived to be strategic or even duplicitous, then the beneficiary is more likely to feel a negative feeling of obligation, which can also motivate reciprocity (Greenberg, 1980; Greenberg and Westcott, 1983; Watkins et al., 2006). This obligation-based reciprocity is likely driven by external pressures, such as social expectations and potential reputational costs, rather than the communal concern for the benefactor (Rotella et al., 2020). In everyday life, inferences about a benefactor's intentions are often mixed, raising the possibility that the negative feeling of indebtedness in response to favors may be comprised of feelings of communal concern (i.e., guilt) and obligation.

In this study, we propose a theoretical model of indebtedness to characterize how the beneficiaries' appraisals and emotions lead to reciprocal behaviors (Fig. 1). Specifically, we propose that there are two 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 reciprocal behaviors. 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 increases 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). Building on previous models of other-regarding preferences (Dufwenberg and Kirchsteiger, 2004; Fehr and Schmidt, 1999; Rabin, 1993), we model the utility associated with reciprocal behaviors as reflecting the trade-off between these different feelings (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 of this model is that upon receiving a favor from a benefactor (player A), the beneficiary (player B) chooses an action (D_B) that maximizes his/her overall utility (U), where utility is comprised of a mixture of values arising from self-interest (π) weighted by a greed parameter Θ, and feelings of communal concern and obligation ($U_{Communal}$ and $U_{Obligation}$), which are weighted by the parameter Φ. Larger Φ values reflect the beneficiary's higher sensitivity to feelings of communal concern relative to obligation.

In this paper, we validate the predictions of our model across multiple studies. In Study 1 (N = 1619), we explore lay intuitions of indebtedness using large-scale experience sampling. In Study 2 (Study 2a, N = 51; Study 2b, N = 57), we evaluate

how different components of indebtedness are generated and influence behaviors by combining computational modeling with an interpersonal game, in which benefactors choose to spend some amount of their initial endowment to reduce the amount of pain experienced by the participants. In Study 3 (N = 53), we investigate how different feelings of indebtedness are represented in the brain using functional magnetic resonance imaging (fMRI) and how they vary across individuals.

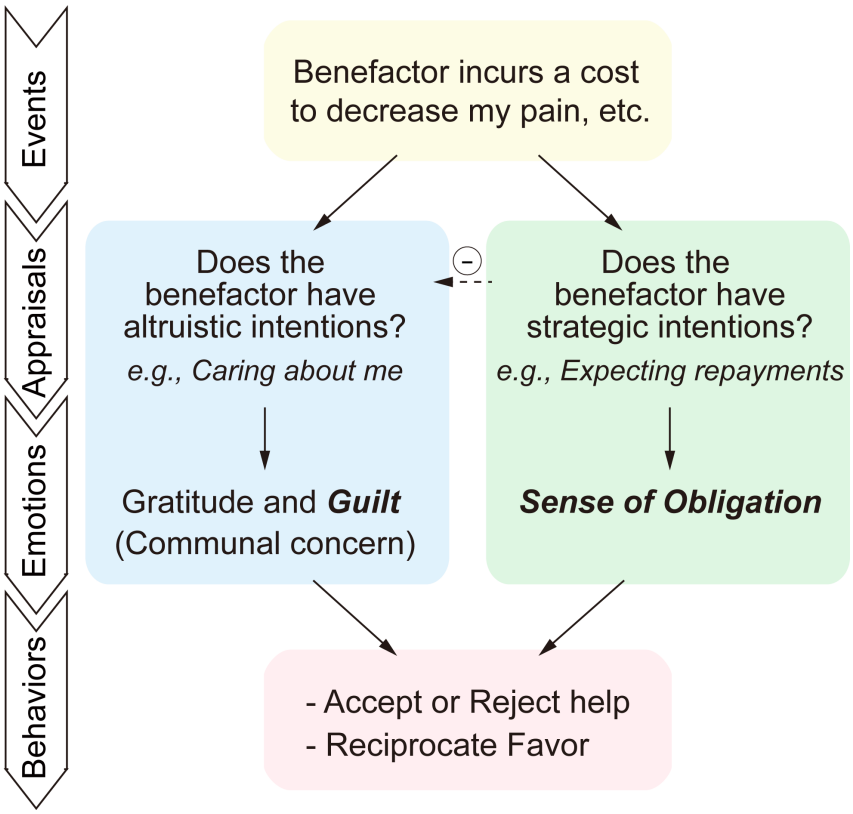

Fig. 1 Theoretical model of indebtedness. We propose that there are two components of indebtedness, guilt and the sense of obligation, which are derived from appraisals about the benefactor's altruistic and strategic intentions and can differentially impact the beneficiary's reciprocal behaviors. The 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 increases 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). The beneficiary makes trade-offs between communal and obligation feelings to determine the reciprocal behaviors to favors (e.g., accept or reject the help and reciprocity after receiving help).

Results

Indebtedness is a mixed feeling comprised of guilt and obligation

In Study 1, we used an online questionnaire to characterize the subjective experience of indebtedness in Chinese participants. First, participants (N = 1,619) described specific experiences, 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 explained indebtedness ratings (β_{guilt} = 0.70 ± 0.02, t = 40.08, p < 0.001; $\beta_{obligation}$ = 0.40 ± 0.02, t = 2.31, p = 0.021; Fig. 2A-I). Models with both guilt and obligation ratings outperformed models with only a single predictor (Full model vs. guilt-only model: F = 5.34, p = 0.021, Full model vs. obligation-only model: F = 1606.1, p < 0.001, Table S1). Second, participants were asked to select sources of indebtedness in their daily lives and 91.9% attributed the guilt for burdening the benefactor and 39.2% indicated the sense of obligation resulting from the benefactor's ulterior intention as the source of indebtedness (Fig. 2A-II, Fig. S1A). Third, participants were asked to describe their own personal definitions of indebtedness. The 100 words with the highest frequency in the definitions of indebtedness were annotated by an independent sample of participants (N = 80) to extract the emotion-related words. We applied Latent Dirichlet Allocation (LDA) based topic modeling (Blei and Lafferty, 2006) to the emotion words to demonstrate that indebtedness is comprised of 2 latent topics (Fig. S1B). Topic 1 accounted for 77.0% of the emotional words, including communal-concern-related words such as "guilt," "feel," "feel sorry," "feel indebted," and "gratitude". In contrast, Topic 2 accounted for 23.0% of the emotional words, including words pertaining to burden and negative bodily states, such as "uncomfortable," "uneasy," "trouble," "pressure," and "burden" (Fig. 2A-III, see supplementary materials). These results suggest that the subjective experience of indebtedness is comprised of feelings of both guilt and obligation.

Next we examined how participants' emotions ratings were related to their self-reported response to the help (Fig. 2B). We found that gratitude, indebtedness, guilt, and the sense of obligation positively predicted participants' reported need to repay after receiving help (gratitude: $\beta = 0.45\pm0.03$, $t = 9.52$, $p < 0.001$; indebtedness: $\beta = 0.34\pm0.03$, $t = 12.86$, $p < 0.001$; guilt: $\beta = 0.32\pm0.03$, $t = 11.13$, $p < 0.001$; obligation: $\beta = 0.19\pm0.04$, $t = 4.90$, $p < 0.001$). However, decisions to reject help were negatively predicted by anticipatory gratitude ($\beta = -0.71\pm0.05$, $t = 9.52$, $p < 0.001$), but positively predicted by anticipatory indebtedness, guilt, and obligation (indebtedness: $\beta = 0.40\pm0.06$, $t = 7.16$, $p < 0.001$; guilt: $\beta = 0.54\pm0.05$, $t = 9.97$, $p < 0.001$; obligation: $\beta = 0.55\pm0.05$, $t = 10.99$, $p < 0.001$). These results suggest the dual components of indebtedness (i.e., guilt and the sense of obligation) along with gratitude influence the behavioral responses to other's favors.

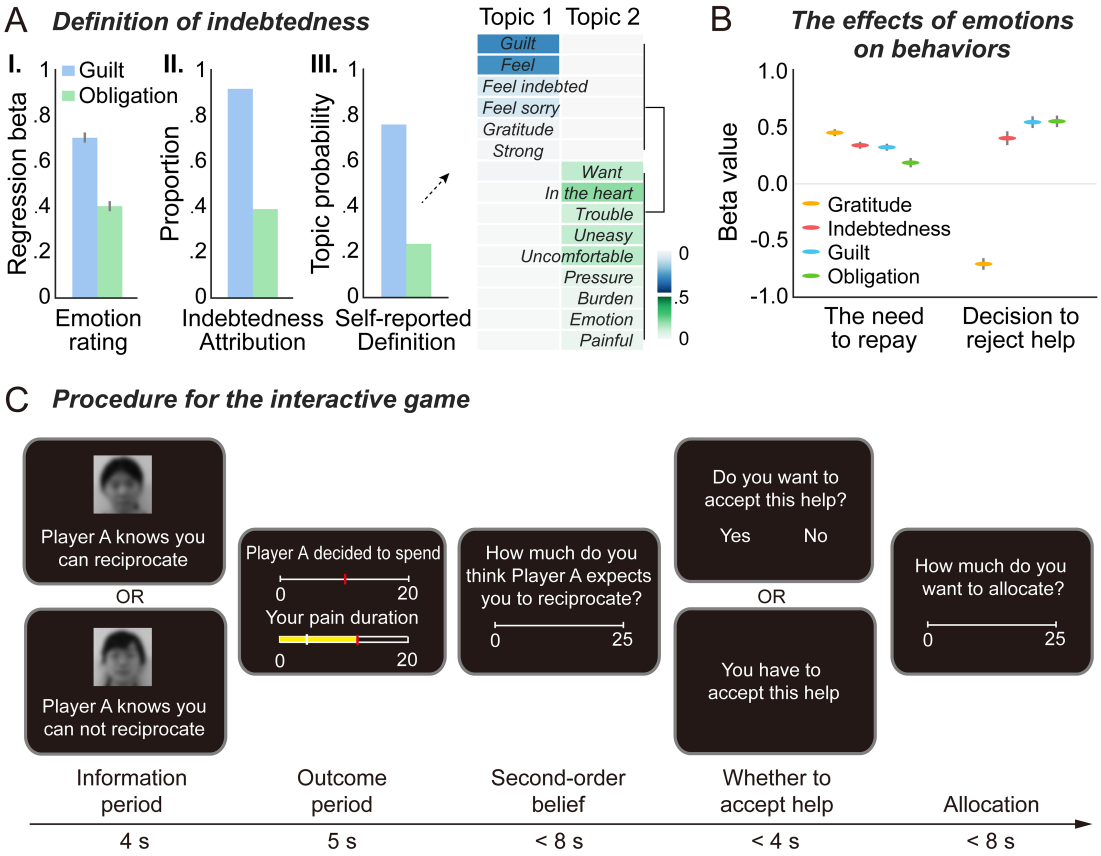

Fig. 2 Subjective experiences of indebtedness. (A) Contributions of guilt and obligation to indebtedness in Study 1 in (I) the emotion ratings in the daily event recalling, (II) attribution of guilt and obligation as source of indebtedness, and (III)

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. **(C)** Procedure for the interactive game. In each round, the participant was paired with a different anonymous co-player, who decided how much endowment to spend (i.e., benefactor's cost) to reduce the participant's pain duration. Participants indicated how much they thought this co-player expected them to reciprocate (i.e., second-order belief of the benefactor's expectation for repayment). In half of the trials, participants could decide whether to accept the help; in the remaining trials, participants had to accept help and could reciprocate by allocating monetary points to the co-player. We manipulated the perception of the benefactor's intentions by providing information about whether the co-player knew the participant could (Strategic condition), or could not (Altruistic condition) reciprocate after receiving help. After the experiment, all trials were displayed again and participants recalled their perceived care, gratitude, indebtedness, sense of obligation and guilt when they received the help. Error bars represent the standard error of means.

Benefactor's intentions lead to diverging components of indebtedness.

Next, we tested the predictions of the theoretical model of indebtedness using a laboratory-based task involving interactions between participants (Fig. 2C). In each round of the task, the participant was paired with a different anonymous co-player, who decided how much of their endowment to spend (i.e., benefactor's cost) to reduce the participant's duration of pain (i.e., electrical stimulation). Unbeknownst to participants, co-players' decisions were pre-determined by the computer program (Table S2). Participants indicated how much they thought this co-player expected them to reciprocate (i.e., second-order belief of the benefactor's expectation for repayment). We manipulated perceptions of the benefactor's intentions by providing information about whether the benefactor knew that the participant could (Strategic condition) or could not (Altruistic condition) reciprocate after receiving help. In half of the trials, participants could decide whether to accept the help; in the remaining trials, participants were only allowed to accept help and could reciprocate by allocating monetary points to the co-player regardless of the condition. After the experiment, participants recalled how much they believed the benefactor cared for

them, as well as their feelings of gratitude, indebtedness, sense of obligation, and guilt when they received the help for each trial. We manipulated information about the benefactor's intentions and benefactor's cost in Study 2a (N = 51), and further manipulated the exchange rate between the benefactor's cost and the participant's benefit (i.e., the help efficiency) in Study 2b (N = 57) (Table S2). As results were replicated in studies 2a and 2b (Table S3), for brevity we combine these datasets when reporting results in the main text.

Our theoretical 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 benefactor. Consistent with this prediction, participants reported feeling indebted in both conditions, but slightly more in the Altruistic compared to the Strategic condition (Fig. 3A, Fig. S2A, $\beta = 0.09 \pm 0.03$, $t = 2.98$, $p = 0.004$). Moreover, our manipulation successfully impacted participants' appraisals, as participants reported increased second-order beliefs of the benefactor's expectations for repayment ($\beta = 0.53 \pm 0.03$, $t = 15.71$, $p < 0.001$) and decreased perceived care ($\beta = -0.31 \pm 0.02$, $t = -13.90$, $p < 0.001$) in the Strategic compared to the Altruistic condition (Fig. 3A, see Table S3 for a summary of results). Both of these effects were magnified as the benefactor's cost increased (Fig. 3, B-C; second-order belief: $\beta = 0.22 \pm 0.02$, $t = 13.13$, $p < 0.001$; perceived care: $\beta = -0.08 \pm 0.01$, $t = -6.65$, $p < 0.001$). In addition, perceived care was negatively associated with second-order beliefs ($\beta = -0.44 \pm 0.04$, $t = -11.29$, $p < 0.001$) controlling for the effects of experimental variables (benefactor's intention, cost, and efficiency).

The manipulation of information regarding benefactors' intentions not only impacted the participants' appraisals, but also their emotions. Participants reported feeling a greater sense of obligation (Fig. 3A, Fig. S2B, $\beta = 0.30 \pm 0.03$, $t = 9.28$, $p < 0.001$), but less gratitude and guilt (Fig. 3A, Fig. S2, C-D; gratitude: $\beta = -0.27 \pm 0.02$, $t = -13.18$, p

< 0.001; guilt: β = -0.25±0.02, t = -10.30, $p < 0.001$), in the Strategic condition relative to the Altruistic condition. 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, $p < 0.001$; gratitude: β = -0.06±0.01, t = -4.20, $p < 0.001$; guilt: β = -0.05±0.01, t = -4.28, $p < 0.001$). A principal component analysis (PCA) on the subjective appraisals and emotion ratings revealed that 77% of the variance in ratings could be explained by two principal components (PCs) (Fig. 3, D-E, and Fig. S2E), which appeared to reflect two distinct subjective experiences. PC 1 reflected participants' perception that the benefactor cared about their welfare and resulted in emotions of gratitude and guilt, while PC2 reflected participants' second-order beliefs about the benefactor's expectation for repayment and the sense of obligation. Interestingly, indebtedness moderately loaded on both PCs. This interpretation was further supported by mediation analyses. Second-order beliefs mediated the effects of the experimental variables (benefactor's intention, cost, and efficiency) on obligation (Indirect effect = 0.34±0.03, $Z = 11.729$, $p < 0.001$, Fig. S3, A-B), whereas perceived care mediated the effects of experimental variables on gratitude and guilt (Indirect effect = 0.34±0.04, Z = 10.00, $p < 0.001$, Fig. S3, C-D). Together, these results provide further support for the predictions of our theoretical model that indebtedness is comprised of two distinct feelings. The guilt component of indebtedness, along with gratitude, arises from the belief that the benefactor acts from altruistic intentions, while the obligation component of indebtedness arises when the benefactor's intentions are perceived to be strategic.

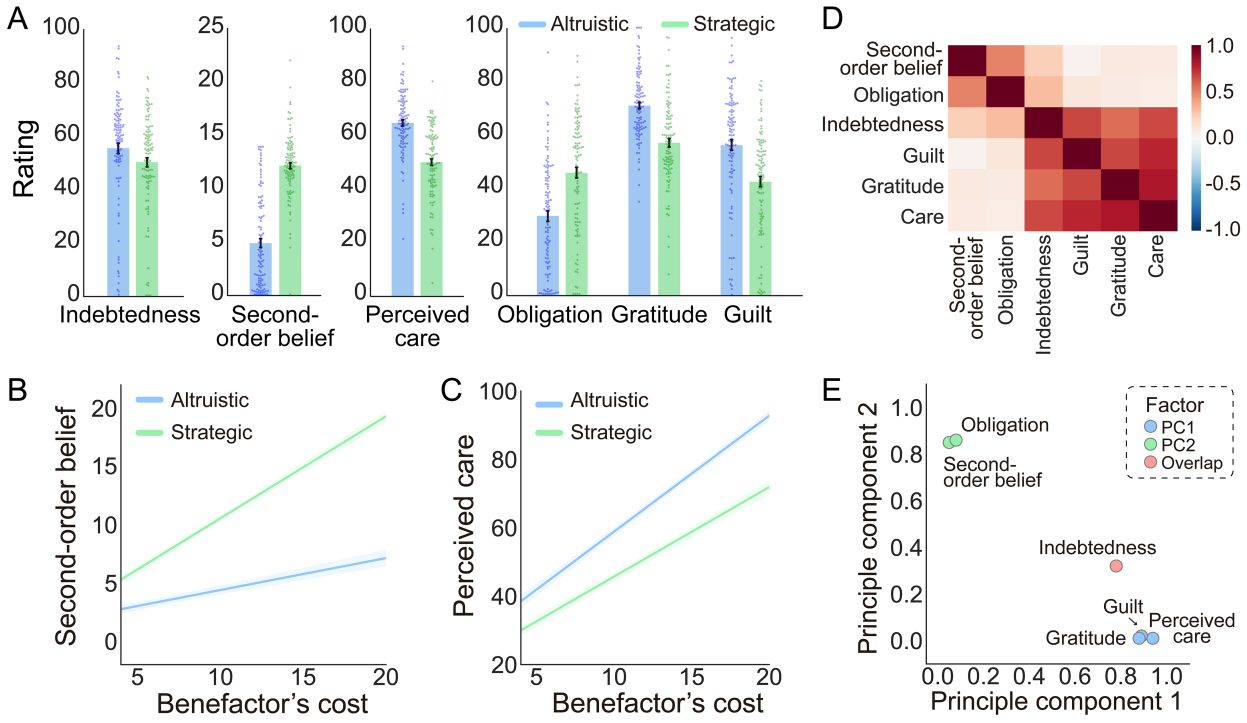

Fig. 3 Appraisals and emotional responses to benefactor's help with altruistic versus strategic intentions. (A) Participant's appraisal (i.e., second-order belief of how much the benefactor expected for repayment and perceived care) and emotion ratings (indebtedness, the sense of obligation, gratitude and guilt) in Altruistic and Strategic 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 repayment and perceived care plotted as functions of the benefactor's intention and cost. **(D)** Correlation matrix between participant's appraisal and emotion ratings. **(E)** Principal component analysis showed that participants' appraisals and emotions could be reduced to two principal components (PCs), which appeared to reflect two distinct subjective experiences. PC 1 reflects participants' perception that the benefactor cared about their welfare and resulted in emotions of gratitude and guilt, while PC2 reflects participants' second-order beliefs about the benefactor's expectation for repayment and the sense of obligation. Error bars represent the standard error of means.

Behavioral responses to help are influenced by benefactor's intentions

Next, we examined participant's behaviors in response to receiving help from a benefactor. Specifically, we were interested in whether participants would reciprocate the favor by sending some of their own money back to the beneficiary and also whether they might outright reject the beneficiary's help given the opportunity. These

behaviors comprise two crucial reciprocal responses in the beneficiary indicated by previous studies on indebtedness (Greenberg, 1980; Greenberg and Shapiro, 1971; Greenberg and Westcott, 1983). We found that participants reciprocated more money as the benefactor's cost increased in both conditions, β = 0.64±0.02, t = 25.77, p < 0.001. This effect was slightly enhanced in the Altruistic relative to the Strategic condition, β = 0.03±0.01, t = 3.02, p = 0.003 (Fig. 4A). A logistic regression revealed that when given the chance to reject the help, participants were more likely to reject help in the Strategic condition where they reported more sense of obligation (rejection rate = 0.37±0.10), compared to the Altruistic condition (rejection rate = 0.30±0.03), β = 0.28±0.10, z = 617.00, p < 0.001 (Fig. 4B).

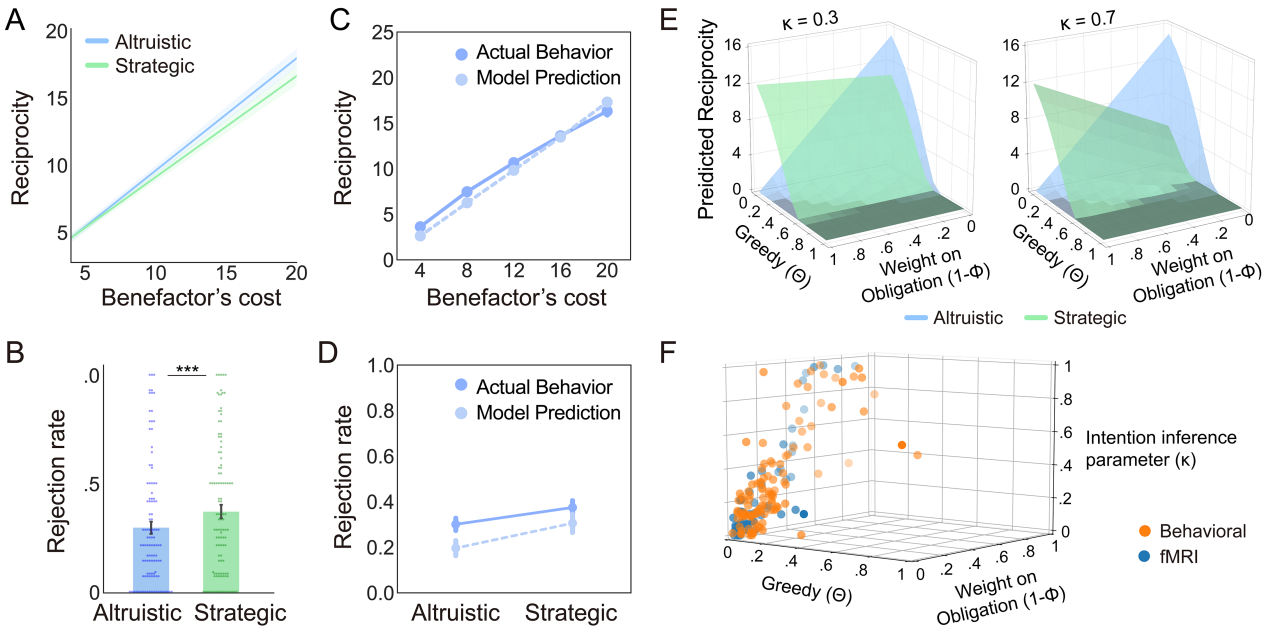

Fig. 4 Computational model of indebtedness. (A) Participants' reciprocity behavior in each trial plotted as function of the benefactor's intention and cost. **(B)** Overall rate of rejecting help in Altruistic and Strategic conditions, *** p < 0.001. Each dot represents the average rejection rate in the corresponding condition for each participant. **(C)** The observed amounts of reciprocity after receiving help and predictions generated by computational model at each level of the benefactor's cost. **(D)** The observed rates of rejecting help and predictions generated by computational model in Altruistic and Strategic conditions. **(E)** Model simulations for predicted reciprocity behavior in Altruistic and Strategic conditions at different parameterizations. **(F)** Best fitting parameter estimates of the computational model of indebtedness for each participant. Error bars represent the standard error of means.

Computational model captures feelings underlying responses to receiving favors

Next we evaluated how well our computational model (Eq. 1) could account for the behavioral data. Since our results above suggested that communal and obligation feelings are derived from the appraisals of perceived care (ω_B) and second-order belief (E_B'') of the benefactor's expectation for repayment respectively, we modeled these two appraisals to index communal and obligation feelings. The parameter κ_B captures the process of inferring intentions representing the degree to which the perceived strategic intention reduced the perceived altruistic intention (see Methods and Supplemental Materials for more details). We found that our model was able to successfully capture the patterns of participants' reciprocity after receiving help (r^2 = 0.81, $p < 0.001$; Fig. 4C) and decisions of whether to accept help (accuracy = 80.00%; Fig. 4D). In addition, each term of our model was able to accurately capture self-reported appraisals of second-order belief of the benefactor's expectation for repayment ($\beta = 0.68 \pm 0.03$, $t = 21.48$, $p < 0.001$; Fig. S4, A-B) and perceived care ($\beta = 0.64 \pm 0.02$, $t = 26.76$, $p < 0.001$; Fig. S4, C-D), which provides further validation that we are accurately modeling the intended psychological processes. In addition, the indebtedness model with both communal and obligation feelings outperformed other plausible models, such as: (a) models that only include a single term, (b) models with separate parameters for each term, (c) a model that assumes participants reciprocate as a function of the cost to the benefactor, and (d) a model that assumes that participants are motivated to minimize inequity in payments (Fehr and Schmidt, 1999) (Table S5 and S6). Furthermore, parameter recovery tests indicated that the parameters of the indebtedness model were identifiable (correlation between true and recovered parameters: reciprocity $r = 0.94 \pm 0.07$, $p < 0.001$; decisions of whether to reject help $r = 0.67 \pm 0.36$, $p < 0.001$; Table S7 and S8). See *SI Results* for detailed results of computational modeling and Table S9 and S10 for descriptive statistics for model parameters.

A simulation of the model across varying combinations of the Θ, Φ and κ parameters reveals diverging predictions of the beneficiaries' response to altruistic and strategic favors (Fig. 4E). Not surprisingly, greedier individuals (higher Θ) are less likely to reciprocate others' favors. However, reciprocity changes as a function of the tradeoff

between communal (Φ) and obligation (1 - Φ) feelings and interacts with the intention inference parameter (κ). As the emphasis on obligation increases, the amount of reciprocity to strategic favors increases whereas that to altruistic favors decreases; this effect is enhanced 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. 4F). 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-Φ) 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.

Communal and obligation feelings are associated with distinct neural processes

Next we explored the neural basis of indebtedness guided by our computational model and behavioral findings. Participants (N = 53) in Study 3 completed the same task as Study 2 while undergoing fMRI scanning, except that they were unable to reject help. We successfully replicated all of the behavioral results observed in Study 2 (Table S4; Fig. S6). We were specifically interested in brain processes during the Outcome period, where participants learned about the benefactor's decision to help. Using a model-based fMRI analytic approach (O'doherty et al., 2007), 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 of: (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''), defined using a linear contrast (Strategic_Lowcost +1, Strategic_Midcost +2, Strategic_Highcost +3, and Altruistic_condition -6) (Chang et al., 2011). We found that trial-by-trial reciprocity behavior correlated with activity in bilateral dorsal lateral prefrontal cortex (dlPFC, peak MNI coordinates: [-45, 5, 29] and [45, 11, 35]), bilateral inferior parietal lobule (IPL, [-54, -40, 53] and [51, -28, 47]), precuneus [6,

-64, 41], and bilateral inferior temporal gyrus (ITG, [-45, -61, -13] and [51, -52, -13]) (Fig. 5A, Table S11). Trial-by-trial communal feelings tracked with activity in the ventromedial prefrontal cortex (vmPFC, [0 33 -22]), anterior insula (aINS, [-24, 11, -16]), precuneus [3, -46, 38], bilateral dlPFC ([-48, 20, -26] and [45, 11, 38]) and bilateral ITG ([-54, -76, -7] and [48, -46, -16]) (Fig. 5B; Tables S11). Linear contrasts of obligation revealed significant activations in dorsomedial prefrontal cortex (dmPFC, [-9, 47, 41]) and left temporo-parietal junction (TPJ, [-57, -61, 26]) (Fig. 5C, Tables S11).

To aid in interpreting these results, we performed meta-analytic decoding (Chang et al., 2013) using Neurosynth (Yarkoni et al., 2011). 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. 5D). Together, these neuroimaging results reveal differing neural bases underlying feelings of communal concern and obligation and support the role of intention inference in the generation of these feelings. The processing of communal feelings was associated with activity in vmPFC, an area in default mode network that has been linked to gratitude (Fox et al., 2015; Yu et al., 2017; Yu et al., 2018), positive social value and kind intention, (Cooper et al., 2010; Ruff and Fehr, 2014a) as well as the insula, which has been previously related to guilt (Chang et al., 2011; Koban et al., 2013; Yu et al., 2014). In contrast, the processing of obligation was associated with the activations of theory of mind network, including dmPFC and TPJ, which is commonly observed when representing other peoples' intentions or strategies (Hampton et al., 2008; Ruff and Fehr, 2014a; Van Overwalle and Baetens, 2009).

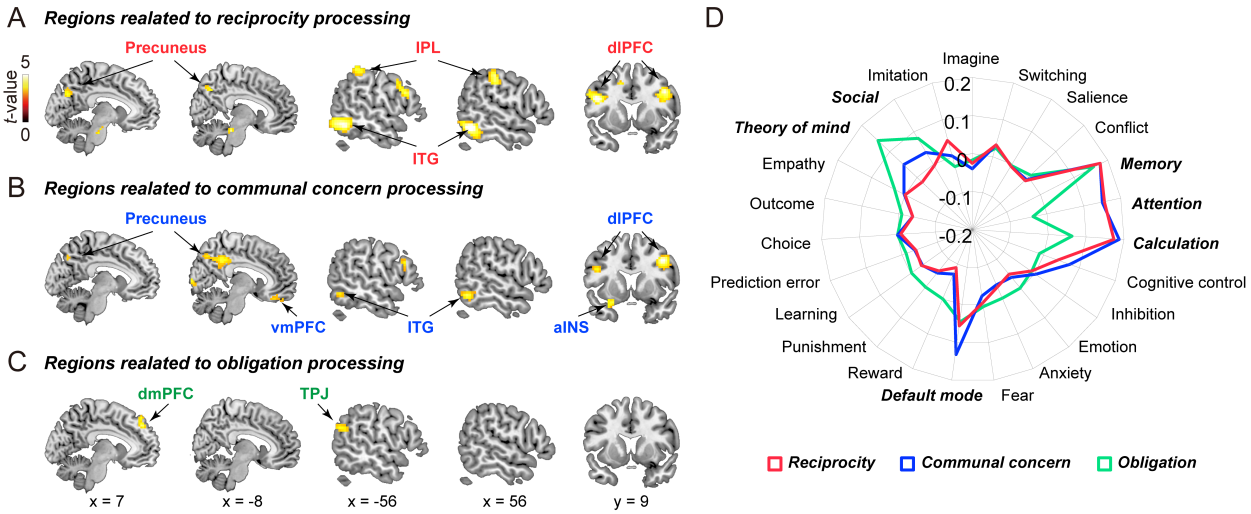

Fig. 5 Neural processes associated with reciprocity, communal concern and obligation. (A) Brain regions responding parametrically to trial-by-trial amounts of reciprocity. **(B)** Brain regions responding parametrically to trial-by-trial communal concern, which depended on the perceived care from the help (ω_B). **(C)** Brain regions identified in the parametric contrast for obligation ($E_B{''}$), the responses of which monotonically increased in the strategic condition relative to the altruistic condition. **(D)** Meta-analytical decoding for the neural correlates of reciprocity, communal concern and obligation, respectively.

Neural utility model of indebtedness predicts reciprocity behavior

Having established that our model of indebtedness was able to accurately capture the psychological processes underlying feelings of communal concern and obligation, we next sought to test whether we could use signals directly from the brain to construct a utility function and predict reciprocity behavior (Fig. 6A). We trained two whole-brain models using principle components regression with 5-fold cross-validation (Chang et al., 2015; Wager et al., 2013; Woo et al., 2017) to predict feelings of communal concern (ω_B) and obligation ($E_B{''}$) using brain activity during the Outcome period of the task 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 ± 0.03, *fisher-z* = 0.20 ± 0.02, *permutation* $p < 0.001$;

obligation pattern: average r = 0.10 ± 0.03, *fisher-z* = 0.09 ± 0.02, *permutation p* = 0.004).

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. 2):

$$U(D_B) = \theta_B * \pi_B + (1-\theta_B) * (\phi_B * \vec{\beta}_{map} \cdot \vec{Communal}_{map} + (1-\phi_B) * \vec{\beta}_{map} \cdot \vec{Obligation}_{map}), \quad \textbf{Eq. 2}$$

where $\vec{\beta}_{map}$ refers to the vector of brain intensities observed during the Outcome phase and $\vec{Communal}_{map}$ and $\vec{Obligation}_{map}$ refer to the multivariate brain models predictive of communal and obligation feelings 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 = 0.10 ± 0.01, *fisher-z* = 0.10 ± 0.01, *permutation p* = 0.013, AIC = 324.04 ± 4.93). The brain-based predictions of the weights on obligation were closely correlated with those estimated by directly fitting the model to behavior, r = 0.88, $p < 0.001$. As a benchmark, this model performed slightly worse than our overall ability to directly predict reciprocity behavior from multivariate patterns of brain activity (Fig. 6A, reciprocity pattern: average r = 0.18 ± 0.03, *fisher-z* = 0.17 ± 0.03, *permutation p* < 0.001, AIC = 321.07 ± 4.81; paired t test for AIC, t_{52} = 5.26, $p < 0.001$), which we take as the upper bound of the neural signal that can predict behavior.

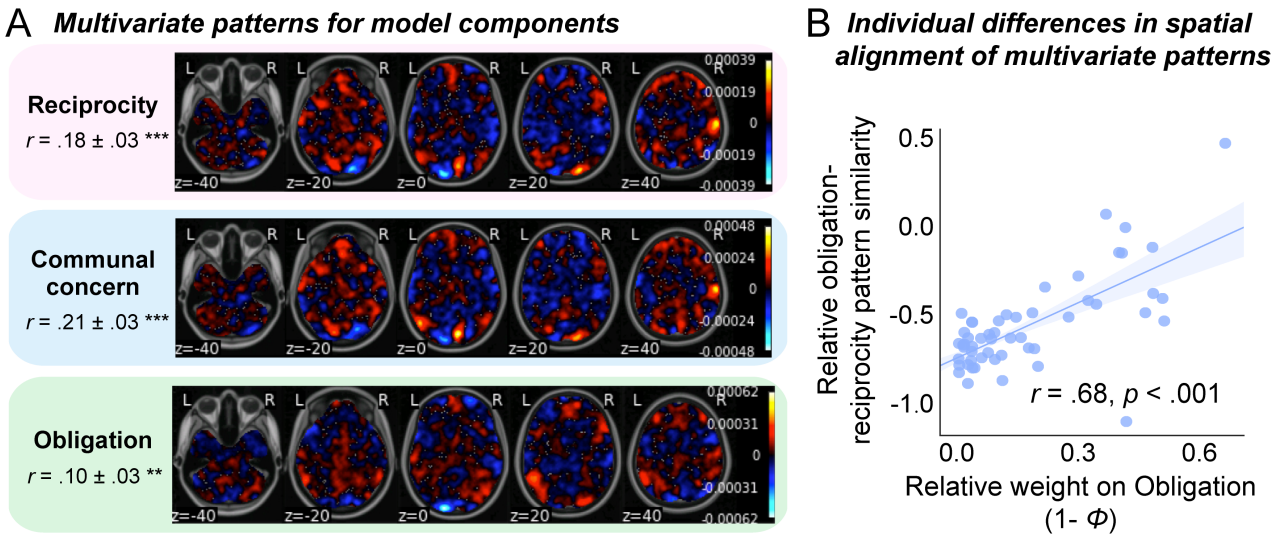

Fig. 6 Neural utility model of indebtedness. (A) Unthresholded multivariate patterns used to predict the amounts of reciprocity, trial-by-trial communal concern (ω_B) and obligation ($E_B{''}$) separately. **(B)** The relationship between the relative weight on obligation (1 - Φ) derived from behavior and a neurally derived metric of how much obligation vs. communal feelings influenced reciprocity behavior (Eq. 3).

Finally, we examined if brain activity reflected individual differences in the degree to which participants were motivated by obligation relative to communal concern in their decisions based on spatial alignment of the multivariate brain patterns (Kriegeskorte et al., 2008). The relative influence of a particular feeling on behavior should be reflected in the spatial similarity of the corresponding brain patterns. For example, if a participant weights obligation more than communal concern, then there should be a corresponding relationship reflected in the spatial similarity of the brain patterns of each construct. We operationalized relative pattern similarity as:

$$relative\ pattern\ similarity = corr(\vec{Obligation}_{map}, \vec{Reciprocity}_{map}) - corr(\vec{Communal}_{map}, \vec{Reciprocity}_{map}) \quad \textbf{Eq. 3}$$

We found strong support for this hypothesis. Participants with higher relative weights on obligation estimated from the computational model of behavior (1 - Φ) also exhibited increased relative similarity between their predictive reciprocity brain

representation and their predictive obligation brain representation (Eq. 3), $r = 0.68$, $p < 0.001$ (Fig. 6B). 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

In this study, we sought to develop and validate a theoretical model of indebtedness across three separate experiments, by combining large-scale experience sampling, behavioral measurements in an interpersonal game, computational modeling, and neuroimaging. These studies provide consistent evidence suggesting that indebtedness is comprised of two distinct components - guilt and the sense of obligation. When participants believe that a benefactor cares for them and has altruistic intentions, they are more likely to feel guilt, which along with gratitude, contributes to feelings of communal concern. Alternatively, when participants believe a benefactor possesses strategic intentions and expects something in return, they are more likely to experience a sense of obligation. Both communal concern and obligation motivate the beneficiary to reciprocate, while obligation is more likely to lead to rejection of help when a benefactor has strategic intentions.

An important contribution of this work is our use of different types of experimental designs to test the predictions of our theory. First, we used an open-ended survey to capture lay intuitions about indebtedness based on past experiences from a relatively large sample. Overall, we find strong support that the feeling of indebtedness resulting from receiving help from others can be attributed to two distinct emotions – 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 words pertaining to burden and negative bodily states

load on a separate topic. Second, we used a laboratory task designed to elicit indebtedness in the context of a social 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 (Ellsworth and Scherer, 2003; Frijda, 1993; Frijda et al., 1989; Lazarus and Smith, 1988; Scherer, 1999; Smith and Ellsworth, 1985), these shifts in appraisals influenced participants' subjective emotions and ultimately their behavior. Altruistic intentions lead to increased guilt and gratitude, while strategic intentions increase feelings of obligation. All three feelings were associated with increased monetary reciprocation back to the benefactor after receiving help. However, only obligation increased the rejection of help when that option was available to the participant.

One of the most notable contributions of this work is the development and validation of a computational model of indebtedness. The majority of research on emotions relies on self-reported subjective feelings (Lench et al., 2011; Lindquist et al., 2013), which has a number of limitations, such as its dependence on participants' ability to introspect (Larsen and Fredrickson, 1999; Nisbett and Wilson, 1977). Formalizing emotions using computational models is critical to advancing theory, charactering their impact on behavior, and identifying neural and physiological substrates (Chang and Jolly, 2018; Chang and Smith, 2015; Jolly and Chang). Our model provides a demonstration of how emotion appraisal theory (Ellsworth and Scherer, 2003; Frijda, 1993; Frijda et al., 1989; Lazarus and Smith, 1988; Scherer, 1999; Smith and Ellsworth, 1985) can be integrated with psychological game theory (Dufwenberg and Kirchsteiger, 2004; Geanakoplos et al., 1989) to predict behavior (Chang and Smith, 2015). We model emotions as arising from appraisals about perceived care and beliefs

about the beneficiary's expectations, which both ultimately increase the likelihood of the benefactor selecting actions to reciprocate the favor. This model contributes to a growing family of game theoretic models of social emotions such as guilt (Battigalli and Dufwenberg, 2009; Chang et al., 2011), gratitude (Khalmetski et al., 2015), and anger (Battigalli et al., 2015; Chang and Sanfey, 2013).

We provide a rigorous validation of our indebtedness model across behaviors in the task, subjective experiences, and neural correlates. First, our model does remarkably well at predicting participants' reciprocity behavior. It also captures our theoretical predictions that participants would be more likely to reject help when they perceived the benefactor to have strategic intentions than when they perceived the benefactor to have altruistic intentions. Second, the parameters of our model were able to accurately capture self-reported appraisals of second-order belief of the benefactor's expectation for repayment and perceived care, which validates our model from subjective experiences. Third, our brain imaging analyses provide an additional level of validation 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 (Chang et al., 2011; Koban et al., 2013; Krajbich et al., 2009; Yu et al., 2014) and gratitude (Fox et al., 2015; Yu et al., 2017; Yu et al., 2018), our model representation of feelings 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 (Hampton et al., 2008; Ruff and Fehr, 2014a; Van Overwalle and Baetens, 2009). These brain results are particularly noteworthy as we are unaware of any prior work that has probed the neural basis of indebtedness. Fourth, our computational modeling reveals that individuals who are more sensitive to obligation tend to reciprocate more to strategic favors than to altruistic favors,

555 indicating a greater susceptibility to hidden costs when receiving strategic favors (Bal,
556 2005; Fehr and Gächter, 2000; Malmendier and Schmidt, 2012). This quantitative
557 measure might be more sensitive than self-report measures and could be used as an
558 individual difference measure in future work.
559
560 We provide an even stronger test of our ability to characterize the neural processes
561 associated with indebtedness by deriving a "neural utility" model. Previous work has
562 demonstrated that it is possible to build brain models of preferences that can predict
563 behaviors (Knutson et al., 2007; Smith et al., 2014). In this series of analyses, we
564 trained multivoxel patterns of brain activity to predict participants' communal and
565 obligation feelings. We then used these brain-derived predictions of communal and
566 obligation feelings to predict how much money they ultimately reciprocated to the
567 beneficiary. Remarkably, we found that this neural utility model of indebtedness was
568 able to predict individual decisions entirely from brain activity and almost as well as a
569 control brain-model that was designed to directly predict reciprocity behavior. In
570 addition, we find that the more the neural activity during reciprocity resembled brain
571 patterns predictive of obligation compared with communal concern, the more our
572 computational model attributed obligation to behavior, providing a direct link
573 between these distinct feelings and patterns of brain activity.
574
575 Our study has several potential limitations, which are important to acknowledge. First,
576 though we directly and conceptually replicate our key findings across multiple
577 samples, all of our experiments recruit experimental samples from a Chinese
578 population. It is possible that there exist cultural differences in the experience of
579 indebtedness, which may not generalize to other parts of the world. For example,
580 compared with Westerners who commonly express gratitude when receiving
581 benevolent help, Japanese participants often respond with "Thank you" or "I am
582 sorry" (Benedict, 1946; Kotani, 2002). However, we think this is unlikely as both

583 guilt toward favor-doers (e.g., the organ transplant patients' guilt) (Achille et al., 2006;
584 Annema et al., 2013; Látos et al., 2016; Shemesh et al., 2017) and the sense of
585 obligation to repay (Watkins et al., 2006) have been consistently observed in various
586 Western populations. Second, our laboratory-based task was designed to test a key
587 assumption in our theory that individuals trade-off feelings of communal concern and
588 obligation when responding to receiving help. Although we found compelling
589 evidence distinguishing between feelings of communal concern and obligation, our
590 current task was unable to distinguish between guilt and gratitude. Theoretically, we
591 predicted that both guilt and gratitude arise from altruistic favors and are part of a
592 broader encompassing construct of communal concern (Baumeister et al., 1994; Le et
593 al., 2018). This construct is related to communal relationships described by
594 psychologists, sociologists, and anthropologists (Algoe, 2012; Algoe et al., 2008;
595 Clark and Mills, 1993; Clark and Mills, 2012; Elfers and Hlava, 2016; McCullough et
596 al., 2001; Nowak and Sigmund, 2005), while obligation, in contrast, corresponds
597 more to transactional exchange relationships (Greenberg, 1980; Greenberg and
598 Westcott, 1983; Watkins et al., 2006). Future work might design tasks that can better
599 differentiate between gratitude and guilt to explore whether these two emotions of
600 communal concern have shared or differential neurocognitive mechanisms (Chang et
601 al., 2011; Fox et al., 2015; Koban et al., 2013; Krajbich et al., 2009; Yu et al., 2017;
602 Yu et al., 2018; Yu et al., 2014).
603
604 Gift-giving, favor-exchanges, and providing assistance are behaviors reflective of the
605 relationship between individuals or groups. On the one hand, while altruistic favors
606 often engender reciprocity and gratitude, they can also elicit guilt in a recipient who
607 feels burdensome to a benefactor. On the other hand, favors in transactive
608 relationships in which reciprocity is expected, can engender a feeling of obligation for
609 a recipient. Previous studies have independently investigated these two components of
610 indebtedness (Benedict, 1946; Greenberg, 1980; Greenberg and Shapiro, 1971;

611 Greenberg and Westcott, 1983; Kotani, 2002; Naito and Washizu, 2015; Tsang, 2006;
612 Watkins et al., 2006). Here, by developing a comprehensive and systematic model of
613 indebtedness, our work emphasizes how appraisals about the intentions behind a favor
614 are critical to the generation of these distinct emotions, which in turn motivates how
615 willing individuals are to accept or reject help and ultimately reciprocate the favor.
616 Our model provides not only a general framework that integrates previous
617 independent findings, but also a theoretical bases for future investigations. For
618 example, although we test our theory primarily in an interpersonal task on favors,
619 which involve unsolicited help between strangers to reduce pain, we believe these
620 processes will generalize more broadly to receiving help in most interpersonal
621 contexts. This work highlights the importance of considering the psychological and
622 neural mechanisms underlying the hidden costs of receiving help (Bal, 2005; Fehr and
623 Gächter, 2000; Malmendier and Schmidt, 2012).

Methods

Participants. In total, the data of 1,619 (812 females, 18.9 ± 2.0 (SD) years), 51 (33 females, 19.9 ± 1.6 years), 57 (45 females, 20.1 ± 1.8 years), and 53 (29 females, 20.9 ± 2.3 years) healthy graduate and undergraduate students were included for Study 1 (experience sampling), Studies 2a and 2b (behavioral studies) and Study 3 (fMRI study), respectively. In addition, 80 participants (45 females, 22.6 ± 2.58 years) were recruited for the word classification task to extract emotion-related words in the definition of indebtedness. All of the experiments were carried out in accordance with the Declaration of Helsinki and were approved by the Ethics Committee of the School of Psychological and Cognitive Sciences, Peking University. Informed written consent was obtained from each participant before each experiment. Consent to publish was obtained for each image in Fig. 2C.

Topic Modeling. For the self-reported definition of indebtedness analysis, we used the "Wordcloud" (https://amueller.github.io/word_cloud/index.html) and "Jieba" (https://github.com/fxsjy/jieba) packages to conduct text segmentation. We excluded stop words using Wordcloud dataset and extracted the 100 words with the highest weight/frequency in the definitions of indebtedness using Term Frequency-Inverse Document Frequency (TF-IDF) (Neto et al., 2000; Salton and Buckley, 1988). These 100 words were then classified by an independent sample of participants (N = 80) into levels of appraisal, emotion, behavior, person and other. Because Chinese retains its own characters of various structures, synonym combinations were implemented before topic modeling (Liu, 2016). We conducted Latent Dirichlet Allocation (LDA) based topic modeling on only the emotional words of indebtedness using collapsed Gibbs sampling implemented in the lda package (https://lda.readthedocs.io/en/latest/) (Blei et al., 2003). We then selected the model with the best model fit using topic numbers ranging from 2 to 10, and found that the two-topic solution performed the best.

Modeling of each utility term. Each item in Eq. 1 (π_B, $U_{Communal}$ *and* $U_{Obligation}$) was defined according to the corresponding context of decision-making. We modeled the utility of self-interest (π_B) as Eq. 4. For each amount of reciprocity (D_B), the self-interest was defined as the percentage of money the participant receives from the total endowment (γ_B). For the decisions of whether to accept or reject help, the self-interest from accepting help was defined as the percentage of pain reduction from the total amount of the maximum pain reduction, 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).

$$\pi_B = \begin{cases} \frac{\gamma_B - D_B}{\gamma_B} & \text{Reciprocity} \\ \frac{D_A * \mu}{\max(D_A * \mu)} & \text{Accept/Reject help} \end{cases} \qquad \textbf{Eq. 4}$$

Participant's second-order beliefs of how much the benefactor expected in each trial were determined by the benefactor's intention and benefactor's cost (D_A) (Eq. 5). In the altruistic condition, participants knew that the benefactor did not expect them to reciprocate, so we fixed the second-order belief as zero (E_B''). However, in the strategic 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.

$$E_B'' = \begin{cases} 0 & \text{Altruistic condition} \\ D_A & \text{Strategic condition} \end{cases} \qquad \textbf{Eq. 5}$$

The participant's perceived care (ω_B) in each trial was defined as a function of the benefactor's cost and second-order belief (Eq. 6). Specifically, we assumed that the perceived care from the help increased as a linear function of how much the benefactor spent (D_A) from his/her endowment (γ_A); however, this effect was reduced by the second-order belief of the benefactor's expectation for repayment (E_B''). Here, the parameter kappa (κ) is a free parameter ranging from 0 and 1 that represents the extent to which the benefactor's expectation for repayment reduced the participant's perceived care.

$$\omega_B = \frac{D_A - \kappa_B * E_B''}{\gamma_A} \quad \textbf{Eq. 6}$$

We defined $U_{Communal}$ *and* $U_{Obligation}$ as functions of ω_B and E_B'' respectively, but the formulations were slightly different for predicting reciprocity and rejection decisions (Eq. 7 and Eq. 8).

$$U_{Communal} = \begin{cases} -(\frac{\omega_B * \gamma_B - D_B}{\gamma_B})^2 & \text{Reciprocity} \\ \omega_B & \text{Accept/Reject help} \end{cases} \quad \textbf{Eq. 7}$$

$$U_{Obligation} = \begin{cases} -(\frac{E_B'' - D_B}{\gamma_B})^2 & \text{Reciprocity} \\ -\frac{E_B''}{\gamma_B} & \text{Accept/Reject help} \end{cases} \quad \textbf{Eq. 8}$$

Specifically, for reciprocity, $U_{Communal}$ *and* $U_{Obligation}$ were defined as functions of ω_B and E_B''. Participants maximized utility of communal concern ($U_{Communal}$) by minimizing the difference between the benefactor's reciprocity (D_B) and the amount of money the participant was willing to repay the benefactor's kindness, which depended on the perceived care (ω_B) and the endowment size (γ_B). In contrast, participants maximized utility of obligation ($U_{Obligation}$) 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''). For decisions of whether to reject help, $U_{Communal}$ and $U_{Obligation}$ were defined as the linear functions of ω_B and E_B''.

We modeled the utility of reciprocating $U(D_B)$ as:

$$U(D_B) = \theta_B * \frac{\gamma_B - D_B}{\gamma_B} - (1 - \theta_B) * (\phi_B * (\frac{\omega_B * \gamma_B - D_B}{\gamma_B})^2 + (1 - \phi_B) * (\frac{E_B'' - D_B}{\gamma_B})^2) \quad \textbf{Eq.9}$$

Where Φ is defined as a free parameter between 0 and 1, which captures the trade-off between feelings of communal concern and obligation. We estimated the model parameters for Eq. 9 by minimizing the sum of squared error of the percentiles. To

713 minimize the possibility of the algorithm getting stuck in a local minimum, we used
714 the best fitting model over 1000 random starting values.
715

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

716 **Eq. 10**
717

718 In contrast to reciprocity, decisions of whether to accept or reject help might be more
719 complex. The sense of obligation may motivate rejecting the help to avoid being in
720 the benefactor's debt (Greenberg, 1980; Greenberg and Shapiro, 1971; Greenberg and
721 Westcott, 1983). For communal concern, while gratitude may motivate one to accept
722 help to build a communal relationship (Algoe, 2012; Algoe et al., 2008), guilt may
723 motivate one to reject to avoid burdening a benefactor (Battigalli and Dufwenberg,
724 2009; Chang et al., 2011). We model the utility of accepting help as:
725

$$U(Accept) - U(Reject) = \theta_B * \frac{D_A * \mu}{\max(D_A * \mu)} + (1 - \theta_B) * (\phi_B * \omega_B - (1 - |\phi_B|) * \frac{E''_B}{\gamma_B})$$

726
727 **Eq. 11**
728

729 Where Φ lies on the interval of [-1, 1]. Specifically, $\Phi < 0$ indicates that the
730 communal concern motives the participants to reject the help, while $\Phi > 0$ indicates
731 that the communal concern motives the participants to accept the help. The individual
732 weight on obligation is captured by *1*- $|\Phi|$, which ranges from 0 to 1. We estimated
733 the parameters for Eq. 10, by maximizing the log-likelihood.
734

$$LLE = -\sum_{t=1}^{n} log(P(D_B(t)))$$

735 **Eq. 12**
736

737 We conducted parameter recovery analyses to ensure that our model was robustly
738 identifiable (Fareri et al., 2015). To this end, we simulated data for each participant
739 using our models and the data from each trial of the experiment and compared how
740 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.

FMRI Data Acquisition and Analysis. Images were acquired using a 3-T Prisma Siemens scanner (Siemens AG, Erlangen, Germany). We used standard preprocessing in SPM12 (Wellcome Trust Centre for Neuroimaging) and estimated three general linear models for each participant that focused on the neural responses during the Outcome phase at which participants saw the benefactor's help decisions. As our model hypothesizes that feelings of communal concern and obligation arise from the perceived care from the help (ω_B) the second-order belief of the benefactor's expectation for repayment (E_B'') respectively, we used ω_B and E_B'' in the computational model as indices for communal and obligation feelings and conducted parametric analyses. Brain responses to ω_B and E_B'' reflected how much information in neural patterns was associated with each type of feeling in the brain. An alternative approach is to use the $U_{Communal}$ and the $U_{Obligation}$ from our computation model as parametric modulators when estimating brain responses. However, in our model, $U_{Communal}$ and the $U_{Obligation}$ were defined as negative quadratic functions, the maximum values of which were zero. As we predicted and observed, participants behaved to maximize their $U_{Obligation}$ by minimizing the differences between the amount of reciprocity and E_B'', and to maximize their $U_{Communal}$ by minimizing the differences between the amount of reciprocity and ω_B. Therefore, in a large proportion of trials, the $U_{Obligation}$ and $U_{Communal}$ were 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. In contrast, ω_B and E_B'' better captured the inferences that comprised participants' feelings and

were more suitable for testing our hypotheses about brain responses. For whole brain analyses, all results were corrected for multiple comparisons using the threshold of voxel-level $p < 0.001$ (uncorrected) combined with cluster-level threshold $p < 0.05$ (FWE-corrected). This threshold provides an acceptable family error control (Eklund et al., 2016; Flandin and Friston, 2017). 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 (Yarkoni et al., 2011) (http://neurosynth.org). This allowed us to quantitatively evaluate the spatial similarity (Chang et al., 2013) 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) (Barrett and Satpute, 2013), social cognition (Empathy, Theory of mind, Social, and Imitation) (Adolphs, 2009) and decision-making (Reward, Punishment, Learning, Prediction error, Choice, and Outcome) (Ruff and Fehr, 2014b).

Neural Utility Model of Indebtedness. We applied multivariate pattern analysis (MVPA) (Haynes and Rees, 2006) and trained two whole-brain models to predict the communal concern (ω_B) and obligation (E_B'') terms in our behavioral model separately for each participant using principle components regression with 5-fold cross-validation (Chang et al., 2015; Wager et al., 2013; Woo et al., 2017), which was carried out in Python 3.6.8 using the NLTools package (Chang et al., 2018) version 0.3.14 (http://github.com/cosanlab/nltools). We used whole-brain single-trial beta maps of the Outcome period for each participant to separately predict ω_B and E_B''. 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 predict the corresponding term. 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. 2). We trained a whole-brain model to predict trial-by-trial reciprocity for each participant as a benchmark comparison. Finally, for each participant, we estimated the whole-brain spatial similarity (Kriegeskorte et al., 2008) between the two prediction maps of communal and obligation feelings and the reciprocity prediction map. The relative obligation-reciprocity similarity was defined as Eq. 3 and was used to examine whether this neural alignment could predict individual relative weight on obligation during reciprocity.

All the statistical tests in the current study are two-tailed tests. A detailed description of methods including participants, procedures, computational modeling, and fMRI data analyses are given in *SI Appendix*.

Data availability

All data needed to evaluate the conclusions in the current study are present in the paper and the *SI Appendix*. Original materials are available from the corresponding author upon reasonable request.

Code availability

The codes used in the current study are available from the corresponding author upon reasonable request.

Acknowledgements

We thank Dr. Christian C. Ruff for his comments and suggestions on this article, Ms. Yunyan Duan's for her advice in topic modeling, and Ms. Zhewen He for the preparation of the manuscript. This work was supported by National Basic Research Program of China (973 Program: 2015CB856400), National Natural Science Foundation of China (91232708, 31170972, 31630034, 31900798, 71942001), China Postdoctoral Science Foundation (2019M650008), the National Science Foundation of USA (CAREER 1848370), and the National Institute of Health (R01MH116026). Thanks are also due to Graduate School of Peking University to support Dr. Gao for visiting Dartmouth College.

Reference

Achille, M.A., Ouellette, A., Fournier, S., Vachon, M., and Hébert, M.J. (2006). Impact of stress, distress and feelings of indebtedness on adherence to immunosuppressants following kidney transplantation. Clin Transplant 20, 301-306.

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

Akerlof, G.A. (1982). Labor contracts as partial gift exchange. The quarterly journal of economics 97, 543-569.

Algoe, S.B. (2012). Find, remind, and bind: The functions of gratitude in everyday relationships. Social and Personality Psychology Compass 6, 455-469.

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

Annema, C., Roodbol, P.F., Stewart, R.E., and Ranchor, A.V. (2013). Validation of the Dutch version of the transplant effects questionnaire in liver transplant recipients. Res Nurs Health 36, 203-215.

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

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

Battigalli, P., Corrao, R., and Dufwenberg, M. (2019). Incorporating belief-dependent motivation in games. Journal of Economic Behavior & Organization.

Battigalli, P., and Dufwenberg, M. (2009). Dynamic psychological games. Journal of Economic Theory 144, 1-35.

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

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

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

Blei, D.M., and Lafferty, J.D. (2006). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (ACM), pp. 113-120.

Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent dirichlet allocation. Journal of machine Learning research 3, 993-1022.

Carmichael, H.L., and MacLeod, W.B. (1997). Gift giving and the evolution of cooperation. International Economic Review, 485-509.

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

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

Chang, L.J., Jolly, E., Cheong, J.H., Burnashev, A., and Chen, A. (2018). cosanlab/nltools: 0.3.11. (Zenodo).

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

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

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

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

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

Clark, M.S., and Mills, J.R. (2012). A theory of communal (and exchange)

relationships. In Handbook of theories of social psychology, Vol 2 (Thousand Oaks, CA: Sage Publications Ltd), pp. 232-250.

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

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

Eklund, A., Nichols, T.E., and Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. Proc Natl Acad Sci U S A 113, 7900-7905.

Elfers, J., and Hlava, P. (2016). The Spectrum of Gratitude Experience (Springer).

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

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

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

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

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

Flandin, G., and Friston, K.J. (2017). Analysis of family‐wise error rates in statistical parametric mapping using random field theory. Hum Brain Mapp hbm.23839.

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

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

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

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

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

Greenberg, M.S. (1980). A theory of indebtedness. In Social exchange (Springer), pp. 3-26.

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

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

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

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

Jolly, E., and Chang, L.J. The Flatland Fallacy: Moving Beyond Low-Dimensional Thinking. Top Cogn Sci.

Khalmetski, K., Ockenfels, A., and Werner, P. (2015). Surprising gifts: Theory and laboratory evidence. Journal of Economic Theory 159, 163-208.

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

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

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

Kotani, M. (2002). Expressing gratitude and indebtedness: Japanese speakers' use of

"I'm sorry" in English conversation. Research on Language and Social Interaction 35, 39-72.

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

Kriegeskorte, N., Mur, M., and Bandettini, P.A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience 2, 4.

Larsen, R.J., and Fredrickson, B.L. (1999). Measurement issues in emotion research. In Well-being: The foundations of hedonic psychology (New York, NY, US: Russell Sage Foundation), pp. 40-60.

Látos, M., Lázár, G., Horváth, Z., Wittman, V., Szederkényi, E., Hódi, Z., Szenohradszky, P., and Csabai, M. (2016). Psychological rejection of the transplanted organ and graft dysfunction in kidney transplant patients. Transplant Research and Risk Management 8, 15-24.

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

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

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

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

Liu, Q. (2016). A novel Chinese text topic extraction method based on LDA. In

International Conference on Computer Science & Network Technology.

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

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

Naito, T., and Washizu, N. (2015). Note on cultural universals and variations of gratitude from an East Asian point of view. The Journal of Behavioral Science 10, 1-8.

Neilson, W.S. (1999). The economics of favors. Journal of economic behavior & organization 39, 387-397.

Neto, J.L., Santos, A.D., Kaestner, C.A., Alexandre, N., and Santos, D. (2000). Document clustering and text summarization.

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

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

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

Rabin, M. (1993). Incorporating fairness into game theory and economics. The American economic review, 1281-1302.

Regan, D.T. (1971). Effects of a favor and liking on compliance. J Exp Soc Psychol 7, 627-639.

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

Ruff, C.C., and Fehr, E. (2014a). The neurobiology of rewards and values in social decision making. Nat Rev Neurosci 15, 549.

Ruff, C.C., and Fehr, E. (2014b). The neurobiology of rewards and values in social

decision making. Nat Rev Neurosci 15, 549-562.

Salton, G., and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management 24, 513-523.

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

Shemesh, Y., Peles‐Bortz, A., Peled, Y., HarZahav, Y., Lavee, J., Freimark, D., and Melnikov, S. (2017). Feelings of indebtedness and guilt toward donor and immunosuppressive medication adherence among heart transplant (HT x) patients, as assessed in a cross‐sectional study with the Basel Assessment of Adherence to Immunosuppressive Medications Scale (BAASIS). Clin Transplant 31, e13053.

Sherry Jr, J.F. (1983). Gift giving in anthropological perspective. Journal of consumer research 10, 157-168.

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

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

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

Trivers, R.L. (1971). The evolution of reciprocal altruism. The Quarterly review of biology 46, 35-57.

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

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

Wager, T.D., Atlas, L.Y., Lindquist, M.A., Roy, M., Woo, C.-W., and Kross, E. (2013). An fMRI-based neurologic signature of physical pain. N Engl J Med 368, 1388-1397.

Watkins, P.C., Scheer, J., Ovnicek, M., and Kolts, R. (2006). The debt of gratitude:

Dissociating gratitude and indebtedness. Cognition Emotion 20, 217-241.

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

Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., and Wager, T.D. (2011). Large-scale automated synthesis of human functional neuroimaging data. Nat Meth 8, 665.

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

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

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