From cognition to compensation: Neurocomputational mechanisms of guilt-driven and shame-driven altruistic behavior

Ruida Zhu[1], Huanqing Wang[4], Chunliang Feng[5], Linyuan Yin[1],
Ran Zhang[2, 3], Yi Zeng[3, 6, 7, 8, 9], Chao Liu[2, 3, *]

[1]Department of Psychology, Sun Yat-sen University, Guangzhou, China
[2]State Key Laboratory of Cognitive Neuroscience and Learning, and IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China
[3]Beijing Key Laboratory of Safe AI and Superalignment, China
[4]Department of Psychology, The Ohio State University, Columbus, United States
[5]Key Laboratory of Brain, Cognition and Education Sciences, Ministry of Education, School of Psychology, Center for Studies of Psychological Application, Guangdong Key Laboratory of Mental Health and Cognitive Science, South China Normal University, Guangzhou, China
[6]Beijing Institute of AI Safety and Governance, China
[7]Brain-inspired Cognitive AI Lab, Institute of Automation, Chinese Academy of Sciences, China
[8]University of Chinese Academy of Sciences, China
[9]Long-term AI, China.

[*]Correspondence to: Chao Liu (liuchao@bnu.edu.cn)

Author contributions: Ruida Zhu and Chao Liu designed the experiment. Huanqing Wang implemented the study design and collected the data. Ruida Zhu carried out the analyses. Ruida Zhu and Chao Liu wrote the paper. All authors provided critical revisions and approved the final paper for submission.

Data availability: The behavioral data and a part of fMRI data are available on Open Science Framework (https://osf.io/sve7h/?view_only=10ee55af054d46a9907a827d3f042d8f). Raw fMRI data are available from the corresponding authors upon request due to privacy concern.

Code availability: The code used in the current study are available on Open Science Framework (https://osf.io/sve7h/?view_only=10ee55af054d46a9907a827d3f042d8f).

Competing interests: The authors declare no competing interests

Funding: This project has received funding from the Scientific and Technological Innovation (STI) 2030-Major Projects (2021ZD0200500 to Chao Liu), the National Natural Science Foundation of China (32200884 to Ruida Zhu; 32441109, 32271092, 32130045 to Chao Liu), the Open Research Fund of the State Key Laboratory of Cognitive Neuroscience and Learning (CNLYB2404 to Ruida Zhu and Chao Liu), the Start-up Project for Support of Young Doctors (SL2023A04J00351 to Ruida Zhu), the Fundamental Research Funds for the Central Universities (Sun Yat-sen University; 2024qntd90 to Ruida Zhu), the Beijing Major Science and Technology Project under Contract No. Z241100001324005 (to Chao Liu), and the Opening Project of the State Key Laboratory of General Artificial Intelligence (SKLAGI20240P06 to Chao Liu).

Abstract: Guilt and shame are key moral emotions that influence mental health and regulate social behavior. Although prior research has examined the psychological and neural correlates of these emotions, the cognitive antecedents that trigger them, as well as their transformation into social behavior, remain insufficiently understood. In this study, we developed a novel task to investigate how two crucial cognitive antecedents, harm and responsibility, elicit guilt and shame, and how these emotions subsequently drive compensatory behavior, by combining functional magnetic resonance imaging (fMRI) with computational modeling. Behaviorally, we found that harm had a stronger impact on guilt than on shame, whereas responsibility had a stronger impact on shame

than guilt, which supports the functionalist theory of emotion. Moreover, compared to shame, guilt exerted a greater effect on compensation. Computational modeling results indicated that the integration of harm and responsibility by individuals is consistent with the phenomenon of responsibility diffusion. The fMRI results revealed that brain regions associated with inequity representation (posterior insula) and value computation (striatum) encode this integrated measure. Furthermore, individual differences in responsibility-driven shame sensitivity were associated with activity in theory-of-mind regions (temporoparietal junction and superior temporal sulcus). Guilt- and shame-driven compensatory behavior recruited distinct neural substrates, with shame-driven compensatory sensitivity being more strongly linked to activity in the lateral prefrontal cortex, a region implicated in cognitive control. Our findings provide computational, algorithmic, and neural accounts of guilt and shame.

Keywords: guilt; shame; compensation; computational modeling; functional magnetic resonance imaging; functionalist theory

Introduction

Guilt and shame, two prominent moral emotions, underpin a wide array of social behaviors and phenomena, including norm compliance, cooperation, face-saving strategies, suicides, and even large-scale conflicts like wars (Haidt, 2003; Sznycer et al., 2021). Although guilt and shame often coexist after moral violation and serve to inhibit further transgressions (Eisenberg, 2000; Lewis, 1971), they differ in their associations with psychological and behavioral issues (Tangney & Dearing, 2003). For instance, shame is commonly associated with higher levels of anxiety, depression, stress, eating disorders, and aggression, whereas guilt is typically unrelated or negatively related to these issues (Blythin et al., 2020; Caldwell et al., 2021; Cândea & Szentagotai-Tătar, 2018; Elison et al., 2014; Kim et al., 2011; Schuster et al., 2021).

Recognizing their distinctions, a growing body of research in psychology and neuroscience has sought to elucidate the psychological and neural mechanisms differentiating guilt and shame, providing insights into how these emotions are processed and regulated. Psychological studies have revealed that guilt involves more concerns on one's actions on others (i.e., negative behavioral impacts), whereas shame involves more concerns on one's self-image (i.e., negative self-evaluations) (Lewis, 1971; Tangney et al., 2007; Tangney & Dearing, 2003). Echoing and complementing these findings, neuroscience research showed that neural activity in brain regions related to other-oriented theory-of-mind processing (temporoparietal junction (TPJ) and dorsomedial prefrontal cortex (DMPFC)), self-oriented self-referential processing (anterior cingulate cortex (ACC) and DMPFC), and cognitive control (lateral prefrontal cortex (LPFC)) can distinguish guilt and shame (Piretti et al., 2023; Zhu, Feng, et al., 2019).

Although the psychological processes and neural activities related to guilt and shame experience are relatively well-documented, the cognitive antecedents of these emotions and their neural representation remain insufficiently understood. Existing research has identified harm (i.e., severity of harm) and responsibility (i.e., responsibility for harm) as cognitive antecedents influencing guilt (Abrams & Doosje, 2011; Berndsen et al., 2004; Čehajić-Clancy et al., 2011; Gao et al., 2021; Iyer et al., 2007; Koban et al., 2013; Li et al., 2020; Tangney, 1991; Yu et al., 2014) and shame (Iyer et al., 2007; Koban et al., 2013; Tangney, 1991). However, it remains unclear whether these factors differ in the strength of their influence on guilt and shame. Functionalist theories propose that guilt functions to minimize and repair undue harm on valued others, thereby addressing the adaptive problem of insufficiently valuing others—a behavior that can indirectly harms oneself (e.g., harming one's partner can disrupt cooperation, ultimately damaging one's own interests) (Sznycer, 2019). In contrast, shame functions to prevent and mitigate the threat of being devalued by others, thus addressing the adaptive problem of reputational damage to the self (e.g., being perceived as deficient in competence or morality by others) (Landers, Sznycer, & Al-Shawaf, 2024; Sznycer, 2019; Sznycer et

al., 2021). If the functionalist theories are correct, the intensity of guilt and shame should more closely track the information that reflects the adaptive problems they are meant to address (e.g., Sznycer et al., 2016). Accordingly, severity of harm—a factor reflecting inadequate valuation of others—is expected to have a stronger impact on guilt than on shame, whereas responsibility for harm—a factor tied to devaluation by others, particularly in open and transparent social contexts—is anticipated to exert a greater influence on shame than guilt (see the Supplementary Information for illustrative examples).

Previous neuroscience studies have examined the neural response to harm or responsibility. Greater harm inflicted on others is associated with stronger activation in the DMPFC, TPJ, and insula (Crockett et al., 2017; Koban et al., 2013). Greater responsibility for harm is linked to stronger activation in the anterior middle cingulate cortex (aMCC) (Li et al., 2020; Yu et al., 2014). Despite these findings, it remains elusive how human brain integrate harm and responsibility when both factors are simultaneously present and dynamically vary. One possibility is that distinct brain regions separately encode harm and responsibility, with their signals converging in other regions responsible for integrating these two factors (e.g., Hu et al., 2017; Yu et al., 2018). Alternatively, some brain regions may directly encode an integrated representation of harm and responsibility, potentially through computational processes such as their product or quotient (e.g., Gray et al., 2002).

According to appraisal theory, emotions are not directly elicited by cognitive antecedents (e.g., harm and responsibility) but arise from the appraisal processes applied to these antecedents (Lazarus & Smith, 1988; Moors et al., 2013). Since appraisal processes vary across individuals, the same stimuli can evoke emotions of distinct intensities and even distinct types across different individuals (Ellsworth, 2013; Moors et al., 2013). Individual differences in emotional experiences, such as guilt and shame, may originate from the variations in neural responses to their cognitive antecedents (Hamann & Canli, 2004; Morawetz & Basten, 2024). However, the neural

substrates that determine the extent to which harm and responsibility are transformed into guilt and shame remain largely unknown.

Shifting the focus from the origins of guilt and shame to their consequences, existing evidence on the distinct associations between these emotions and behavior remains incomplete. Guilt is widely recognized as a powerful motivator of altruistic behaviors, such as offering apologies and making amends (Graton & Ric, 2017; Howell et al., 2012; Ketelaar & Tung Au, 2003; Zhu et al., 2017). In comparison, shame appears to have a weaker promotive effect on altruistic behaviors compared to guilt (de Hooge et al., 2007, 2008; Declerck et al., 2014) and is commonly associated with non-cooperative and antisocial behaviors, including hiding, evasion, self-improvement, externalizing blame, and aggression (de Hooge et al., 2010; Gausel & Leach, 2011; Landers, Sznycer, & Al-Shawaf, 2024; Tangney et al., 1996; Zhu, Xu, et al., 2019). Most studies examined only guilt or only shame, restricting direct comparisons of their behavioral effects. The few studies that have directly compared these emotions relied on participants' recollections of past personal events (de Hooge et al., 2007) or imagined scenarios (Ghorbani et al., 2013) to elicit guilt and shame. Variations in recollections and imagined contexts may introduce potential confounds, making it difficult to discern whether the observed behavioral differences are caused by guilt and shame or by irrelevant contextual factors.

Neurally, whether the transformation of guilt and shame into behavior depends on distinct neural bases remains an open question. Yu et al. (2014) found that the aMCC and midbrain nuclei are involved in linking guilt to compensatory behavior. To the best of our knowledge, the neural bases underlying the conversion of shame into behavior have yet to be explored. Naturally, this also means that no studies have directly compared the neural correlates of guilt-driven and shame-driven behavior.

To address the research gaps outlined above and provide computational, algorithmic, and neural accounts of guilt and shame (see Yu et al., 2024), we combined computational modeling and functional magnetic resonance imaging (fMRI) with a novel interpersonal game. In this game, we independently manipulated the harm

inflicted on a victim and the responsibility of participants. Throughout the experiment, participants engaged in compensatory decision-making while undergoing fMRI scanning, followed by post-experimental self-reports of their guilt and shame feelings. This paradigm allowed us to explore the associations among harm, responsibility, guilt, shame, and compensation, as well as to uncover relevant neural substrates. Based on the functionalist theories (Baumeister et al., 1994; Gilbert, 1997; Landers, Sznycer, & Durkee, 2024; Sznycer, 2019; Sznycer et al., 2016), we predicted that harm would have a greater impact on guilt than shame, whereas responsibility has a greater impact on shame than guilt. Based on previous research on the relationships between guilt, shame, and altruistic behavior (de Hooge et al., 2007), we predicted that guilt would exert stronger influence on compensation than shame. Drawing from existing knowledge of the neural correlates of guilt and shame (Bastin et al., 2016; Li et al., 2020; Yu et al., 2020; Zhu, Feng, et al., 2019), we predicted that brain regions involved in emotional processing (e.g., insula), self-referential processing (e.g., DMPFC), theory-of-mind processing (e.g., TPJ), and cognitive control (e.g., LPFC) would play important roles in the neural representation of harm and responsibility, and the emergence of guilt-driven and shame-driven compensation.

Methods

Participants

We recruited 49 college students. All participants had normal or corrected-to-normal vision and reported no history of psychological or neurological disorders. They provided written consent and gained monetary payment for their participation. The experimental protocol (protocol number: ICBIR_A_0071_010) was approved by the local research ethics committee at the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University and was conducted in accordance with the Declaration of Helsinki.

Data from four participants weren't correctly recorded due to a machine malfunction. One participant dropped out of the experiment for personal reasons, one missed more than half of the experimental trials, and one fell asleep in the fMRI scanner. These seven participants were removed from all analyses. In addition, one participant who didn't compensate the receiver in any trial (i.e., all compensation decisions were 0) was excluded from the computational modeling because the model parameters couldn't be reliably estimated under such a condition (Zhong et al., 2016). Thus, unless otherwise specified, analyses unrelated to computational modeling were conducted with 42 participants (18 females, 24 males; age: $M \pm SD = 21.71 \pm 2.30$ years), while analyses related to computational modeling were conducted with 41 participants (18 females, 23 males; age: $M \pm SD = 21.71 \pm 2.33$ years).

Experimental procedures

The experiment comprised two phases (Figure 1A). In the first phase, participants completed personality questionnaires online, including the Guilt and Shame Proneness scale (GASP; Cohen et al., 2011), Gratitude Questionnaire–6 scale (GQ-6; McCullough et al., 2002), and Social Value Orientation Slider Measure (SVO Slider Measure; Murphy et al., 2011). The 16-item GASP measures trait guilt and trait shame. Participants read various scenarios and indicated the likelihood of specific responses on a 7-point Likert scale (1 = very unlikely, 7 = very likely). The scale includes two guilt subscales—negative behavior evaluations (e.g., "What is the likelihood that you would feel terrible about the lies you told?") and repair action tendencies (e.g., "What is the likelihood that you would try to act more considerately toward your friends?")—and two shame subscales—negative self-evaluations (e.g., "What is the likelihood that this would make you would feel like a bad person?") and withdrawal action tendencies (e.g., "What is the likelihood that you would avoid the guests until they leave?"). The six-item GQ-6 is a widely used measure of trait gratitude. Each item (e.g., "I am grateful to a wide variety of people") is answered on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). The six-item SVO Slider Measure assesses social value orientation

(SVO), defined as the tendency to prioritize and balance outcomes for oneself and others in interdependent situations (Murphy & Ackermann, 2014). Participants chose among several self-other payoff combinations. Based on their choices, an SVO score was calculated. Larger SVO scores indicate more prosocial orientation. The participants filled out the questionnaires online at least one day before the second phase.

We measured trait gratitude and SVO for exploratory purposes. Previous research has shown that both are linked to altruistic behavior, particularly in donation contexts (Van Lange et al., 2007; Yost-Dubrow & Dunham, 2018) and reciprocity contexts (Ma et al., 2017; Yost-Dubrow & Dunham, 2018). Here, we explored whether they also exert significant effects in a compensatory context.

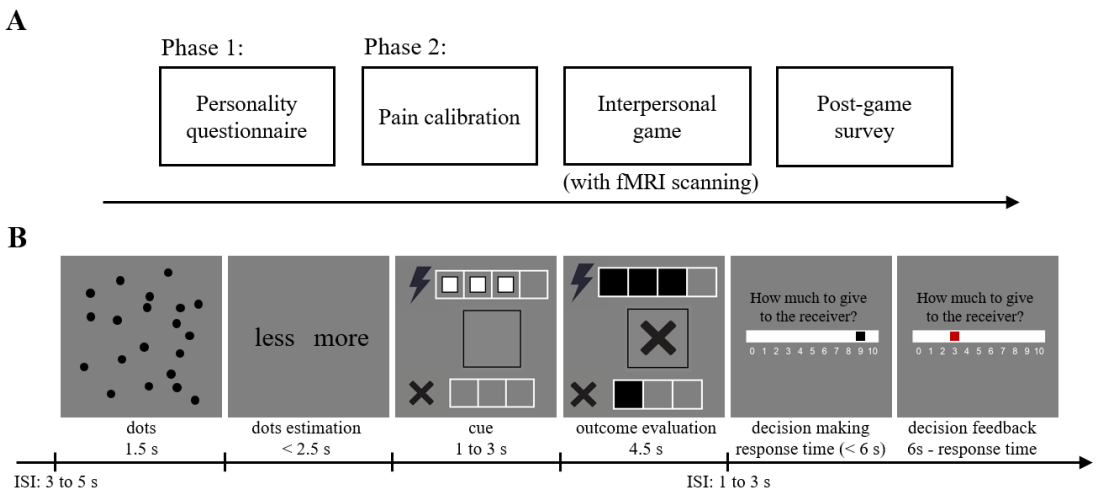

Figure 1. Experimental procedure. (**A**) Timeline for the whole experiment. (**B**) Timeline for the interpersonal game. In this example, the participant, who was one of the four deciders, made an incorrect estimation, as did one of the other deciders. ISI, inter-stimulus interval.

In the second phase, each participant arrived at our laboratory individually. Upon arrival, the participant was introduced in person to three co-players. These co-players, however, were confederates, and they were purported to interact with the participant later via an internal network. At least one confederate shared the participant's gender,

and at least one was of a different gender. The participant was then led to a separated room and completed a series of tasks.

Pain calibration. First, the participant underwent an individual pain calibration procedure with an SXC-4A multichannel electrical stimulator. After introducing the calibration process and precautions, we cleaned the participant's left forearm and placed two disposable electrodes on the back of their left wrist. The initial electric shock was set as 12 repeated square waveform electrical stimulation pulses (current intensity of each pulse: 0.2 mA; duration of each pulse: 0.5 ms; interval between pulses: 10 ms) (Yu et al., 2018). We gradually increased or decreased the current intensity of each pulse in small increments (0.5 mA) with a 3:1 ratio (Crockett et al., 2014). The participant rated their subjective experience after each shock on an 11-point Likert scale (0 = no sensation, 10 = maximum tolerable pain). The calibration procedure continued until the participant reported a rating of 8. The current intensity value corresponding to this self-reported pain rating of 8 was recorded and used in the following steps. We then adjusted the number of pulses and applied electric shocks involving 2, 4, 8 and 12 repeated pulses to the participant (Yu et al., 2018). The participant rated the subjective pain intensity for each of the four shock levels. They were informed that these four pain levels would be used in the upcoming task. Providing the participant with the opportunity to experience the electric shocks beforehand enhanced the experiment's authenticity.

Interpersonal game (with fMRI scanning). After the pain calibration, the participant played a novel interpersonal game while their blood-oxygen-level-dependent (BOLD) signals were measured using fMRI. The development of this game was inspired by several previous studies on guilt and shame (Li et al., 2020; Yu et al., 2014; Zhu, Feng, et al., 2019; Zhu, Wu, et al., 2019). In the game, there were five players: four deciders and one receiver. At the beginning of each trial, a picture of dots was presented on each decider's computer screen (dots period, 1.5 s) (Figure 1B). Each decider indicated whether the number of dots was more than or less than 20 based on their own estimation by pressing a corresponding button (dots estimation period, < 2.5 s) and was unaware

of the estimations made by other deciders. Unbeknownst to the participant, the picture always contained 20 dots and the positions of the dots varied across trials. The four deciders wouldn't receive any electric shock regardless of the correctness of their estimations. However, the receiver would receive an electric shock, if any decider made an incorrect estimation. The electric shock had four levels corresponding to the pain levels in the pain calibration procedure (pain levels 1 to 4). It was emphasized that all electric shocks were within safe limits and we had measured the pain sensitivity of all players to ensure that the deciders could experience the receiver's pain sensation caused by different levels of electric shocks. The pain level of each trial was randomly determined. After making their estimations, the deciders saw the pain level of the current trial (cue period, 1 to 3 s), which was indicated by the number of white squares in the four white frames next to the electric shock symbol (e.g., three white squares indicating pain level 3). Afterwards, the deciders viewed the outcome (outcome evaluation period, 4.5 s). The first row displayed whether the receiver was subjected to an electric shock and its pain level. If the receiver wouldn't receive any electric shock (i.e., all deciders made correct estimations), the white squares in the white frames disappeared. If the receiver would receive an electric shock, the white squares turned into black and filled the white frames. The second row showed whether the current decider made a correct estimation. A tick appeared in the black frame for a correct estimation, while a cross appeared for an incorrect estimation. The third row revealed how many other deciders, excluding the current decider, made incorrect estimations. It was represented by the number of black squares in the three white frames next to the small cross symbol. At the end of each trial, each decider was endowed 10 tokens (1 token = 0.5 Chinese yuan) and could freely distribute the tokens between themselves and the receiver (minimum unit: 1 token) by moving a black block using a response box (decision making period, < 6 s). The initial position of the block was randomized across trials, which helped minimize stable anchoring effects across trials. When the decider finalized their decision by pressing a button, the black block turned red as feedback for their response (decision feedback period, 6 s - response time). Each

decider made their distribution independently and didn't know how many tokens the other deciders gave to the receiver. The receiver could receive a maximum of 40 tokens in a trial. Five trials would be randomly chosen and actualized (i.e., implementing the electric shocks and monetary rewards) at the end of the interpersonal game.

The participant and three confederates played as the deciders. We told the participant a cover story that the receiver was played by another college student who was not present in the laboratory at the time. They had visited the laboratory, undergone pain calibration, spent time understanding the game rules, and agreed to participate as the receiver. The receiver would visit the laboratory again to see the behaviors of all the deciders and face the consequences, including electric shocks and monetary rewards. In fact, the receiver did not actually exist. Not having participants meet the receiver helped prevent excessive guilt and shame that might produce a ceiling effect, while also eliminating the need to recruit an additional confederate.

The outcomes of the interpersonal game were predetermined. The experimental trials were those in which the participant as a decider was informed that they made an incorrect estimation (i.e., the self-error condition). We had a 4 (harm: levels 1 to 4) × 4 (responsibility: levels 1 to 4) within-subjects design. To create different levels harm and responsibility, we manipulated the pain levels of the electric shock [harm level 1 (pain level 1), harm level 2 (pain level 2), harm level 3 (pain level 3), and harm level 4 (pain level 4)] and the number of other deciders who also made incorrect estimates [responsibility level 1 (3 deciders), responsibility level 2 (2 deciders), responsibility level 3 (1 deciders), and responsibility level 4 (0 decider), respectively. There were 16 possible combinations of harm levels and responsibility levels. We repeated these combinations 4 times and thus built 64 experimental trials. Unless otherwise specified, only data from these experimental trials were included in our analyses. To enhance the authenticity of the experiment, we also involved 48 filler trials, in which the participant was informed that they made a correct estimation (see Supplementary Table 1). Thus, the interpersonal game contained a total of 112 trials. We divided the trials into 4 sessions. Each session involved 16 experimental trials and 12 filler trials. The trials of

a specific session were presented in a pseudo-random order. The 4 sessions were assigned to 4 fMRI scanning runs in a balanced manner across participants.

Participants did not proceed to the interpersonal game until they had fully understood the experimental rules and passed a comprehension test.

Post-game survey (guilt and shame feelings recalling). After the interpersonal game, the outcomes of the experimental trials were re-presented in a random order. All the participants were required to recall and rate their feelings of guilt and shame (emotion ratings) when they saw those outcomes during the game (0 = not at all, 10 = very strong; 11-point Likert scale). They gave a rating of guilt and a rating of shame for each harm-responsibility combination (i.e., each participant provided 16 guilt ratings and 16 shame ratings).

Post-game survey (responsibility manipulation check). The participants were also asked to rate their perceived responsibility for causing the receiver's harm (1) when they were the only decider who made an incorrect estimation; (2) when they were one of the two deciders who made incorrect estimations; (3) when they were one of the three deciders who made incorrect estimations; (4) when they were one of the four deciders who made incorrect estimations; and (5) when they made a correct estimation, but the other deciders made wrong estimations (1 = no responsibility, 9 = high responsibility, 9-point Likert scale).

Post-game survey (additional measures and procedures). Participants rated the perceived difficulty of the dot estimation task (1 = not at all, 4 = very much, 4-point Likert scale; M = 3.07, SD = 0.46) (Yu et al., 2014) and the extent of variation in task difficulty across trials (1 = not at all, 4 = very much, 4-point Likert scale; M = 2.43, SD = 0.59). Additionally, they reported to what extent they wanted to make correct estimations in the dot estimation task (1 = not at all, 9 = very much, 9-point Likert scale; M = 7.45, SD = 1.56) and to what extent they wanted to help the receiver to avoid electric shocks (1 = not at all, 9 = very much, 9-point Likert scale; M = 7.40, SD = 1.99). The results suggest that the participants found the task moderately challenging with some variation in difficulty across trials, while remaining highly motivated to perform

accurately and positively impact the receiver, reflecting their active engagement in the task.

To examine whether the participants had any suspicion of the experiment, we conducted an interview at the very end of the experiment. No participant expressed doubts about the authenticity of the experiment.

Magnetic resonance imaging data acquisition. Whole-brain imaging data were acquired using a 3-Tesla Prisma magnetic resonance scanner at the State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University. During the interpersonal game, T2-weighted echo-planar functional images were acquired (TR/TE: 2000/30 ms; FoV: [224 mm (RL) 224 mm (AP) 143 mm (FH)]; flip angle: 90°; voxel size: 2 × 2 × 2 mm; number of slices: 62; slice thickness: 2 mm; slice gap: 0.3 mm). Additionally, a T1-weighted 3-dimensional magnetization-prepared rapid acquisition gradient echo structural image was collected for image registration (TR/TE: 2530/2.98 ms; FoV: [192 mm (RL) 224 mm (AP) 256 mm (FH)]; flip angle: 7°; voxel size: 0.5 × 0.5 × 1 mm; number of slices: 192; slice thickness: 1 mm; slice gap: 0.5 mm) and gradient-echo field map images were acquired for correcting deformations (TR/TE1/TE2: 620/4.92/7.38 ms; FoV: [224 mm (RL) 224 mm (AP) 143 mm (FH)]; flip angle: 60°; voxel size: 2 × 2 × 2 mm; number of slices: 62; slice thickness: 2 mm; slice gap: 0.3 mm).

Behavioral analyses

Harm and responsibility manipulation checks. Painful feelings caused by electric shocks can be regarded as a form of harm (e.g., Crockett et al., 2017). To examine whether the electric shocks induced distinguishable perceptions of pain, we compared participants' pain ratings pairwise across four levels of harm by conducting repeated measures analyses of variance (ANOVA). To examine whether the number of other deciders who made incorrect estimates with the participant caused distinguishable perceptions of responsibility, we compared participants' responsibility ratings pairwise across four levels of responsibility by conducting repeated measures ANOVA.

Additionally, to confirm that participants' perceptions of responsibility were influenced by their correctness of estimates, we compared participants' responsibility ratings between conditions where they made incorrect and correct estimates using repeated measures ANOVA.

The effects of harm and responsibility on guilt and shame. To test the influence of harm and responsibility on guilt and shame, we performed two separate linear mixed-effect regressions. In the first regression, we regressed participants' emotion ratings onto harm levels, emotion types (guilt vs. shame), and their interaction. In the second regression, we regressed participants' emotion ratings onto responsibility levels, emotion types (guilt vs. shame), and their interaction. The significance of the interaction effects in these two models allowed us to evaluate whether harm and responsibility had distinct influence on guilt and shame, which is a key question we are concerned with. We also conducted simple slope analyses for each emotion type to test whether harm and responsibility had significant effects on guilt and shame, respectively. As a complementary analysis, we conducted another linear mixed-effect regression, in which participants' emotion ratings were regressed onto harm levels, responsibility levels, emotion types (guilt vs. shame), and all possible interaction. We used it to test whether the interaction between harm and emotion type, as well as the interaction between responsibility and emotion type, remained significant when controlling for the effects of other regressors. For all the regressions above, participant-specific random constants and random slopes for each fixed effect were involved as random effects.

The effects of guilt and shame on compensation. To examine whether and how guilt and shame affect compensation, we built four linear mixed-effect regression models. In all models, the dependent variable remained the same: the amount of compensation (i.e., the number of tokens that participants distributed to the receiver), while the fixed-effect regressors varied. Model I included guilt rating as the sole regressor. Model II included shame rating as the sole regressor. Model III included both guilt and shame ratings.

Model IV included guilt rating, shame rating, and their interaction. In all the regression models, participant-specific random constants and random slopes for each fixed effect were involved as random effects. The model with the lowest Bayesian information criterion (BIC) value was selected as the best model. Model III won, which suggest guilt and shame each uniquely contribute to compensatory behavior. To further understand the effects of guilt and shame, we used Wald chi-square tests to directly compared their influence (i.e., regression coefficients) on compensation based on the best model (i.e., Model III). Besides, to test whether the difference in the effects of guilt and shame on compensation was due the difference in their intensity, we compare participants' average guilt and shame ratings across trials using repeated measures ANOVA.

Correlation analyses. Although the current study focused on state guilt and shame, we were aware of prior findings suggesting that trait guilt is associated with moral behavior, whereas trait shame seems not (Cohen et al., 2011; Tangney & Dearing, 2003). To replicate these findings, we calculated Pearson correlation coefficients to examine the relationship between the amount of compensation and trait guilt and trait shame. Additionally, we conducted exploratory analyses by calculating Pearson correlation coefficients between the amount of compensation and trait gratitude and social value orientation (SVO).

Computational modeling

Model development. After validating the associations among harm, responsibility, guilt, shame, and compensation through separate linear mixed-effect regressions, we sought to advance mechanistic understanding of guilt-driven and shame-driven compensatory decision-making through computational modeling. Drawing from the phenomenon of responsibility diffusion (Darley & Latané, 1968), we hypothesized that individuals perceive responsibility as being diluted when multiple wrongdoers are involved. This implies a cognitive mechanism wherein individuals mentally distribute harm across all

wrongdoers, thereby reducing their altruistic behavior (compensation). Based on this Diffusion Hypothesis, we constructed Model family 1, where harm and the number of wrongdoers are integrated in the form of a quotient. All models within this family share this psychological mechanism, differing only in whether self-interest was accounted for (e.g., Wu et al., 2024) and whether a compensatory baseline was incorporated (e.g., Feng et al., 2023), both of which are common psychological components involved in social decision-making.

The models of Model family 1 are described as below:

Model family 1 (Diffusion Hypothesis):

Model 1.1 (self-interest considered and compensatory baseline exists):

$$U(D) = \theta * (10 - D) - (1 - \theta) * |\kappa * \frac{H}{W} + \eta - D|$$

where $U(D)$ is the utility for a compensatory decision. D is the number of tokens distributed to the receiver by participants and $10 - D$ is the number of tokens left for participants themselves (i.e., self-interest). H and W are the level of harm and the number of wrongdoers, respectively. W is the complement transformation of the level of responsibility (W = 5 – the level of responsibility). $\frac{H}{W}$ denotes the average harm for which participants are responsible, suggesting participants integrated harm and the number of wrongdoers in the form of a quotient. The parameter κ represents compensatory sensitivity. Larger κ values indicate that participants are willing to allocate more tokens to compensate for the harm they are responsible for. The parameter η represents compensatory baseline. Larger η values indicate participants are inclined to allocate more tokens to the receivers for compensation, irrespective of the level of harm and responsibility. $\kappa * \frac{H}{W} + \eta$ is the total number of tokens that participants believe should be distributed to the receiver as compensation. $-|\kappa * \frac{H}{W} + \eta - D|$ means that participants are averse to providing the receiver less compensation (undercompensation) or more compensation (overcompensation) they believe the receiver deserves (i.e., improper compensation aversion). The parameter θ represents

greed. Larger θ values indicate participants put more weight on self-interest (relative to improper compensation aversion). Model 1.1 is the full model in model family 1, while the following Model 1.2, Model 1.3, and Model 1.4 are simplified versions, each omitting specific components.

Model 1.2 (self-interest considered and compensatory baseline doesn't exist):

$$U(D) = \theta * (10 - D) - (1 - \theta) * |\kappa * \frac{H}{W} - D|$$

Model 1.3 (self-interest not considered and compensatory baseline exists):

$$U(D) = -|\kappa * \frac{H}{W} + \eta - D|$$

Model 1.4 (self-interest not considered and compensatory baseline doesn't exist):

$$U(D) = -|\kappa * \frac{H}{W} - D|$$

It is also plausible that individuals might perceive their responsibility as intensifying the impact of harm (Moore, 2015). Under this mechanism, harm and responsibility interact in a multiplicative manner, resulting in an amplified sense of moral obligation and increased compensatory behaviors. Based on this Amplification Hypothesis, we built Model family 2. The models in the Model family 2 were the same as the Model family 1, except that the quotient was replaced by the product of harm and responsibility. The models of Model family 2 are described as below:

Model family 2 (Amplification Hypothesis):

Model 2.1 (self-interest considered and compensatory baseline exists):

$$U(D) = \theta * (10 - D) - (1 - \theta) * |\kappa * H * R + \eta - D|$$

where R is the level of responsibility (R = 5 - W). Model 2.1 is the full model in model family 2, whereas the following Model 2.2, Model 2.3, and Model 2.4 are simplified versions.

Model 2.2 (self-interest considered and compensatory baseline doesn't exist):

$$U(D) = \theta * (10 - D) - (1 - \theta) * |\kappa * H * R - D|$$

Model 2.3 (self-interest not considered and compensatory baseline exists):

$$U(D) = -|\kappa * H * R + \eta - D|$$

Model 2.4 (self-interest not considered and compensatory baseline doesn't exist):

$$U(D) = -|\kappa * H * R - D|$$

We note that Model Families 1 and 2 reflect distinct psychological mechanisms and differ fundamentally in their mathematical formulations ($\frac{H}{W}$ vs. $H * R$). Because the correlation between these two mathematical expressions is substantially below 1, model comparisons can thus empirically determine which model more accurately captures participants' compensatory decision-making.

For all the models, we transformed the trial-by-trial utility into choice probabilities (P) using the softmax function:

$$P(D) = \frac{e^{\lambda * U(D)}}{\sum_{j \in J} e^{\lambda * U(D_j)}}$$

where D_j represents a possible number of tokens that distributed to the receiver by participants (from 0 to 10 tokens, in increments of 1). J is the full set of j. λ is the inverse temperature parameter indicating the level of stochasticity in the compensatory decision.

Parameter estimation. For each model, we used the fmincon function in MATLAB to estimate the parameters for each participant. To decrease the possibility of the model converging on a local minimum, we constructed a coarse grid in the parameter space to choose the best start location. Maximum likelihood estimation was performed by maximizing the log likelihood function over each participant i and trial t:

$$\sum_i \sum_t \log\left(P_{i,t}(D)\right)$$

Model comparison. We compared the models adopting the Bayesian information criteria, which considers both goodness of fit and model complexity. Model 1.3 was selected as the winning model. It assumes that individuals neglect their self-interest, have a compensatory baseline, and integrate harm and the number of wrongdoers in the form of a quotient.

Model validation. We performed a series of analyses to validate the winning model (Model 1.3). To evaluate predictive accuracy, we used the estimated parameters of the winning model to simulated each participant's compensatory decisions 50 times, generating 50 sets of simulated data. Predictive accuracy for each participant was calculated as the proportion of simulated decisions that matched their real decisions. We tested whether the predictive accuracies of our participants exceeded the chance level (9%) using a one-sample *T* test. We also visualized the effects of harm level and responsibility level on both simulated and real decisions, checking whether the behavioral patterns of the simulated decision aligned with those of the real decisions (Figures 3A and 3B).

To demonstrate that the parameters of the winning model are identifiable, we conducted parameter recovery analyses. We refitted the model using the 50 sets of simulated data. We then evaluated the degree to which the parameter could be recovered by calculating the Pearson correlation coefficients between the parameters estimated from the real data and the parameters estimated from the simulated data.

To explore the relationships among the parameters of the winning model, we calculated the Pearson correlation coefficients among the parameters estimated from the real data.

fMRI analyses

fMRI data preprocessing. Imaging data were preprocessed using the Statistical Parametric Mapping 12 (SPM 12, https://www.fil.ion.ucl.ac.uk/spm/software/spm12/). Functional images were realigned to the first volume for head-motion correction, corrected for geometric distortions using field maps, slice-time corrected to the acquisition time of the middle slice. The mean functional image was co-registered with each participant' T1-weighted image using parameters from the segmentation performed on both type of images. The functional images were then normalized into the Montreal Neurological Institute (MNI) space, resampled to 3-mm isotropic voxels, and smoothed with an isotropic 8-mm full width at half-maximum Gaussian kernel.

Activation analysis (neural representation of cognitive antecedents). We built several separate general linear models (GLMs) to explore how the cognitive antecedents of the guilt and shame are represented in the brain during outcome evaluation. They were designed to test different possibilities while addressing the problems of multicollinearity among regressors.

GLM 1 aimed to identify brain areas responding parametrically to the level of harm and level of responsibility. At the first level, we modeled the periods of dots (duration: 1.5 s), dots estimation (duration: 2.5 s), cue (duration: 1 to 3 s), outcome evaluation (duration: 4.5 s), decision making (duration: response time), decision feedback (duration: 6 s - response time), and missing trials (if existed) (Figure 1B). The period of outcome evaluation was modeled using two event regressors: one for experimental trials where participants made incorrect dot estimations (i.e., the self-error condition), and another for filler trials where participants made correct dot estimations. All other periods were modeled using a single regressor that encompassed all trials. Our analysis focused on the period of outcome evaluation in the self-error condition. For this corresponding event regressor, we included two parametric modulators: the level of harm (H) and the number of wrongdoers (W) (number of wrongdoers = 5 – responsibility level). Notably, in our experimental design, we ensured that the values of these two modulators were orthogonal, effectively avoiding multicollinearity issues. No additional orthogonalization adjustments were applied during the fMRI analysis. The regressors were convolved with a canonical hemodynamic response function. To control a motion effect of no interest, six head movement parameters from the realignment procedure were involved in the GLM as nuisance regressors. The inter-stimulus interval (i.e., jitter), which was not modeled in the GLM, served as an implicit baseline. We defined two contrasts at the first level, corresponding to each parametric modulator: (1) the level of harm, (2) the number of wrongdoers, each compared to the implicit baseline.

As complement analyses, we built another two GLMs to test the extent to which including the level of harm and the number of wrongdoers in the separate GLMs affected the neural findings compared to modeling them in same GLM. The new GLMs (GLM 2 and GLM 3) were identical to GLM 1, except that each of them had only one parametric modulator (GLM 2: the level of harm; GLM 3: the number of wrongdoers) instead of two (GLM 1: the level of harm and the number of wrongdoers). For each of GLMs 2 and 3, we defined one contrast at the first level, with the parametric modulator compared to the implicit baseline.

GLM 4 targeted to identify brain areas responding to the average harm for which the participant is responsible (i.e., the quotient of the level of harm divided by the number of wrongdoers). It is same as the GLM 3, except that the parametric modulator was replaced with the quotient ($\frac{H}{W}$). For GLM 4, we defined one contrast at the first level, with the parametric modulator compared to the implicit baseline.

GLM 5 set out to identify brain areas responding to the product of the level of harm and the level of responsibility (R). It is same as the GLM 4, except that the parametric modulator was replaced with the product (H×R). For GLM 5, we defined one contrast at the first level, with the parametric modulator compared to the implicit baseline.

At the second level, we submitted the first-level contrast images from all the GLMs above to one-sample *T* tests. The statistical threshold was set as $P < 0.05$ [family-wise error correction (FWE) at the cluster level] with an initial cluster-defining voxel-level threshold of $P < 0.001$ for the whole-brain analysis. This statistical threshold was applied to all fMRI data analyses unless stated otherwise.

Activation analysis (neural basis of emotion sensitivity). The harm-driven guilt sensitivity, responsibility-driven guilt sensitivity, harm-driven shame sensitivity, and responsibility-driven shame sensitivity refer to the degree to which harm and responsibility elicit guilt and shame for an individual. We quantified them by the individual-specific coefficient estimates of harm and responsibility on guilt and shame

from the linear mixed-effect regressions (Supplementary Tables 4 and 5). To examine whether these emotion sensitivities are correlated with neural responses to the quotient of harm divided by the number of wrongdoers, we submitted the first-level contrast images from GLM 4 to one-sample *T* tests and involved the four types of emotion sensitivity as covariates at the second level.

The results showed that the neural responses in the temporoparietal junction/superior temporal sulcus (TPJ/STS) and precentral cortex/postcentral cortex/supplementary motor area (PRC/POC/SMA) were negatively correlated with the responsibility-driven shame sensitivity. To test whether these regions were more involved in responsibility-driven shame sensitivity than in other types of emotion sensitivity, we implemented a leave-one-subject-out (LOSO) cross-validation procedure (e.g., Esterman et al., 2010). In each fold, clusters in the TPJ/STS and PRC/POC/SMA showing significant correlations with responsibility-driven shame sensitivity were identified at the group level based on *N*-1 participants. These clusters, defined as regions of interest (ROI), were then applied to the left-out participant, from whom we extracted the mean parameter estimates (i.e., neural response values). If, in a given fold, no suprathreshold cluster was detected within the TPJ/STS or PRC/POC/SMA after correction, or if the two regions merged into a single cluster that could not be separated, the corresponding value was coded as missing. Repeating this procedure across all folds yielded an independent set of ROI-based estimates for each participant. In the LOSO cross-validation procedure, the TPJ/STS and PRC/POC/SMA merged into a single inseparable cluster in two folds, and no suprathreshold cluster was detected within the TPJ/STS in one fold. These instances were coded as missing, resulting in valid data from 39 participants for the TPJ/STS and 40 participants for the PRC/POC/SMA. We then correlated these estimates with all four types of emotion sensitivities and compared the correlation with responsibility-driven shame sensitivity against those with the other sensitivities using *Z* tests (Pearson and Filon's *Z*).

Activation analysis (neural basis of compensatory sensitivity). The guilt-driven compensatory sensitivity and shame-driven compensatory sensitivity refer to the extent to which guilt and shame are converted into compensation for an individual. They can be indicated by the individual-specific coefficient estimates of guilt and shame on compensation from the linear mixed-effect regression (Supplementary Table 8). To identify whether these two types of compensatory sensitivity are associated with brain activity during outcome evaluation, an additional contrast was defined at the first level based on GLM 4. This contrast examined average brain activity during outcome evaluation (effect of the event regressor: viewing the outcome in the self-error condition) (Losin et al., 2020). At the second level, we submitted the first-level contrast images from GLM 4 to one-sample *T* tests and involved the two types of compensatory sensitivity as covariates.

We found that neural activity in the left temporal pole (TP), bilateral inferior parietal lobe (IPL), and left lateral prefrontal cortex (LPFC) was only correlated with guilt-driven or shame-driven compensatory sensitivity (Figures 5A and 5B). To directly test whether these regions were more involved in one of the two types of compensatory sensitivity, we applied the same LOSO cross-validation procedure described above. In this procedure, no suprathreshold cluster was detected within the LPFC in one fold and within the TP in 27 folds. These cases were coded as missing, resulting in valid data from 42 participants for the bilateral IPL, 41 participants for the LPFC, and 15 participants for the TP. The limited sample size for the TP likely reflects that its effect was only marginally above the correction threshold, such that the reduced power in cross-validation often rendered it nonsignificant. Because the sample size for the TP was too small and the results may therefore be unreliable, we did not pursue further analyses for this region. The independent ROI-based estimates were then correlated with both guilt-driven and shame-driven compensatory sensitivities, and the strength of the correlations was compared using *Z* tests (Pearson and Filon's *Z*).

Compensation may be driven by emotion or motivation other than guilt and shame (e.g., indebtedness) (Gao et al., 2024). The parameter κ from our winning computational

model (i.e., Model 1.3) is supposed to capture the combined influence of various psychological processes on compensation, including guilt and shame. To confirm the relationships between κ and guilt-driven and shame-driven compensatory sensitivities, we conducted a linear mixed-effects regression. κ was regressed onto guilt-driven and shame-driven compensatory sensitivities, with participant-specific random intercepts and random slopes for each fixed effect included as random effects. Besides, to more comprehensively explore the neural basis of compensatory sensitivity, we submitted the first-level contrast images from GLM 4 to one-sample *T* tests, incorporating the values of κ as covariates.

Activation analysis and mediation analysis (neural correlates of trait guilt and compensation). As we observed significant correlations between trait guilt scores and compensatory behavior, we further explored brain activity that mediates their relationships. At the first step, we attempted to identify the neural correlates of the two dimensions of trait guilt by submitting first-level contrast images from GLM 4 to one-sample *T* tests and involved the negative behavior-evaluations and repair action tendencies scores as covariates at the second level. The aMCC and midbrain nuclei are defined as the ROIs, given that they have been found to be associated with guilt and compensation (Yu et al., 2014). The ROIs for aMCC and midbrain nuclei were respectively defined as a 6-mm-radius sphere centered at the Montreal Neurological Institute (MNI) coordinates [0, 34, 16] and [-2, -20, -20] (Yu et al., 2014). Brain activities that revealed significant correlations with trait guilt scores were considered as candidate mediators (e.g., Losin et al., 2020). The statistical threshold was set as $P < 0.05$ [family-wise error correction (FWE) at the cluster level] with an initial cluster-defining voxel-level threshold of $P < 0.001$ for the small-volume correction analysis or the whole-brain analysis.

At the second step, we examined whether the candidate mediators mediated the relationship between trait guilt and compensation using the PROCESS macro based on the SPSS software (http://www.processmacro.org/index.html). We tested the mediation

effect, with participants' trait guilt scores as the predictor variable, compensation as the outcome variable, and mean estimates extracted from the clusters we identified in the first stage or the ROIs (Supplementary Table 19) as the mediator variable (one candidate mediator per mediation analysis). We used a bootstrap procedure (randomly sampling 5000 observations with replacement) to obtain 95% confidence intervals (CI) of path coefficients for significance testing. If the CI of the path coefficient did not cover zero, the effect was considered as significant.

For completeness, we explored the neural correlates of other trait by submitting first-level contrast images from GLM 4 to one-sample *T* tests and involved the trait shame, trait gratitude, or SVO scores as covariates at the second level.

Results

Manipulation checks. Participants' pain ratings differed significantly across the four harm levels (harm level 1 vs. harm level 2: $F(1,41) = 324.63$, $P < 0.001$, partial $\eta^2 = 0.888$; harm level 2 vs. harm level 3: $F(1,41) = 181.26$, $P < 0.001$, partial $\eta^2 = 0.816$; harm level 3 vs. harm level 4 $F(1,41) = 247.41$, $P < 0.001$, partial $\eta^2 = 0.858$). Their responsibility ratings differed significantly across the four responsibility levels (responsibility level 1 vs. responsibility level 2: $F(1,41) = 36.41$, $P < 0.001$, partial $\eta^2 = 0.470$; responsibility level 2 vs. responsibility level 3: $F(1,41) = 46.47$, $P < 0.001$, partial $\eta^2 = 0.531$; responsibility level 3 vs. responsibility level 4: $F(1,41) = 139.24$, $P < 0.001$, partial $\eta^2 = 0.773$) (Supplementary Figure 1; Supplementary Tables 2 and 3). In addition, participants reported higher responsibility ratings when they made incorrect estimates compared to correct estimates (making right estimates vs. responsibility level 1: $F(1,41) = 56.63$, $P < 0.001$, partial $\eta^2 = 0.580$; making right estimates vs. responsibility level 2: $F(1,41) = 148.25$, $P < 0.001$, partial $\eta^2 = 0.783$; making right estimates vs. responsibility level 3: $F(1,41) = 237.52$, $P < 0.001$, partial $\eta^2 = 0.853$; making right estimates vs. responsibility level 4: $F(1,41) = 495.73$, $P < 0.001$, partial

η^2 = 0.924). These results suggest our manipulations of harm and responsibility were successful.

Distinct impacts of harm and responsibility on guilt and shame. To confirm and compare the effects of harm and responsibility on guilt and shame, we ran two linear mixed-effect regression analyses. In the first analysis, participants' emotion ratings were regressed onto harm levels, emotion types (guilt vs. shame), and their interaction. In the second analysis, participants' emotion ratings were regressed onto responsibility levels, emotion types (guilt vs. shame), and their interaction. To determine whether harm and responsibility had significant effects on guilt and shame respectively, we performed simple slope analyses for each emotion type. We found that harm significantly increased participants' guilt ratings (β_{guilt} = 0.74, $T(67)$ = 8.92, P < 0.001) and shame ratings (β_{shame} = 0.23, $T(67)$ = 2.77, P = 0.007). Similarly, responsibility significantly increased participants' guilt ratings (β_{guilt} = 0.57, $T(61)$ = 5.27, P < 0.001) and shame ratings (β_{shame} = 0.93, $T(61)$ = 8.54, P < 0.001). Of importance, harm had a stronger effect on guilt than on shame, as indicated by the interaction effect between harm and emotion type (β_{guilt} - β_{shame} = 0.51, $T(41)$ = 5.96, P < 0.001) (Figure 2A; Supplementary Table 4). Conversely, responsibility exerted a greater influence on shame than on guilt, as shown by the interaction effect between responsibility and emotion type (β_{guilt} - β_{shame}) = -0.35, $T(41)$ = 3.60, P < 0.001) (Figure 2B; Supplementary Tables 5). To control for the potential influence of other interaction effects, we conducted an additional linear mixed-effect regression. We regressed participants' emotion ratings onto harm levels, responsibility levels, emotion types (guilt vs. shame), and all their interactions. The results showed that the interaction between harm and emotion type, as well as the interaction between responsibility and emotion type, remained significant (Supplementary Table 6). Together, these findings demonstrate the differential impacts of harm and responsibility on guilt and shame.

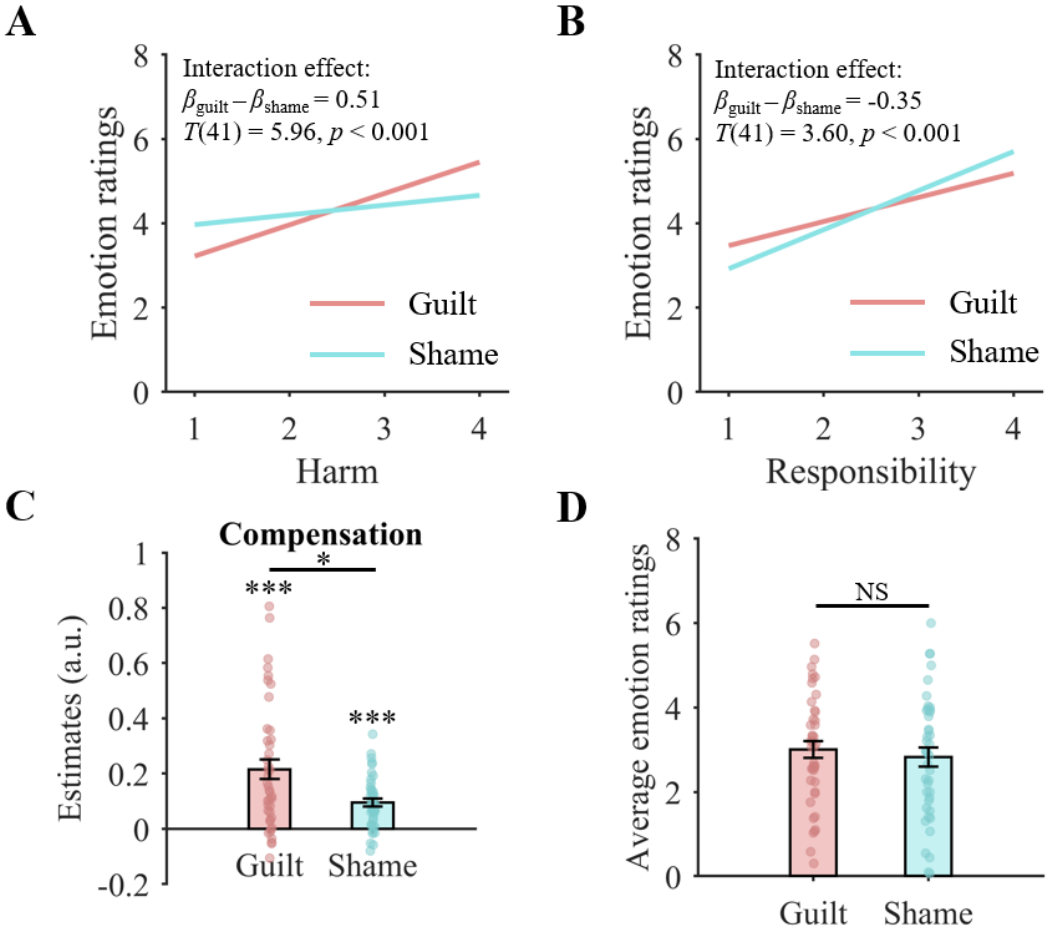

Figure 2. Behavioral results. (**A**) Harm had a stronger effect on guilt than on shame. (**B**) Responsibility revealed a stronger effect on shame than on guilt. (**A, B**) We created regression equations for guilt and shame ratings using the coefficient estimates from the linear mixed-effect regression analyses. To illustrate the impacts of harm and responsibility, we visualized the regression lines based on these equations. (**C**) The coefficient estimates from the linear mixed-effect regression model showed that, compared with shame, guilt exerted a larger effect on compensation. (**D**) Participants' average guilt and shame feelings showed no significant difference. (**C, D**) Data are shown as the mean ± standard error with overlaid dot plots. $^{*}P < 0.05$, $^{***}P < 0.001$; NS, not significant.

Guilt and shame differed in their effects on compensatory behavior. To investigate whether and how guilt and shame contributed to compensation, we tested 4 linear

mixed-effect regression models. The models regressed the number of tokens that participants distributed to the receiver on different fixed-effect regressors (see Methods). The results of the model comparison showed that the best model (Model III, Supplementary Table 7) included guilt and shame ratings (but not their interaction) as the regressors. The statistical results of the best model revealed that both guilt ($\beta = 0.22$, $T(41) = 5.16$, $P < 0.001$) and shame ($\beta = 0.10$, $T(36) = 3.69$, $P < 0.001$) were significantly associated with compensation (Supplementary Table 8). Furthermore, a direct comparison between the coefficient estimates indicated that guilt had a stronger effect on compensation than shame ($\chi^2(1) = 6.48$, $P = 0.011$) (Figure 2C). This difference in the effects of guilt and shame on compensation was not attributable to differences in the intensity of guilt and shame, as participants' average guilt and shame ratings across trials showed no significant difference ($F(1,41) = 1.11$, $P = 0.297$, partial $\eta^2 = 0.026$) (Figure 2D). Our results provide solid evidence supporting that guilt, compared to shame, is more effectively converted into compensation.

Personality traits and compensatory behavior. Consistent with previous findings (e.g., Cohen et al., 2011), we found that two dimensions of trait guilt scores were significantly correlated to participants' compensatory behavior (negative behavior-evaluations: $R = 0.39$, $P = 0.010$; repair action tendencies: $R = 0.33$, $P = 0.030$), whereas two dimensions of trait shame scores were not (negative self-evaluations: $R = 0.20$, $P = 0.213$; withdrawal action tendencies: $R = 0.16$, $P = 0.315$). These findings again demonstrate the distinct behavioral impacts of guilt and shame. The trait gratitude scores and SVO scores were not significantly correlated with compensatory behavior (trait gratitude: $R = 0.18$, $P = 0.255$; SVO: $R = 0.24$, $P = 0.121$).

Computational models of compensatory behavior. After confirming their associations through separate linear mixed-effect regression analyses, we used utility models to characterize how individuals integrate harm with responsibility during guilt-driven and shame-driven compensatory decision-making within a unified computational

framework. We constructed 8 models across 2 model families, which aimed at capturing individuals' distinct latent psychological processes. They varied in whether self-interest was considered, whether compensatory baseline existed, and how harm and responsibility were integrated (see Methods). Based on the Bayesian information criteria, Model 1.3 outperformed other alternative models (Supplementary Table 9). This model assumes that, during compensatory decision-making, individuals disregard their self-interest, adopt a compensatory baseline, and integrate harm and the number of wrongdoers in the form of a quotient (rather than integrating harm and responsibility in the form of a product). The utility (U) of a compensatory decision is described as follows (Model 1.3):

$$U(D) = -|\kappa \times \frac{H}{W} + \eta - D|$$

where H and W are the level of harm and the number of wrongdoers, respectively. $\frac{H}{W}$ represents the average harm for which participants are responsible. The parameter $\boldsymbol{\kappa}$ reflects compensatory sensitivity. Higher κ values indicate that participants are inclined to spend more tokens to compensate for the harm they are responsible for. The parameter η denotes compensatory baseline. Higher η values indicate that participants believe they should distribute more tokens to the receiver for compensation, regardless of the level of harm and the number of wrongdoers. D is the number of tokens distributed to the receiver by participants. It is assumed that participants are averse to giving the receiver less compensation (undercompensation) and more compensation (overcompensation) they believe the receiver deserves.

We validated the winning model (Model 1.3) through a variety of analyses. Simulation tests indicated that the model could replicate the influence of harm level and responsibility level on compensation through simulation (see Figures 3A and 3B) and predict participants' compensatory decisions with significantly higher accuracy than the chance level (mean accuracy: 31%, 95% confidence interval (CI): [27%, 36%], chance level: 9%). Parameter recovery tests revealed that the parameters were highly identifiable (Pearson correlation between real and recovered parameters: r_κ = 0.95, CI

= [0.95, 0.96]; r_η = 0.99, CI = [0.99, 0.99]). No evidence supports that the psychological processes represented by the parameters κ and η overlap, as the Pearson correlations between the two parameters were not significant (R = -0.17, P = 0.289) (Figure 3C).

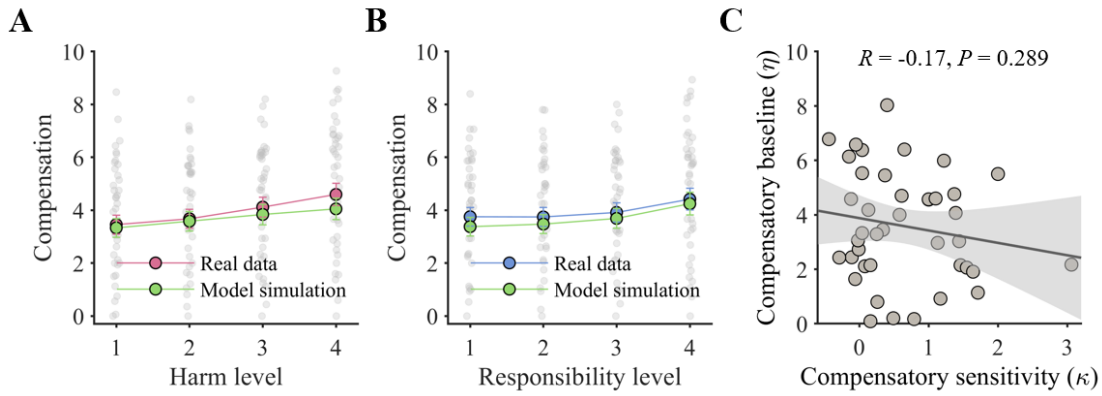

Figure 3. Computational modeling results. (**A, B**) Model simulations reproduced the behavioral patterns of compensatory decisions as influenced by harm (**A**) and responsibility (**B**). Data are shown as the mean ± standard error with overlaid dot plots. (**C**) The compensatory sensitivity (κ) and compensatory baseline (η) had no significant correlation. The line represents the least squares fit with shading showing the 95% confidence interval.

Neural representation of cognitive antecedents. To investigate the neural representation of the cognitive antecedents of guilt and shame during outcome evaluation, we constructed general linear models (GLMs) to identify brain regions that respond parametrically to harm, the number of wrongdoers, the quotient of harm divided by the number of wrongdoers, and the product of harm and responsibility (responsibility = 5 – the number of wrongdoers) (see Methods). No brain region showed significant responses to harm, which implies the brain doesn't represent harm in isolation. The precentral and postcentral cortices, known to be involved in movement preparation (Porro et al., 1996; Thoenissen et al., 2002), exhibited positive parametric responses to the number of wrongdoers, potentially reflecting motor preparation for subsequent button-press action related to compensation (Supplementary Table 10). The results of

the neural representation of the level of harm and the number of wrongdoers remained consistent, regardless of whether these two parametric modulators were put in the same GLM or in separated GLMs (Supplementary Tables 10 and 11).

The quotient of harm divided by the number of wrongdoers correlated parametrically with the activation in brain regions involved in social cognition. Specifically, the striatum linked to value computation (Bartra et al., 2013; Clithero & Rangel, 2013) and the posterior insula (pINS) associated with inequity aversion (Gao et al., 2018; Hsu et al., 2008) exhibited negative parametric responses to the quotient (Figure 4A; Supplementary Table 12). In contrast, no brain region showed significant responses to the product of harm and responsibility. Together with the findings from computational modeling, these fMRI results suggest that individuals integrate harm and the number of wrongdoers in the form of a quotient, rather than integrating harm and responsibility in the form of a product.

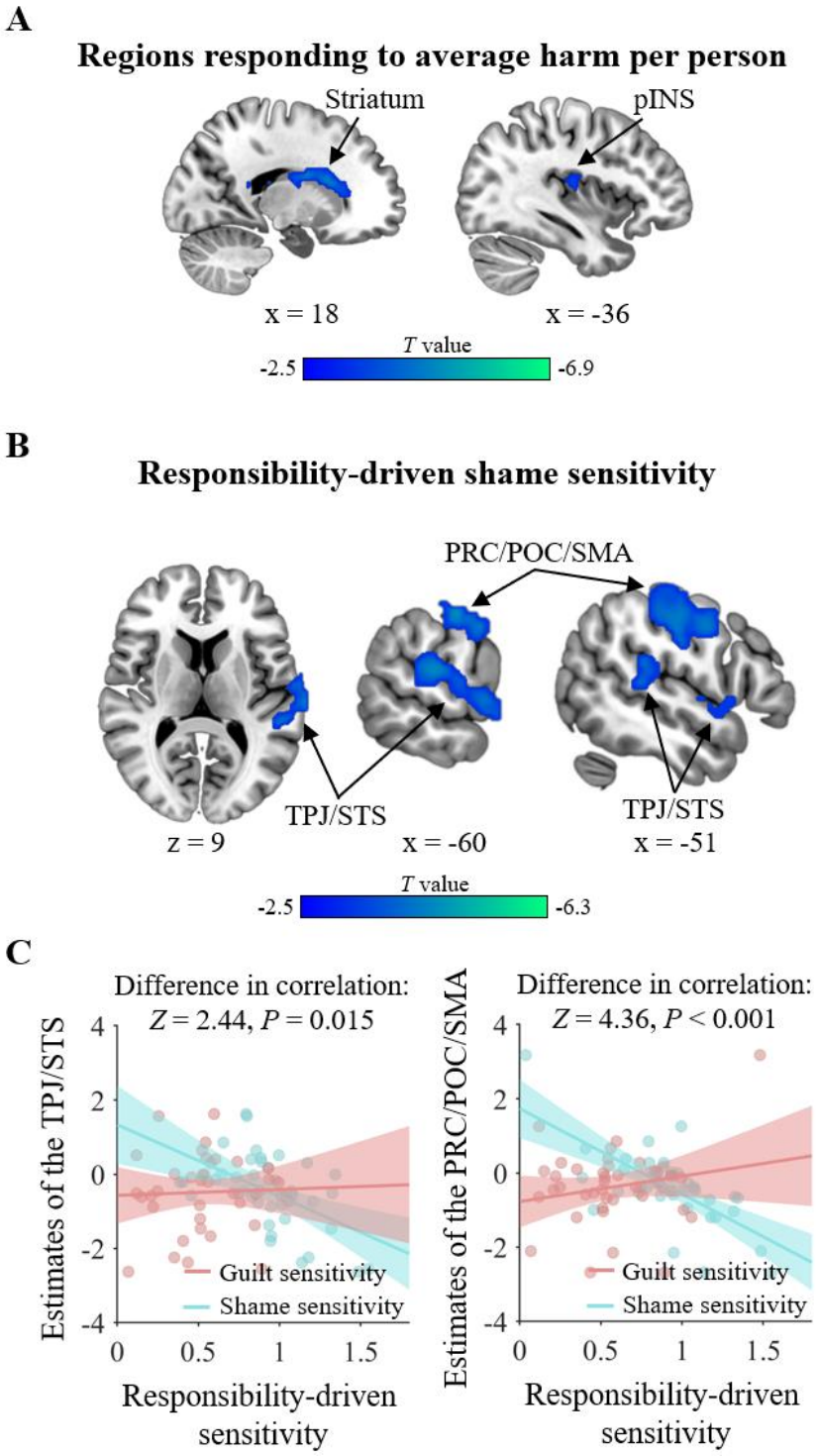

Figure 4. Neural representation of cognitive antecedents and neural basis of emotion sensitivity. (**A**) The quotient of harm divided by the number of wrongdoers (i.e., average harm per person) is represented by the striatum and posterior insula (pINS). (**B**) The neural responses to average harm per person in the two clusters containing temporoparietal junction (TPJ)/ superior temporal sulcus (STS) and

precentral cortex (PRC)/postcentral cortex (POC)/ supplementary motor area (SMA) were negatively correlated with responsibility-driven shame sensitivity. (**A, B**) Negative *T* values indicate negative correlation. Whole-brain FWE-cluster correction at $P < 0.05$ after cluster-forming-threshold at $P < 0.001$. (**C**) The TPJ/STS and PRC/POC/SMA clusters showed a significantly stronger negative correlation with shame-driven sensitivity than with guilt-driven sensitivity. Each line represents the least squares fit with shading showing the 95% confidence interval.

Neural basis of emotion sensitivity. To examine the neural basis of the individual harm-driven guilt sensitivity, responsibility-driven guilt sensitivity, harm-driven shame sensitivity, and responsibility-driven shame sensitivity, we correlated the coefficient estimates of harm and responsibility on guilt and shame from the linear mixed-effect regressions (Supplementary Tables 4 and 5) with the whole-brain neural activity. The whole-brain analysis revealed that the parametric responses of the theory-of-mind areas (i.e., TPJ and superior temporal sulcus (STS)) to the quotient of harm divided by the number of wrongdoers were negatively correlated with responsibility-driven shame sensitivity (Figure 4B; Supplementary Table 13). Namely, participants who more intensively converted responsibility into shame had weaker neural responses to the average harm per person in these regions. To assess whether these brain regions were specifically involved in responsibility-driven shame sensitivity, we compared the Pearson correlations between their activity and all types of emotion sensitivities. The results demonstrated the domain specificity of these regions, by revealing that the TPJ/STS cluster had significantly stronger negative responses to responsibility-driven shame sensitivity than to responsibility-driven guilt sensitivity ($Z = 2.44$, $P = 0.015$) and harm-driven shame sensitivity ($Z = 3.38$, $P < 0.001$), and a marginally stronger negative response to harm-driven guilt sensitivity ($Z = 1.87$, $P = 0.062$) (Figure 4C; Supplementary Table 14). In addition, the sensorimotor areas (i.e., precentral cortex (PRC), postcentral cortex (POC), and supplementary motor area (SMA)) exhibited the

similar activation pattern as the TPJ/STS (Figure 4B and 4C; Supplementary Tables 13 and 14).

No brain response showed significant correlation with harm-driven guilt sensitivity, responsibility-driven guilt sensitivity, and harm-driven shame sensitivity.

Neural basis of compensatory sensitivity. To examine the neural basis of the individual guilt-driven and shame-driven compensatory sensitivities, we correlated the coefficient estimates of guilt and shame on compensation from the linear mixed-effect regression (Supplementary Table 8) with the whole-brain neural activity. The whole-brain analysis showed that the activation in the DMPFC/SMA was significantly correlated with both guilt-driven and shame-driven compensatory sensitivities (Supplementary Table 15). Besides, the activity in the left temporal pole (TP) was significantly correlated with guilt-driven compensatory sensitivity but not with shame-driven compensatory sensitivity (Figure 5A), whereas activity in the bilateral inferior parietal lobe (IPL), and left LPFC clusters was significantly correlated with shame-driven compensatory sensitivity but not with guilt-driven compensatory sensitivity (Figure 5B). To directly assess whether these brain regions were more involved in guilt-driven or shame-driven compensatory sensitivity, we compared the Pearson correlations between brain activity and the two types of compensatory sensitivities. The results revealed that the left LPFC was more engaged in shame-driven compensatory sensitivity (Z = 1.93, P = 0.053), as its activity showed a marginally stronger positive correlation with shame-driven sensitivity than with guilt-driven sensitivity (Figure 5C). No significant difference was found in the Pearson correlations between the activity of the bilateral IPL and the two types of sensitivities (Supplementary Table 16). For the TP, the effective sample size was too small to yield reliable results (see Methods).

In addition to guilt and shame, other emotion or motivation may also contribute to the transformation of harm and responsibility into compensatory behavior. The parameter κ from our winning computational model ought to capture the combined effects of various psychological processes on compensation, including guilt and shame.

Confirming their associations, the parameter κ significantly correlated to both guilt-driven ($\beta = 2.06$, $T(38) = 5.85$, $P < 0.001$) and shame-driven ($\beta = 2.43$, $T(38) = 2.82$, $P = 0.008$) compensatory sensitivities (Supplementary Table 17). To gain a more comprehensive understanding of the neural basis of compensatory sensitivity, we correlated κ with the whole-brain neural activity. The neural correlates of the parameter κ largely overlapped with those associated with guilt-driven and/or shame-driven compensatory sensitivities, including the dorsomedial prefrontal cortex (DMPFC), supplementary motor area (SMA), left temporal pole (TP), left LPFC (LPFC), bilateral inferior parietal lobe (IPL) (Figures 5A, 5B, and 5D; Supplementary Tables 14 and 18). Notably, κ is also associated with activation in the bilateral anterior insula (aINS).

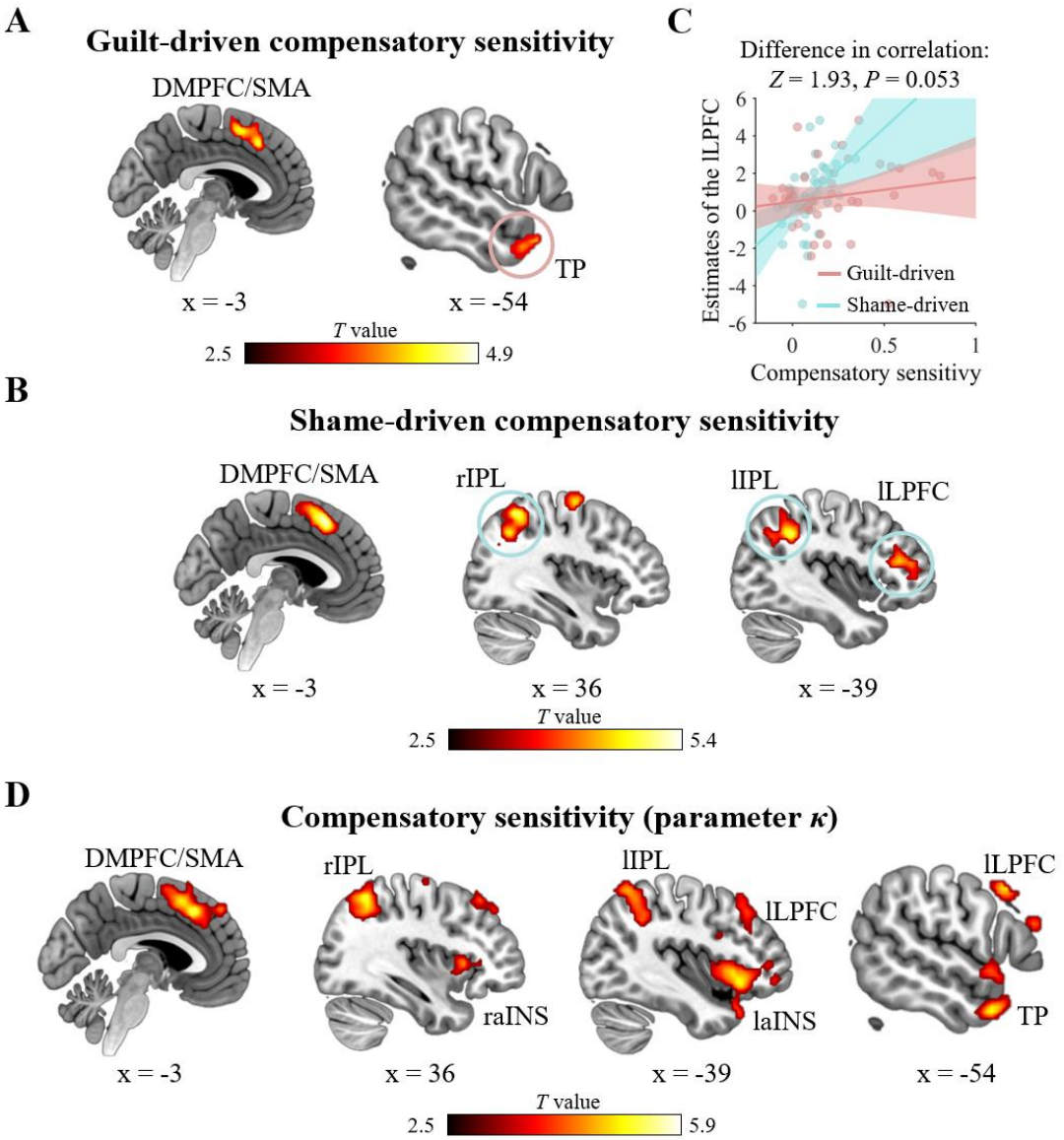

Figure 5. Neural basis of compensatory sensitivity. (**A**) The dorsomedial prefrontal cortex (DMPFC), supplementary motor area (SMA), and temporal pole (TP) showed significant activity associated with guilt-driven compensatory sensitivity. A red circle marked the region showed significant activity associated with guilt-driven compensatory sensitivity but not with shame-driven compensatory sensitivity. (**B**) The DMPFC, SMA, right and left inferior parietal lobe (IPL), and left lateral prefrontal cortex (LPFC) showed significant activity associated with shame-driven compensatory sensitivity. Blue circles marked the region showed significant activity associated with shame-driven compensatory sensitivity but not with guilt-driven compensatory

sensitivity. (**C**) The left LPFC showed a marginally stronger positive correlation with shame-driven sensitivity than with guilt-driven sensitivity. Each line represents the least squares fit with shading showing the 95% confidence interval. (**A, B, D**) The neural correlates of the parameter κ largely overlapped with the regions linked to guilt-driven and shame-driven compensatory sensitivities. The parameter κ is additionally associated with activation in the bilateral anterior insula (aINS). r, right; l, left; Whole-brain FWE-cluster correction at $P < 0.05$ after cluster-forming-threshold at $P < 0.001$.

Neural correlates of trait guilt and compensation. As we observed significant correlations between trait guilt scores and compensatory behavior, we further our investigation by examining their neural correlates. A small-volume correction analysis revealed that repair action tendencies (i.e., a dimension of trait guilt) were significantly associated with the aMCC's responses to the quotient of harm divided by the number of wrongdoers (peak MNI coordinates: [-3, 30, 15]; cluster size: 12 voxels; P_{FWE} = 0.003, small volume corrected; Figure 6A). The results of the aMCC, along with other clusters, retained significance after whole-brain correction (Figure 6B; Supplementary Table 19). Moreover, only the activity in the aMCC's mediated the relationship between repair action tendencies and compensation (indirect effect: $\beta = 0.10$, CI = [0.01, 0.21]) (Figure 6C).

No neural activity showed a significant correlation with negative behavior evaluations (i.e., another dimension of trait guilt), trait shame, trait gratitude, or SVO.

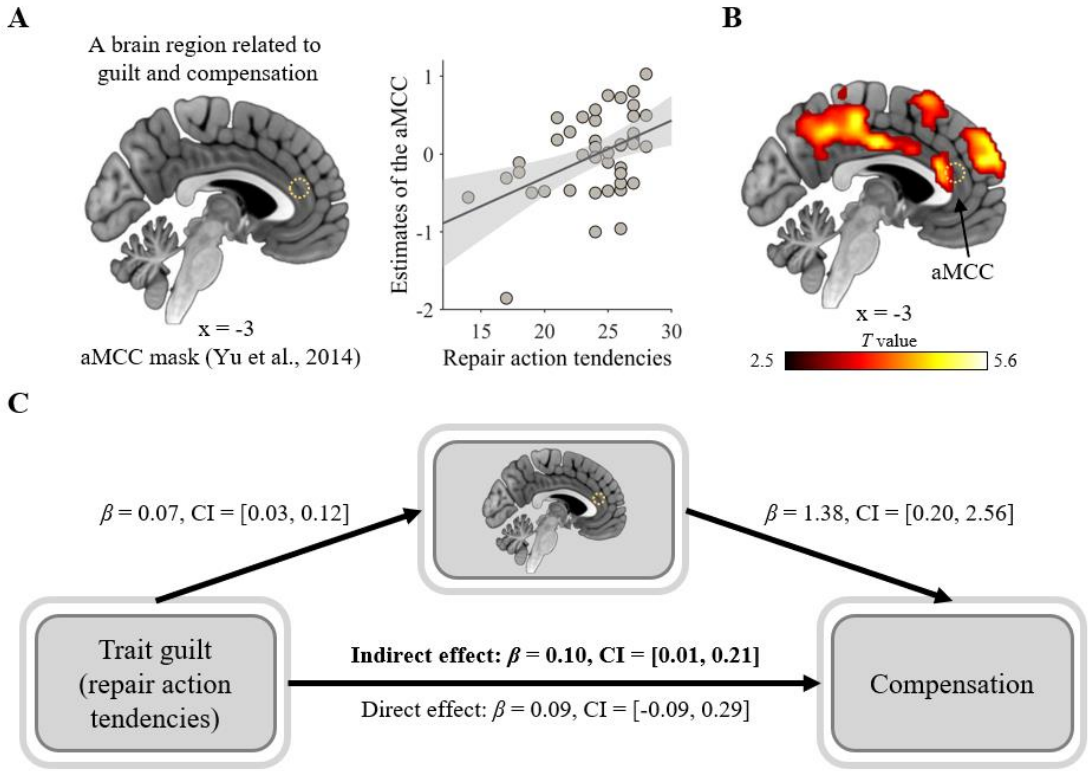

Figure 6. Neural correlates of trait guilt and compensation. (**A**) A small-volume correction analysis showed that participants with higher guilt trait scores (i.e., repair action tendencies) have more positive parametric responses to the quotient of harm divided by the number of wrongdoers in the anterior middle cingulate cortex (aMCC). The scatter plot is for presenting the positive correlation relationship between repair action tendencies and neural responses of the aMCC. The line represents the least squares fit with shading showing the 95% confidence interval. (**B**) The neural finding of the aMCC remained significant after whole-brain correction. Whole-brain FWE-cluster correction at $P < 0.05$ after cluster-forming-threshold at $P < 0.001$. (**A, B**) A circle of dots indicated the position of the aMCC mask. (**C**) The aMCC parametric responses mediated the relationship between repair action tendencies and compensation. β, path coefficient; CI, 95% confidence interval of β; bold font, significant indirect effect.

Discussion

Guilt and shame have long been a focal point of research across various disciplines, including psychology, behavioral economics, neuroscience, and psychiatry (e.g., Chang et al., 2011; Gao et al., 2018; Landers, Sznycer, & Durkee, 2024; Mottershead et al., 2024; Schuster et al., 2021; Sznycer, 2019; Tangney et al., 2007). Extensive research has investigated the psychological and neural activities associated with the experience of guilt and shame to enhance emotion regulation and improve behavioral prediction (Bastin et al., 2016; Miceli & Castelfranchi, 2018; Michl et al., 2014; Piretti et al., 2023; Pulcu et al., 2014; Takahashi et al., 2004; Tangney & Dearing, 2003; Z. Xu et al., 2022; Zhu, Feng, et al., 2019). We extended this research by shifting the focus from the mere states of guilt and shame to their associations with the cognitive antecedents and behavioral consequences. Our findings advance the understanding of the psychological and neural mechanisms that underlie both the formation of guilt and shame and their subsequent transformation into compensatory behavior.

Consistent with previous studies (Gao et al., 2021; Iyer et al., 2007; Koban et al., 2013; Li et al., 2020; Tangney, 1991; Yu et al., 2014), we observed that both harm and responsibility increase individuals' feelings of guilt and shame. Of importance, for the first time, we discovered that harm exerts a stronger effect on guilt than on shame, whereas responsibility has a stronger effect on shame than on guilt. These findings provide empirical support for the proposition that guilt and shame serve distinct social functions. According to functionalist theories (e.g., Baumeister et al., 1994; Gilbert, 1997; Sznycer, 2019; Sznycer et al., 2021), guilt functions to curb the harm imposed on valued others, whereas shame functions to mitigate the cost of reputational damage to oneself. If these theories hold true, guilt should be more sensitive than shame to the information related to harm inflicted on valued others, where shame should be more sensitive than guilt to the information related to reputational damage to oneself (Sznycer et al., 2016). However, few studies have provided such direct evidence. An exception is the study by Landers et al. (2024), which found that information related to harm inflicted on valued others (e.g., liking of a victim) and concerns about reputational damage (e.g., fear of a victim) were respectively predictive of guilt and shame. Notably,

Landers et al. (2024) employed a vignette-based paradigm and assessed participants' guilt and shame using items that reflected the motivational tendencies characteristic of these emotions (e.g., guilt: "I would go to him and apologize for it"; shame: "I would feel like avoiding him for a while"). In contrast, our study utilized a laboratory-based paradigm with real-time decision-making, directly measured participants' state guilt and shame experiences, and test new psychological factors (harm and responsibility). This approach offers greater ecological validity and provides novel evidence (Yu et al., 2024).

Existing findings suggest that guilt is more strongly linked to altruistic behavior than shame is (de Hooge et al., 2007, 2008; Declerck et al., 2014; Gausel & Leach, 2011; Graton & Ric, 2017; Ketelaar & Tung Au, 2003; Tangney et al., 2007; Tangney & Dearing, 2003). Nevertheless, because of methodological limitations, such as failing to compare guilt and shame directly or employing methods that may introduce confounding variables, conclusive evidence has been lacking. Overcoming these limitations, our study demonstrates that while both guilt and shame promote compensation, guilt is more effective in prompting compensatory behavior. The functionalist theories offer a framework for understanding why guilt exerts a stronger effect on compensation than shame does. Guilt corresponds to the adaptive problem of insufficiently valuing others (Sznycer, 2019). To address it, individuals in guilt must bring benefits to those they have harmed—typically through altruistic behavior—to correct the inequity caused by their wrongdoing. In contrast, shame is tied to reputational damage (Sznycer, 2019; Sznycer et al., 2021). Although altruistic behavior can also mitigate shame by demonstrating their social value, individuals may resort to other strategies—such as avoidance or aggression—to protect themselves from potential devaluation by others (de Hooge et al., 2018; Zhu, Xu, et al., 2019).

For a deeper understanding of social emotions, it is crucial to formally model their cognitive operations and investigate the neural underpinnings of these cognitive operations (Yu et al., 2024). Building on this research line, our computational modeling results reveal that individuals in guilt and shame disregard their self-interest, adopt a

compensatory baseline, and mentally distribute harm across all wrongdoers. The findings not only offer a mechanistic explanation at the behavioral (algorithmic) level for guilt- and shame-driven compensatory decision-making, but also deepen the understanding of the phenomenon of responsibility diffusion by offering a formal mathematical formulation and linking it to compensatory behavior (Darley & Latané, 1968).

Notably, in many computational models of social decision-making, self-interest plays a crucial role (e.g., Wu et al., 2024). However, our computational findings suggest that participants disregarded self-interest during compensatory decision-making. A possible explanation is that the personal stakes in our task were relatively small (a maximum loss of 5 Chinese yuan), whereas the harm inflicted on the receiver was highly stigmatized (i.e., an electric shock). Under conditions where the harm is highly salient and the cost of compensation is low, participants may be inclined to disregard self-interest and focus solely on making appropriate compensation.

At the neural level, our findings demonstrate the involvement of the posterior insula and striatum in representing the cognitive antecedents of guilt and shame. Specifically, the activation in these brain regions decreased as the quotient of harm divided by the number of wrongdoers increased. Harm inflicted on the victim, particularly the portion for which a wrongdoer is responsible, creates a sense of inequity between them. Beyond its well-established role in interoceptive awareness (Craig, 2009; Craig, 2011), the pINS has been implicated in processing economic inequity in allocation tasks (Gao et al., 2018; Hsu et al., 2008). For instance, Hsu et al. (2008) reported that pINS activation negatively correlates with the degree of inequity, suggesting that greater inequity elicits lower pINS activation. Our results extend this role of the pINS beyond economic inequity to encompass harm inequity. Given that the striatum is implicated in value computation (Bartra et al., 2013; Crockett et al., 2017; Rilling et al., 2008), we believe that its activity reflects individuals' perception of the loss (i.e., harm) inflicted on the victim. In contrast, no brain region had significant responses to the product of harm and responsibility. Thus, the fMRI and computational modeling findings offer convergent

evidence indicating that individuals are more likely to integrate these cognitive antecedents in a form of quotient.

In addition, no brain region exhibited significant responses to harm. Only the sensorimotor areas showed significant responses to the number of wrongdoers (i.e., the complement of responsibility, 5 – responsibility level). Although the fMRI findings revealed that no brain region associated with social cognition showed significant responses to harm or responsibility, this does not suggest that the human brain encodes only a unified measure integrating harm and responsibility and does not process them as separate entities. Using more fine-grained techniques, such as intracranial electrophysiological recordings, it may still be possible to observe independent neural representations of harm and responsibility.

As to emotion sensitivity, our findings show that individuals who tend to convert responsibility into shame exhibit reduced activation in brain regions associated with other-oriented theory-of-mind processing, specifically the TPJ and STS. The TPJ and STS have been implicated in inferring others' mental states (Schurz et al., 2014). Lower activation in these regions indicates that individuals with higher responsibility-driven shame sensitivity may be less engaged in considering the victim's experiences and thoughts. This aligns with existing research on shame, which, compared to guilt, is associated with less concerns on one's actions on others (Tangney & Dearing, 2003) and weaker activation in the TPJ (Zhu, Feng, et al., 2019).

Regarding compensatory sensitivity, our results show that both individuals with higher guilt-driven and shame-driven compensatory sensitivity have stronger activation in the DMPFC. This region is central to both theory-of-mind processing (Schurz et al., 2014) and self-referential processing (Northoff et al., 2006), playing a crucial role in combining others' thoughts and feelings with one's own (D'Argembeau et al., 2007; Saxe et al., 2006). Zhu, Feng, et al. (2019) has identified the DMPFC's involvement in the experience of both guilt and shame. Our findings here further highlight its role in translating these emotions into compensatory behavior.

We found that the TP's activity is positively related to individuals' guilt-driven compensatory sensitivity. This region is considered as a core part of the theory-of-mind network (Frith & Frith, 2003). Numerous studies suggest that the activation in this region reflect retrieval of both general conceptual knowledge (Lambon Ralph & Patterson, 2008) and social conceptual knowledge (e.g., social rules) (Ross & Olson, 2010; Sugiura et al., 2006; Tsukiura et al., 2010; Zahn et al., 2007). The retrieval of such information likely facilitates understanding others' thoughts and empathizing with their suffering (Olson et al., 2007; Schurz et al., 2014). Empathy, in turn, has been widely established as a significant driver of altruistic behavior, including compensatory behavior (Ding & Lu, 2016; Eisenberg & Miller, 1987). Our findings confirm the involvement of the TP in translating guilt into compensation.

The IPL's activity has a positive correlation with individuals' shame-driven compensatory sensitivity. This region is associated with various non-social and social cognitive functions, including number processing (Pinel et al., 2004; Sandrini et al., 2004), salience processing (Kahnt et al., 2014), and theory-of-mind processing (Igelström & Graziano, 2017; Tusche et al., 2016). Two recent studies provided direct evidence showing that the IPL plays a role in encoding others' benefits during altruistic decision making (Hu et al., 2017, 2021). Our findings about IPL can be explained by its involvement in generating other-regarding motives (Hu et al., 2021) that facilitate the conversion from shame to compensation.

We did not find a significant difference in the correlations between TP activity and guilt-driven versus shame-driven compensatory sensitivities. Similarly, no significant difference was observed in the correlations between IPL activity and shame-driven versus guilt-driven compensatory sensitivities. These findings suggest that neither of these regions plays a domain-specific role in compensation driven by guilt or shame.

In contrast, LPFC activity exhibited a significantly stronger correlation with shame-driven compensatory sensitivity than with guilt-driven compensatory sensitivity, indicating a domain-specific role of the LPFC in shame-related compensation. The LPFC is implicated in cognitive control (Koechlin, 2003) and the optimization of social

decision-making (Buckholtz & Marois, 2012; Feng et al., 2015). Some brain stimulation studies have demonstrated that disrupting LPFC activity impairs individuals' ability to inhibit selfish or aggressive impulses, which can incur social devaluation and punishment from others (Knoch et al., 2006, 2009; Riva et al., 2014). Further research has extended these findings by emphasizing the LPFC's role in strategic social behavior (Ruff et al., 2013; Strang et al., 2015). For instance, Ruff et al. (2013) found that when individuals face the possibility of being punished for selfish behavior, enhancing LPFC activity suppresses selfish impulses and promotes altruistic behavior. However, in the absence of punishment risk, enhancing LPFC activity instead reduces altruistic behavior and promote self-interest. Considering that guilt is typically alleviated through altruistic behavior (Tang et al., 2019), whereas coping strategies for shame are more varied—ranging from altruistic behavior to aggression and avoidance (Sznycer, 2019; Sznycer et al., 2016)—shame appears to be more closely linked to strategic thinking than guilt. This explains why LPFC activity, which is associated with strategic behavior, is more strongly related to shame-driven compensatory sensitivity than to guilt-driven compensatory sensitivity.

The neural correlates of the parameter κ largely overlapped with those linked to compensatory sensitivities driven by guilt and shame. Intriguingly, beyond that, κ also showed a strong association with aINS activity. Insula, known as a key node in the salience network (Uddin, 2015), engages in during experiencing various negative emotions, including sadness (Wagner et al., 2011), disgust (Craig, 2009), guilt (Yu et al., 2014), shame (Piretti et al., 2023; Zhu, Feng, et al., 2019), and indebtedness (Gao et al., 2024). Social neuroscience research highlights the critical role of the aINS in processing norm violations and guiding behavior accordingly (Bellucci et al., 2018; Zinchenko & Arsalidou, 2018). For example, Chang et al. (2011) found that aINS serves to mitigate anticipated negative feelings triggered by norm violations by facilitating individuals' reciprocity toward their partners' investments in a trust game, thereby maintaining adherence to the norm of reciprocity. Consistently, numerous studies on the ultimatum game reveal aINS's involvement in rejecting unfair offers and

upholding the norm of fairness (Feng et al., 2015; Gabay et al., 2014). In the same line, the findings on the involvement of aINS in social conformity also manifest its role in monitoring norm violations and reinforcing adherence to social norms (Berns et al., 2010; Klucharev et al., 2009). Given the involvement of the aINS in various social-affective processes, our findings suggest that the motivation to uphold social norms might directly shape individuals' compensatory behavior or indirectly influence it through emotions beyond guilt and shame, with aINS activity playing a pivotal role in this process.

Interestingly, we found that the sensorimotor areas were associated with the representation of a shame-related cognitive antecedent (i.e., responsibility) and emotional sensitivity. Our findings align with the result from a fMRI meta-analysis, which identified the involvement of sensorimotor regions in processing shame (Piretti et al., 2023). It has been suggested that sensorimotor activation may reflect typical shame-related action tendencies, such as reduced social presence, speech, and movement (Piretti et al., 2023). However, in our study, participants were required to remain completely still throughout the experiment to maintain MRI data quality and were continuously monitored, eliminating the possibility of physical withdrawal. Therefore, the observed sensorimotor activation may reflect motor preparation for subsequent button-press responses associated with compensation rather than a general tendency toward shame-related avoidance. Future studies that permit participants to engage in actual avoidance behaviors could further clarify the role of sensorimotor areas in shame processing.

In line with previous research (Cohen et al., 2011), our findings reveal that both dimensions of trait guilt were significantly associated with compensatory behavior, whereas neither dimension of trait shame exhibited such an association. Furthermore, we found neural responses in the aMCC mediated the relationship between repair action tendencies (one dimension of trait guilt) and compensation. A substantial body of research has revealed that guilt processing consistently activates the aMCC (see a meta-analysis, Gifuni et al., 2017). It is suggested that, in the context of guilt, the aMCC

plays a role in detecting the conflict between social norms and actual behavior, and signaling this conflict via generating negative emotions (Bastin et al., 2016; Gifuni et al., 2017). In addition, Yu et al. (2014) linked the aMCC activity with compensatory behavior. Accordingly, our fMRI findings suggest that individuals with stronger tendency to engage in compensation across various moral violation scenarios (indicated by their repair action tendencies) are more sensitive to the severity of the violation and therefore engage in greater compensatory behavior. However, the neural correlates of negative behavior evaluations (another dimension of trait guilt) were absent. The reasons underlying the non-significant neural finding may be multifaceted. One possibility is that negative behavior evaluations influence neural responses indirectly through intermediate processes not captured in our study (e.g., specific motivational states).

Although previous research has found that trait gratitude and SVO are significantly associated with altruistic behavior in contexts such as donation (Van Lange et al., 2007; Yost-Dubrow & Dunham, 2018) and reciprocity (Ma et al., 2017; Yost-Dubrow & Dunham, 2018), their associations with compensatory decisions in the present study were not significant. This suggests that the effects of trait gratitude and SVO on altruistic behavior are context-dependent and may not predict all forms of altruistic behavior.

This research has several limitations. First, post-task assessments of guilt and shame, unlike in-task assessments, rely on memory and may thus be less precise, although in-task assessments could have heightened participants' awareness of these emotions and biased their decisions. Second, our measures of guilt and shame depend on participants' conceptual understanding of the two emotions. While this is common practice in studies with adult participants (Michl et al., 2014; Wagner et al., 2011; Zhu, Feng, et al., 2019), it may be less appropriate for research involving children. Third, although we aimed to construct a conceptually comprehensive computational model space informed by prior research and our own understanding, it does not encompass all plausible models. Future research is encouraged to explore additional possibilities. Fourth, fMRI cannot establish

causality. Future studies using brain stimulation techniques (e.g., transcranial magnetic stimulation) are needed to clarify the causal role of brain regions in guilt-driven and shame-driven altruistic behavior. Fifth, we did not explicitly measure emotions similar to guilt and shame (e.g., indebtedness), which would have been helpful for understanding their distinct contributions. Sixth, marginally significant results should be viewed cautiously and warrant further examination in future studies with larger sample sizes.

Our study makes several innovative contributions. First, we developed a novel paradigm that effectively elicits guilt and shame at comparable intensities, enabling researchers to systematically explore the associations among guilt, shame, their cognitive antecedents, and behavioral consequences. Future research could combine this paradigm with other cognitive neuroscience methods, such as electroencephalography (EEG) or magnetoencephalography (MEG), and adapt it to investigate additional behaviors linked to guilt and shame, including donation (J. Xu, 2022), avoidance (Shen et al., 2023), and aggression (Velotti et al., 2014). Second, our behavioral findings provide high-quality empirical evidence for functionalist theory, aligning with the contemporary trend of comprehending emotions through their adaptive functions (Landers, Sznycer, & Durkee, 2024; Sznycer et al., 2021). Third, our computational and neural findings offer a clear delineation of the neurocognitive mechanisms underlying guilt and shame. Building on knowledge that harm and responsibility are related to guilt and shame, our results further reveal how these cognitive antecedents are integrated. While previous studies have broadly identified brain regions associated with guilt and shame processing as a whole (Bastin et al., 2016; Gifuni et al., 2017; Piretti et al., 2023), our study advances this understanding by breaking down guilt and shame processing into distinct processes and precisely mapping the neural correlates of each process.

Our study has potential practical implications. The behavioral findings may help counselors understand how cognitive interventions targeting perceptions of harm and responsibility could influence experiences of guilt and shame. The neural findings

highlight specific brain regions (e.g., TPJ) as potential intervention targets for regulating these emotions. Given the close links between guilt, shame, and various psychological disorders (e.g., Kim et al., 2011; Lee et al., 2001; Schuster et al., 2021), strategies to regulate these emotions may contribute to symptom alleviation. Nevertheless, because this study was conducted with healthy adults, caution is warranted when considering applications to other populations.

In conclusion, our findings support the functionalist theory by demonstrating distinct effects of harm and responsibility on guilt and shame, as well as differences in the efficiency with which guilt and shame translate into compensatory behaviors. Notably, harm and responsibility are integrated in a manner consistent with responsibility diffusion prior to influencing guilt-driven and shame-driven compensation. Furthermore, the distinct stages involved in guilt and shame processing correspond to activities in specific neural regions related to value computation, salience processing, theory-of-mind processing, self-referential processing, and cognitive control. By simultaneously providing computational, algorithmic, and neural accounts of guilt and shame (cf. Yu et al., 2024), our study advances the holistic understanding of these emotions, which provides insights into how guilt and shame can be regulated and informs the treatment of guilt- and shame-related mental disorders.

Reference

Abrams, D., & Doosje, B. (2011). The distinct roles of moral responsibility and group-based guilt in response to ingroup wrongdoing. *European Journal of Social Psychology*, *41*, 825–839. https://doi.org/10.1002/ejsp.826

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage*, *76*, 412–427. https://doi.org/10.1016/j.neuroimage.2013.02.063

Bastin, C., Harrison, B. J., Davey, C. G., Moll, J., & Whittle, S. (2016). Feelings of shame, embarrassment and guilt and their neural correlates: A systematic review. *Neuroscience and Biobehavioral Reviews*, *71*, 455–471. https://doi.org/10.1016/j.neubiorev.2016.09.019

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, *115*(2), 243–267. https://doi.org/10.1037/0033-

2909.115.2.243
Bellucci, G., Feng, C., Camilleri, J., Eickhoff, S. B., & Krueger, F. (2018). The role of the anterior insula in social norm compliance and enforcement: Evidence from coordinate-based and functional connectivity meta-analyses. *Neuroscience and Biobehavioral Reviews*, *92*, 378–389. https://doi.org/10.1016/j.neubiorev.2018.06.024
Berndsen, M., van der Pligt, J., Doosje, B., & Manstead, A. S. R. (2004). Guilt and regret: The determining role of interpersonal and intrapersonal harm. *Cognition and Emotion*, *18*(1), 55–70. https://doi.org/10.1080/02699930244000435
Berns, G. S., Capra, C. M., Moore, S., & Noussair, C. (2010). Neural mechanisms of the influence of popularity on adolescent ratings of music. *NeuroImage*, *49*(3), 2687–2696. https://doi.org/10.1016/j.neuroimage.2009.10.070
Blythin, S. P. M., Nicholson, H. L., Macintyre, V. G., Dickson, J. M., Fox, J. R. E., & Taylor, P. J. (2020). Experiences of shame and guilt in anorexia and bulimia nervosa: A systematic review. *Psychology and Psychotherapy: Theory, Research and Practice*, *93*(1), 134–159. https://doi.org/10.1111/papt.12198
Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: Cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, *15*(5), 655–661. https://doi.org/10.1038/nn.3087
Caldwell, J., Meredith, P., Whittingham, K., & Ziviani, J. (2021). Shame and guilt in the postnatal period: A systematic review. *Journal of Reproductive and Infant Psychology*, *39*(1), 67–85. https://doi.org/10.1080/02646838.2020.1754372
Cândea, D.-M., & Szentagotai-Tătar, A. (2018). Shame-proneness, guilt-proneness and anxiety symptoms: A meta-analysis. *Journal of Anxiety Disorders*, *58*, 78–106. https://doi.org/10.1016/j.janxdis.2018.07.005
Čehajić-Clancy, S., Effron, D. A., Halperin, E., Liberman, V., & Ross, L. D. (2011). Affirmation, acknowledgment of in-group responsibility, group-based guilt, and support for reparative measures. *Journal of Personality and Social Psychology*, *101*(2), 256–270. https://doi.org/10.1037/a0023936
Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron*, *70*(3), 560–572. https://doi.org/10.1016/j.neuron.2011.02.056
Clithero, J. A., & Rangel, A. (2013). Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, *9*(9), 1289–1302. https://doi.org/10.1093/scan/nst106
Cohen, T. R., Wolf, S. T., Panter, A. T., & Insko, C. A. (2011). Introducing the GASP scale: a new measure of guilt and shame proneness. *Journal of Personality and Social Psychology*, *100*(5), 947–966. https://doi.org/10.1037/a0022641
Craig, A. D. (2009). How do you feel—now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, *10*(1), 59–70. https://doi.org/10.1038/nrn2555
Craig, A. D. B. (2011). Significance of the insula for the evolution of human awareness of feelings from the body. *Annals of the New York Academy of Sciences*, *1225*(1), 72–82. https://doi.org/10.1111/j.1749-6632.2011.05990.x

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, *20*(6), 879–885. https://doi.org/10.1038/nn.4557

D'Argembeau, A., Ruby, P., Collette, F., Degueldre, C., Balteau, E., Luxen, A., Maquet, P., & Salmon, E. (2007). Distinct regions of the medial prefrontal cortex are associated with self-referential processing and perspective taking. *Journal of Clinical Neuroscience*, *19*(6), 935–944. https://doi.org/10.1162/jocn.2007.19.6.935

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*(4p1), 377. https://doi.org/10.1037/h0025589

de Hooge, I. E., Breugelmans, S. M., Wagemans, F. M. A., & Zeelenberg, M. (2018). The social side of shame: Approach versus withdrawal. *Cognition and Emotion*, *32*(8), 1671–1677. https://doi.org/10.1080/02699931.2017.1422696

de Hooge, I. E., Breugelmans, S. M., & Zeelenberg, M. (2008). Not so ugly after all: When shame acts as a commitment device. *Journal of Personality and Social Psychology*, *95*(4), 933–943. https://doi.org/10.1037/a0011991

de Hooge, I. E., Zeelenberg, M., & Breugelmans, S. M. (2007). Moral sentiments and cooperation: Differential influences of shame and guilt. *Cognition and Emotion*, *21*(5), 1025–1042. https://doi.org/10.1080/02699930600980874

de Hooge, I. E., Zeelenberg, M., & Breugelmans, S. M. (2010). Restore and protect motivations following shame. *Cognition and Emotion*, *24*(1), 111–127. https://doi.org/10.1080/02699930802584466

Declerck, C. H., Boone, C., & Kiyonari, T. (2014). No place to hide: When shame causes proselfs to cooperate. *Journal of Social Psychology*, *154*(1), 74–88. https://doi.org/10.1080/00224545.2013.855158

Eisenberg, N. (2000). Emotion, regulation, and moral development. *Annual Reviews of Psychology*, *51*, 665–697. https://doi.org/10.1146/annurev.psych.51.1.665

Elison, J., Garofalo, C., & Velotti, P. (2014). Shame and aggression: Theoretical considerations. *Aggression and Violent Behavior*, *19*(4), 447–453. https://doi.org/10.1016/j.avb.2014.05.002

Ellsworth, P. C. (2013). Appraisal theory: Old and new questions. *Emotion Review*, *5*(2), 125–131. https://doi.org/10.1177/1754073912463617

Esterman, M., Tamber-Rosenau, B. J., Chiu, Y. C., & Yantis, S. (2010). Avoiding non-independence in fMRI data analysis: Leave one subject out. *NeuroImage*, *50*(2), 572–576. https://doi.org/10.1016/j.neuroimage.2009.10.092

Feng, C., Luo, Y. J., & Krueger, F. (2015). Neural signatures of fairness-related normative decision making in the ultimatum game: A coordinate-based meta-analysis. *Human Brain Mapping*, *36*(2), 591–602. https://doi.org/10.1002/hbm.22649

Feng, C., Tian, X., & Luo, Y. J. (2023). Neurocomputational substrates underlying the effect of identifiability on third-party punishment. *Journal of Neuroscience*, *43*(47), 8018–8031. https://doi.org/10.1523/JNEUROSCI.0460-23.2023

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*(1431), 459–473. https://doi.org/10.1098/rstb.2002.1218

Gabay, A. S., Radua, J., Kempton, M. J., & Mehta, M. A. (2014). The Ultimatum Game and the brain: A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, *47*, 549–558. https://doi.org/10.1016/j.neubiorev.2014.10.014

Gao, X., Jolly, E., Yu, H., Liu, H., Zhou, X., & Chang, L. J. (2024). The psychological, computational, and neural foundations of indebtedness. *Nature Communications*, *15*(1), 1–17. https://doi.org/10.1038/s41467-023-44286-9

Gao, X., Yu, H., Peng, L., Gong, X., Xiang, Y., Jiang, C., & Zhou, X. (2021). The mutuality of social emotions: How the victim's reactive attitude influences the transgressor's emotional responses. *NeuroImage*, *244*, 118631. https://doi.org/10.1016/j.neuroimage.2021.118631

Gao, X., Yu, H., Sáez, I., Blue, P. R., Zhu, L., Hsu, M., & Zhou, X. (2018). Distinguishing neural correlates of context-dependent advantageous-and disadvantageous-inequity aversion. *Proceedings of the National Academy of Sciences*, *115*(33), E7680–E7689. https://doi.org/10.1073/pnas.1802523115

Gausel, N., & Leach, C. W. (2011). Concern for self-image and social image in the management of moral failure: Rethinking shame. *European Journal of Social Psychology*, *41*(4), 468–478. https://doi.org/10.1002/ejsp.803

Ghorbani, M., Liao, Y., Çayköylü, S., & Chand, M. (2013). Guilt, Shame, and Reparative Behavior: The Effect of Psychological Proximity. *Journal of Business Ethics*, *114*(2), 311–323. https://doi.org/10.1007/s10551-012-1350-2

Gifuni, A. J., Kendal, A., & Jollant, F. (2017). Neural mapping of guilt: A quantitative meta-analysis of functional imaging studies. *Brain Imaging and Behavior*, *11*(4), 1164–1178. https://doi.org/10.1007/s11682-016-9606-6

Gilbert, P. (1997). The evolution of social attractiveness and its role in shame, humiliation, guilt and therapy. *British Journal of Medical Psychology*, *70*(2), 113–147. https://doi.org/10.1111/j.2044-8341.1997.tb01893.x

Graton, A., & Ric, F. (2017). How guilt leads to reparation? Exploring the processes underlying the effects of guilt. *Motivation and Emotion*, *41*(3), 343–352. https://doi.org/10.1007/s11031-017-9612-z

Gray, J. R., Braver, T. S., & Raichle, M. E. (2002). Integration of emotion and cognition in the lateral prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(6), 4115–4120. https://doi.org/10.1073/pnas.062381899

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 852–870). Oxford University Press.

Hamann, S., & Canli, T. (2004). Individual differences in emotion processing. *Current Opinion in Neurobiology*, *14*(2), 233–238. https://doi.org/10.1016/j.conb.2004.03.010

Howell, A. J., Turowski, J. B., & Buro, K. (2012). Guilt, empathy, and apology. *Personality and Individual Differences*, *53*(7), 917–922. https://doi.org/10.1016/j.paid.2012.06.021

Hsu, M., Anen, C., & Quartz, S. R. (2008). The right and the good: Distributive justice and neural encoding of equity and efficiency. *Science*, *320*(5879), 1092–1095. https://doi.org/10.1126/science.1153651

Hu, J., Hu, Y., Li, Y., & Zhou, X. (2021). Computational and neurobiological substrates of cost-benefit integration in altruistic helping decision. *Journal of Neuroscience*, *41*(15), 3545–3561. https://doi.org/10.1523/JNEUROSCI.1939-20.2021

Hu, J., Li, Y., Yin, Y., Blue, P. R., Yu, H., & Zhou, X. (2017). How do self-interest and other-need interact in the brain to determine altruistic behavior? *NeuroImage*, *157*, 598–611. https://doi.org/10.1016/j.neuroimage.2017.06.040

Igelström, K. M., & Graziano, M. S. A. (2017). The inferior parietal lobule and temporoparietal junction: A network perspective. *Neuropsychologia*, *105*, 70–83. https://doi.org/10.1016/j.neuropsychologia.2017.01.001

Iyer, A., Schmader, T., & Lickel, B. (2007). Why individuals protest the perceived transgressions of their country: The role of anger, shame, and guilt. *Personality and Social Psychology Bulletin*, *33*(4), 572–587. https://doi.org/10.1177/0146167206297402

Kahnt, T., Park, S. Q., Haynes, J. D., & Tobler, P. N. (2014). Disentangling neural representations of value and salience in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(13), 5000–5005. https://doi.org/10.1073/pnas.1320189111

Ketelaar, T., & Tung Au, W. (2003). The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition and Emotion*, *17*(3), 429–453. https://doi.org/10.1080/02699930143000662

Kim, S., Thibodeau, R., & Jorgensen, R. S. (2011). Shame, guilt, and depressive symptoms: A meta-analytic review. *Psychological Bulletin*, *137*(1), 68. https://doi.org/10.1037/a0021466

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, *61*(1), 140–151. https://doi.org/10.1016/j.neuron.2008.11.027

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, *314*(5800), 829–832. https://doi.org/10.1126/science.1129156

Knoch, D., Schneider, F., Schunk, D., Hohmann, M., & Fehr, E. (2009). Disrupting the prefrontal cortex diminishes the human ability to build a good reputation. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(49), 20895–20899. https://doi.org/10.1073/pnas.0911619106

Koban, L., Corradi-Dell'Acqua, C., & Vuilleumier, P. (2013). Integration of error agency and representation of others' pain in the anterior insula. *Journal of Cognitive Neuroscience*, *25*(2), 258–272. https://doi.org/10.1162/jocn_a_00324

Koechlin, E. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science*, *302*(5648), 1181–1185. https://doi.org/10.1126/science.1088545

Lambon Ralph, M. A., & Patterson, K. (2008). Generalization and differentiation in semantic memory. *Annals of the New York Academy of Sciences*, *1124*, 61–76. https://doi.org/10.1196/annals.1440.006

Landers, M., Sznycer, D., & Al-Shawaf, L. (2024). Shame. In L. Al-Shawaf & T. K. Shackelford (Eds.), *The Oxford Handbook of Evolution and the Emotions*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197544754.013.8

Landers, M., Sznycer, D., & Durkee, P. (2024). Are self-conscious emotions about the self? Testing competing theories of shame and guilt across two disparate cultures. *Emotion*, *24*(5), 1–12. https://doi.org/10.1037/emo0001321

Lazarus, R. S., & Smith, C. A. (1988). Knowledge and appraisal in the cognition-emotion relationship. *Cognition and Emotion*, *2*(4), 281–300. https://doi.org/10.1080/02699938808412701

Lee, D. A., Scragg, P., & Turner, S. (2001). The role of shame and guilt in traumatic events: A clinical model of shame-based and guilt-based PTSD. *British Journal of Medical Psychology*, *74*(4), 451–466. https://doi.org/10.1348/000711201161109

Lewis, H. B. (1971). *Shame and guilt in neurosis.* International Universities Press.

Li, Z., Yu, H., Zhou, Y., Kalenscher, T., & Zhou, X. (2020). Guilty by association: How group-based (collective) guilt arises in the brain. *NeuroImage*, *209*, 116488. https://doi.org/10.1016/j.neuroimage.2019.116488

Losin, E. A. R., Woo, C. W., Medina, N. A., Andrews-Hanna, J. R., Eisenbarth, H., & Wager, T. D. (2020). Neural and sociocultural mediators of ethnic differences in pain. *Nature Human Behaviour*, *4*(5), 517–530. https://doi.org/10.1038/s41562-020-0819-8

Ma, L. K., Tunney, R. J., & Ferguson, E. (2017). Does gratitude enhance prosociality?: A meta-analytic review. *Psychological Bulletin*, *143*(6), 601–635. https://doi.org/10.1037/bul0000103

McCullough, M. E., Emmons, R. A., & Tsang, J. A. (2002). The grateful disposition: A conceptual and empirical topography. *Journal of Personality and Social Psychology*, *82*(1), 112–127. https://doi.org/10.1037/0022-3514.82.1.112

Miceli, M., & Castelfranchi, C. (2018). Reconsidering the differences between shame and guilt. *Europe's Journal of Psychology*, *14*(3), 710–733. https://doi.org/10.5964/ejop.v14i3.1564

Michl, P., Meindl, T., Meister, F., Born, C., Engel, R. R., Reiser, M., & Hennig-Fast, K. (2014). Neurobiological underpinnings of shame and guilt: A pilot fMRI study. *Social Cognitive and Affective Neuroscience*, *9*(2), 150–157.

Moore, C. (2015). Moral disengagement. *Current Opinion in Psychology*, *6*, 199–204. https://doi.org/10.1016/j.copsyc.2015.07.018

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review*, *5*(2), 119–124. https://doi.org/10.1177/1754073912468165

Morawetz, C., & Basten, U. (2024). Neural underpinnings of individual differences in

emotion regulation: A systematic review. *Neuroscience and Biobehavioral Reviews*, *162*, 105727. https://doi.org/10.1016/j.neubiorev.2024.105727

Mottershead, T., Griffiths, A., Nathan, R., & Cole, J. (2024). A mixed-methods systematic review of offence-related shame and/or guilt in violent offenders. *Aggression and Violent Behavior*, *78*, 1–12. https://doi.org/10.1016/j.avb.2024.101989

Murphy, R. O., & Ackermann, K. A. (2014). Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review*, *18*(1), 13–41. https://doi.org/10.1177/1088868313501745

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. (2011). Measuring social value orientation. *Judgment and Decision Making*, *6*(8), 771–781. https://doi.org/10.1017/S1930297500004204

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain-A meta-analysis of imaging studies on the self. *NeuroImage*, *31*(1), 440–457. https://doi.org/10.1016/j.neuroimage.2005.12.002

Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The Enigmatic temporal pole: A review of findings on social and emotional processing. *Brain*, *130*(7), 1718–1731. https://doi.org/10.1093/brain/awm052

Pinel, P., Piazza, M., Le Bihan, D., & Dehaene, S. (2004). Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgments. *Neuron*, *41*(6), 983–993. https://doi.org/10.1016/S0896-6273(04)00107-2

Piretti, L., Pappaianni, E., Garbin, C., Rumiati, R. I., Job, R., & Grecucci, A. (2023). The Neural signatures of shame, embarrassment, and guilt: A voxel-based meta-snalysis on gunctional neuroimaging dtudies. *Brain Sciences*, *13*(4), 1–15. https://doi.org/10.3390/brainsci13040559

Porro, C. A., Francescato, M. P., Cettolo, V., Diamond, M. E., Baraldi, P., Zuiani, C., Bazzocchi, M., & Di Prampero, P. E. (1996). Primary motor and sensory cortex activation during motor performance and motor imagery: A functional magnetic resonance imaging study. *Journal of Neuroscience*, *16*(23), 7688–7698. https://doi.org/10.1523/jneurosci.16-23-07688.1996

Pulcu, E., Lythe, K., Elliott, R., Green, S., Moll, J., Deakin, J. F. W., & Zahn, R. (2014). Increased amygdala response to shame in remitted major depressive disorder. *PLoS ONE*, *9*(1), 1–9. https://doi.org/10.1371/journal.pone.0086900

Rilling, J. K., King-Casas, B., & Sanfey, A. G. (2008). The neurobiology of social decision-making. *Current Opinion in Neurobiology*, *18*(2), 159–165. https://doi.org/10.1016/j.conb.2008.06.003

Riva, P., Romero Lauro, L. J., DeWall, C. N., Chester, D. S., & Bushman, B. J. (2014). Reducing aggressive responses to social exclusion using transcranial direct current stimulation. *Social Cognitive and Affective Neuroscience*, *10*, 352–356. https://doi.org/10.1093/scan/nsu053

Ross, L. A., & Olson, I. R. (2010). Social cognition and the anterior temporal lobes.

NeuroImage, *49*(4), 3452–3462. https://doi.org/10.1016/j.neuroimage.2009.11.012

Ruff, C. C., Ugazio, G., & Fehr, E. (2013). Changing social norm compliance with noninvasive brain stimulation. *Science*, *342*(6157), 482–484. https://doi.org/10.1126/science.1241399

Sandrini, M., Rossini, P. M., & Miniussi, C. (2004). The differential involvement of inferior parietal lobule in number comparison: A rTMS study. *Neuropsychologia*, *42*(14), 1902–1909. https://doi.org/10.1016/j.neuropsychologia.2004.05.005

Saxe, R., Moran, J. M., Scholz, J., & Gabrieli, J. (2006). Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Social Cognitive and Affective Neuroscience*, *1*(3), 229–234. https://doi.org/10.1093/scan/nsl034

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, *42*, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009

Schuster, P., Beutel, M. E., Hoyer, J., Leibing, E., Nolting, B., Salzer, S., Strauss, B., Wiltink, J., Steinert, C., & Leichsenring, F. (2021). The role of shame and guilt in social anxiety disorder. *Journal of Affective Disorders Reports*, *6*, 100208. https://doi.org/10.1016/j.jadr.2021.100208

Shen, B., Chen, Y., He, Z., Li, W., Yu, H., & Zhou, X. (2023). The competition dynamics of approach and avoidance motivations following interpersonal transgression. *Proceedings of the National Academy of Sciences*, *120*(40), e2302484120. https://doi.org/10.1073/pnas.230248412

Strang, S., Gross, J., Schuhmann, T., Riedl, A., Weber, B., & Sack, A. T. (2015). Be nice if you have to--the neurobiological roots of strategic fairness. *Social Cognitive and Affective Neuroscience*, *10*(6), 790–796. https://doi.org/10.1093/scan/nsu114

Sugiura, M., Sassa, Y., Watanabe, J., Akitsuki, Y., Maeda, Y., Matsue, Y., Fukuda, H., & Kawashima, R. (2006). Cortical mechanisms of person representation: Recognition of famous and personally familiar names. *NeuroImage*, *31*(2), 853–860. https://doi.org/10.1016/j.neuroimage.2006.01.002

Sznycer, D. (2019). Forms and functions of the self-conscious emotions. *Trends in Cognitive Sciences*, *23*(2), 143–157. https://doi.org/10.1016/j.tics.2018.11.007

Sznycer, D., Sell, A., & Lieberman, D. (2021). Forms and functions of the social emotions. *Current Directions in Psychological Science*, *30*(4), 292–299. https://doi.org/doi.org/10.1177/096372142110074

Sznycer, D., Tooby, J., Cosmides, L., Porat, R., Shalvi, S., & Halperin, E. (2016). Shame closely tracks the threat of devaluation by others, even across cultures. *Proceedings of the National Academy of Sciences*, *113*(10), 2625–2630. https://doi.org/10.1073/pnas.1514699113

Takahashi, H., Yahata, N., Koeda, M., Matsuda, T., Asai, K., & Okubo, Y. (2004). Brain activation associated with evaluative processes of guilt and embarrassment:

An fMRI study. *NeuroImage*, *23*(3), 967–974. https://doi.org/10.1016/j.neuroimage.2004.07.054

Tang, M., Li, W., Liu, F., & Yuan, B. (2019). The association between guilt and prosocial behavior: A systematic review and meta-analysis. *Advances in Psychological Science*, *27*(5), 773–788. https://doi.org/10.3724/SP.J.1042.2019.00773

Tangney, J. P. (1991). Situational determinan of shame and guilt in young adulthood. *Personality and Social Psychology Bulletin*, *18*(2), 199–206. https://doi.org/10.1177/0146167292182011

Tangney, J. P., & Dearing, R. L. (2003). *Shame and guilt*. Guilford Press.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology*, *58*, 345–372. https://doi.org/10.1146/annurev.psych.56.091103.070145

Tangney, J. P., Wagner, P. E., Hill-Barlow, D., Marschall, D. E., & Gramzow, R. (1996). Relation of shame and guilt to constructive versus destructive responses to anger across the lifespan. *Journal of Personality and Social Psychology*, *70*(4), 797–809. https://doi.org/10.1037/0022-3514.70.4.797

Thoenissen, D., Zilles, K., & Toni, I. (2002). Differential involvement of parietal and precentral regions in movement preparation and motor intention. *Journal of Neuroscience*, *22*(20), 9024–9034. https://doi.org/10.1523/JNEUROSCI.22-20-09024.2002

Tsukiura, T., Mano, Y., Sekiguchi, A., Yomogida, Y., Hoshi, K., Kambara, T., Takeuchi, H., Sugiura, M., & Kawashima, R. (2010). Dissociable roles of the anterior temporal regions in successful encoding of memory for person identity information. *Journal of Cognitive Neuroscience*, *22*(10), 2226–2237. https://doi.org/10.1162/jocn.2009.21349

Tusche, A., Böckler, A., Kanske, P., Trautwein, F. M., & Singer, T. (2016). Decoding the charitable brain: Empathy, perspective taking, and attention shifts differentially predict altruistic giving. *Journal of Neuroscience*, *36*(17), 4719–4732. https://doi.org/10.1523/JNEUROSCI.3392-15.2016

Uddin, L. Q. (2015). Salience processing and insular cortical function and dysfunction. *Nature Reviews Neuroscience*, *16*(1), 55–61. https://doi.org/10.1038/nrn3857

Van Lange, P. A. M., Bekkers, R., Schuyt, T. N. M., & Van Vugt, M. (2007). From games to giving: Social value orientation predicts donations to noble causes. *Basic and Applied Social Psychology*, *29*(4), 375–384. https://doi.org/10.1080/01973530701665223

Velotti, P., Elison, J., & Garofalo, C. (2014). Shame and aggression: Different trajectories and implications. *Aggression and Violent Behavior*, *19*(4), 454–461. https://doi.org/10.1016/j.avb.2014.04.011

Wagner, U., N'Diaye, K., Ethofer, T., & Vuilleumier, P. (2011). Guilt-specific processing in the prefrontal cortex. *Cerebral Cortex*, *21*(11), 2461–2470. https://doi.org/10.1093/cercor/bhr016

Wu, X., Ren, X., Liu, C., & Zhang, H. (2024). The motive cocktail in altruistic

behaviors. *Nature Computational Science*, *4*, 659–676. https://doi.org/10.1038/s43588-024-00685-6

Xu, J. (2022). The impact of guilt and shame in charity advertising: The role of self-construal. *Journal of Philanthropy and Marketing*, *27*(1). https://doi.org/10.1002/nvsm.1709

Xu, Z., Zhu, R., Zhang, S., Zhang, S., Liang, Z., Mai, X., & Liu, C. (2022). Mortality salience enhances neural activities related to guilt and shame when recalling the past. *Cerebral Cortex*, *32*(22), 5145–5162. https://doi.org/10.1093/cercor/bhac004

Yost-Dubrow, R., & Dunham, Y. (2018). Evidence for a relationship between trait gratitude and prosocial behaviour. *Cognition and Emotion*, *32*(2), 397–403. https://doi.org/10.1080/02699931.2017.1289153

Yu, H., Gao, X., Shen, B., Hu, Y., & Zhou, X. (2024). A levels-of-analysis framework for studying social emotions. *Nature Reviews Psychology*, *3*(3), 198–213. https://doi.org/10.1038/s44159-024-00285-1

Yu, H., Gao, X., Zhou, Y., & Zhou, X. (2018). Decomposing gratitude: Representation and integration of cognitive antecedents of gratitude in the brain. *Journal of Neuroscience*, *38*(21), 4886–4898. https://doi.org/10.1523/JNEUROSCI.2944-17.2018

Yu, H., Hu, J., Hu, L., & Zhou, X. (2014). The voice of conscience: Neural bases of interpersonal guilt and compensation. *Social Cognitive and Affective Neuroscience*, *9*(8), 1150–1158. https://doi.org/10.1093/scan/nst090

Yu, H., Koban, L., Chang, L. J., Wagner, U., Krishnan, A., Vuilleumier, P., Zhou, X., & Wager, T. D. (2020). A generalizable multivariate brain pattern for interpersonal guilt. *Cerebral Cortex*, *30*(6), 3558–3572. https://doi.org/10.1093/cercor/bhz326

Zahn, R., Moll, J., Krueger, F., Huey, E. D., Garrido, G., & Grafman, J. (2007). Social concepts are represented in the superior anterior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(15), 6430–6435. https://doi.org/10.1073/pnas.0607061104

Zhong, S., Chark, R., Hsu, M., & Chew, S. H. (2016). Computational substrates of social norm enforcement by unaffected third parties. *NeuroImage*, *129*, 95–104. https://doi.org/10.1016/j.neuroimage.2016.01.040

Zhu, R., Feng, C., Zhang, S., Mai, X., & Liu, C. (2019). Differentiating guilt and shame in an interpersonal context with univariate activation and multivariate pattern analyses. *NeuroImage*, *186*, 476–486. https://doi.org/10.1016/j.neuroimage.2018.11.012

Zhu, R., Jin, T., Shen, X., Zhang, S., Mai, X., & Liu, C. (2017). Relational utility affects self-punishment in direct and indirect reciprocity situations. *Social Psychology*, *48*(1), 19–27. https://doi.org/10.1027/1864-9335/a000291

Zhu, R., Wu, H., Xu, Z., Tang, H., Shen, X., Mai, X., & Liu, C. (2019). Early distinction between shame and guilt processing in an interpersonal context. *Social Neuroscience*, *14*(1), 53–66. https://doi.org/10.1080/17470919.2017.1391119

Zhu, R., Xu, Z., Tang, H., Liu, J., Wang, H., An, Y., Mai, X., & Liu, C. (2019). The

effect of shame on anger at others: Awareness of the emotion-causing events matters. *Cognition and Emotion*, *33*(4), 696–708. https://doi.org/10.1080/02699931.2018.1489782

Zinchenko, O., & Arsalidou, M. (2018). Brain responses to social norms: Meta-analyses of fMRI studies. *Human Brain Mapping*, *39*(2), 955–970. https://doi.org/10.1002/hbm.23895