# Contributions of insula and superior temporal sulcus to interpersonal guilt and responsibility in social decisions

Maria Gädeke[1], Tom Willems[1,2], Omar Salah Ahmed[1,3,4], Bernd Weber[3,4], René Hurlemann[5], Johannes Schultz[3,4]*

1) Masters in Neuroscience Program, University of Bonn, Bonn, Germany
2) Institute of Psychology, University of Bern, Bern, Switzerland
3) Center for Economics and Neuroscience, University of Bonn, Bonn, Germany
4) Institute of Experimental Epileptology and Cognition Research, Medical Faculty, University of Bonn, Bonn, Germany
5) Department of Psychiatry, University of Oldenburg, Oldenburg, Germany

* Corresponding author: johannes.schultz@ukbonn.de

**Keywords**: social decision-making, decisions under risk, computational modelling, neuroimaging

**Abstract**

This study investigated the neural mechanisms involved in feelings of interpersonal guilt and responsibility evoked by social decisions. In two studies (one during fMRI), participants repeatedly chose between safe and risky monetary outcomes in social contexts. Across conditions, each participant chose for both themselves and a partner (*Social* condition), or the partner chose for both themselves and the participant (*Partner* condition), or the participant chose just for themselves (*Solo* condition, control). If the risky option was chosen in the *Social* or *Partner* condition, participant and partner could each receive either the high or the low outcome of a lottery with 50% probability, independently of each other. Participants were shown the outcomes for themselves and for their partner on each trial, and reported their momentary happiness every few trials. As expected, participant happiness decreased following both low lottery outcomes for themselves and for the partner. Crucially, happiness decreases following low outcomes for the partner were larger when the participant rather than their partner had made the choice, which fits an operational definition of guilt. This guilt effect was associated with BOLD signal increase in the left anterior insula. Connectivity between this region and the right inferior frontal gyrus varied depending on choice and experimental condition, suggesting that this part of prefrontal cortex is sensitive to guilt-related information during social choices. Variations in happiness were well explained by computational models based on participant's and partner's rewards and reward prediction errors. A model-based analysis revealed a left superior temporal sulcus cluster that tracked partner reward prediction errors that followed participant choices. Our findings identify neural mechanisms of guilt and social responsibility during social decisions under risk.

## Introduction

Imagine that you go out to dinner with a friend and it is your turn to choose the restaurant. You can choose between two restaurants: one that you both know well, with very predictable food of good quality, and a new restaurant that neither you nor your friend has eaten in before. You decide to try the new one. Unfortunately, while your dish is nice, your friend's turns out to be worse than the known restaurant's dishes. How would you feel? Would you feel differently if it had been your friend's turn to choose the restaurant? Being responsible for such suboptimal outcomes for others can induce a feeling of interpersonal guilt, formally described as a negative emotional response to harming someone with whom one has a positive social bond (Baumeister, 1998; Baumeister et al., 1994; Berndsen et al., 2004; Tangney et al., 2007; Zeelenberg & Breugelmans, 2008). Guilt influences decisions (e.g., Charness & Dufwenberg, 2006), and abnormal sensitivity to guilt is associated with severe social dysfunctions ranging from psychopathy to depression and anxiety, depending on whether sensitivity to guilt is respectively reduced or increased (Tangney et al., 2007).

Several brain regions have been associated with the feeling of guilt: the anterior insula (aIns) (Bastin et al., 2016; Lamm & Singer, 2010; Piretti et al., 2023), the dorsal cingulate cortex (Bastin et al., 2016; Gifuni et al., 2017) and the left temporo-parietal junction (Bastin et al., 2016; Piretti et al., 2023), the ventromedial prefrontal cortex (Krajbich et al., 2009), and with various other regions and networks involved depending on the method used to induce guilt (Bastin et al., 2016; Gifuni et al., 2017). However, we are interested in the mechanisms involved in a specific, as yet understudied aspect of guilt: the kind that can result from everyday choices that expose others to an uncertain outcome. To study this, we built on findings from experimental social decision-making tasks.

Several studies have used games from behavioural economics or perceptual tasks linked to punishment of a partner. For example, participants behaving in accord with an economic definition of guilt aversion during a trust game showed activation in insula, supplementary motor area, dorsolateral prefrontal cortex (dlPFC), and temporal parietal junction (Chang et al., 2011). In another study, partners received painful stimuli when participants made errors during a difficult perception task. These errors evoked activations in the left aIns and dlPFC in the participants (Koban et al., 2013). In a variation of this task, participants could decide to bear a proportion of their partner's pain (Yu et al., 2014). The level of pain taken, indicative of guilt, and activations in anterior middle cingulate cortex and aIns were higher when the pain followed errors made only by the participant rather than by both players. A multivariate reanalysis of these two datasets revealed a neural signature for

guilt (Yu et al., 2020), with key regions including the anterior medial cingulate cortex, insula, inferior frontal gyrus, inferior temporal cortex, thalamus and cerebellum.

One recent study paired the aforementioned perception-and-pain paradigm with dictator game decisions that allowed the participant to compensate for the partner's outcome (Gao et al., 2018). The guilt context increased advantageous-inequity aversion and decreased disadvantageous-inequity aversion, and affected the neural correlates of these inequities (respectively: mentalizing-related, and emotion- and conflict-related regions). Finally, another study contrasted guilt and shame: confederates either experienced economic loss due to bad participant advice (participants experienced guilt), or experienced no loss when participants' bad advice was correctly refused by the confederate (participants experienced shame) (Zhu et al., 2019). Guilt relative to shame activated supramarginal gyrus and temporo-parietal junction as well as orbitofrontal cortex, ventrolateral and dorsolateral prefrontal cortex. Multivariate analyses revealed that guilt could be distinguished from shame based on activation in ventral anterior cingulate cortex and dorsomedial prefrontal cortex.

One landmark study reported on the mechanisms involved in seeking or avoiding responsibility in decisions affecting a group of individuals, and reported involvement of medial prefrontal cortex, aIns and temporo-parietal junction (Edelson et al., 2018). In this elegant study, participants could delegate their choice between a risky and a safe option to a group or decide themselves; in separate conditions, the choice affected payoff either only for the participant or for all group members. Interestingly, most participants displayed responsibility aversion, and this effect could not be explained by guilt, suggesting people associate a psychological cost with assuming responsibility for others' outcomes. However, in many situations one does not have the possibility to delegate a decision to others, and choosing a risky option in such a case may evoke guilt when the outcome for others is negative. The neural mechanisms underlying guilt evoked during such situations of social responsibility are still unknown. These questions are subject of the present study. Similarly to Edelson et al.'s and more recent studies (Arioli et al., 2023; Fareri et al., 2022), our paradigm leveraged a risky choice task with social conditions.

In our experiment, participants were paired with a partner and played an "ice-breaker" game that created a positive social bond between them, increasing the likelihood of feeling empathy and guilt for each other (Baumeister, 1998; Julle-Danière et al., 2020; Loewenstein et al., 1989). Participants or their partner then chose between risky and safe monetary options in three conditions (**Figure 1**). In the *Social* condition, the outcome of

choices affected both participant and partner; participants were and felt responsible for these outcomes because they had agency over their decisions and knew that they could have chosen otherwise (Frith, 2014). We contrasted this condition with similar choices made by a simple expected-value-maximizing algorithm posing for the partner (*Partner* condition), and participant choices only affecting the participant themselves (*Solo* condition, acting as control). Importantly, we assessed the emotional impact of the outcomes of these choices by monitoring participants' happiness every two trials (Rutledge et al., 2014, 2016). This allowed to fit computational models to happiness data and search for networks sensitive to reward prediction errors resulting from participant or partner choices. We ran two experiments, Study 1 outside the MRI scanner and Study 2 during fMRI, with separate groups of participants.

We analysed our behavioural data using several complementary methods: choices were modelled with mixed-effects regressions serving as manipulation checks; risk preferences expressed in choices were assessed using a comprehensive expected utility model as well as with a simpler, more robust "risk premium" approach; and happiness data were fitted, in addition to the computational models, with several linear mixed models to assess the impact of both the participant's and their partner's rewards, the impact of agency and their interactions. Inspired by findings reported in previous neuroimaging of social emotions, we also used several methods to analyse our fMRI data, including conventional methods (both region-of-interest and mass univariate); mixed-effects regression models; computational model-based analyses (inspired by e.g. Konovalov et al., 2021; Rutledge et al., 2014); and functional connectivity (e.g. Edelson et al., 2018; Konovalov et al., 2021). The behavioural modelling is thus complemented by neuroimaging analyses that offer insight about both the activity in regions associated with guilt as well as their place in a wider network, providing an in-depth comprehensive analysis of the mechanisms behind guilt evoked by social responsibility.

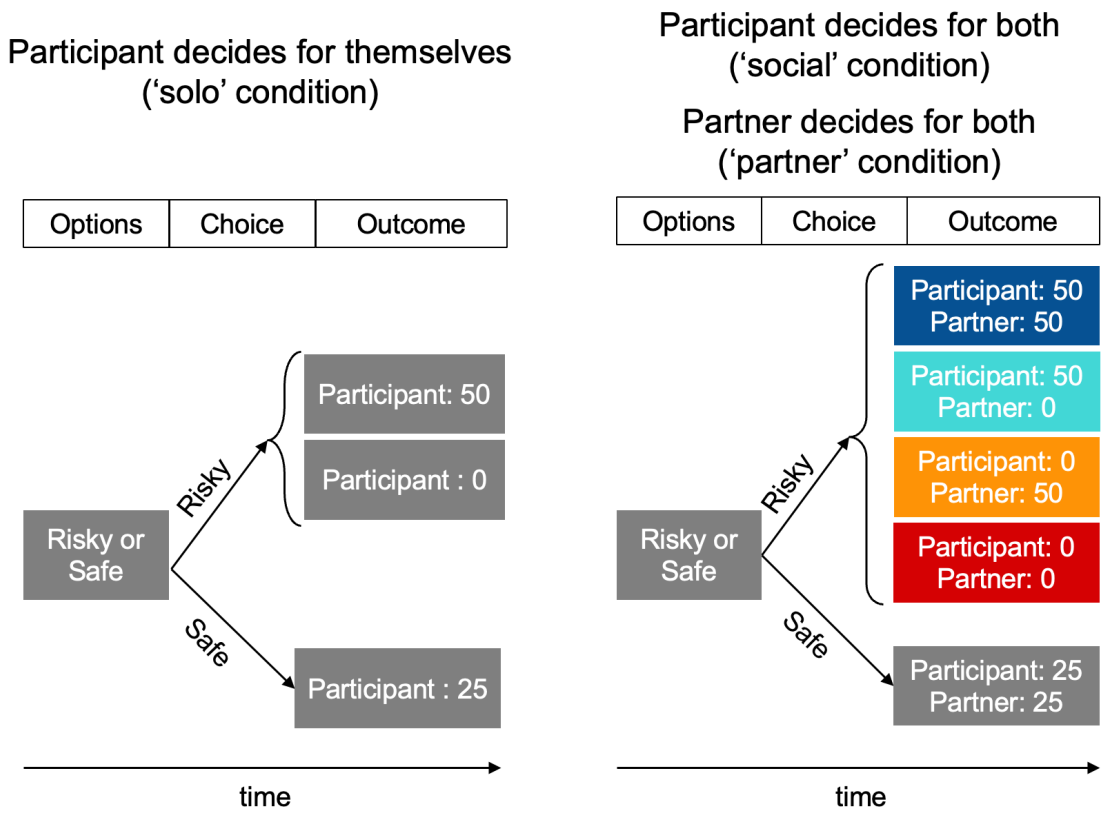


***Figure 1****: Experimental design. In every trial, participants were presented with pairs of monetary options (a safe and a risky option; the risky option was a lottery with equally probable high and low outcomes). There were three conditions: a non-social 'Solo' condition, in which the participant's choice led to an outcome just for themselves (left panel); and two conditions in which choices were made by the participant ('Social' condition) or by their partner ('Partner' condition) and lead to outcomes affecting both players (right panel). Importantly, selecting the risky option in the social or partner conditions lead to the lottery being played out independently for both players; i.e., participant and partner could receive the high or low outcome independently from each other (coloured boxes). Selecting the safe option led to both players receiving an equal outcome.*

## Results

### Behaviour

#### *Overview*

We first verified that participants' choices were reasonable in the *Solo* and *Social* conditions, then assessed whether these choices and the risk preferences they revealed changed depending on whether participants chose just for themselves (*solo*) or for themselves and the interaction partner (*social*). Next, we assessed whether momentary happiness varied with

decision outcomes and whether responsibility influenced this relationship. Lastly, we fitted computational models to the happiness data. While Study 1 (behaviour only) was run before Study 2 (fMRI), we will report the results of both studies together as their results were highly consistent.

*Choices: manipulation check*

As expected, participants' probability of choosing the risky option (lottery) increased with the difference between the expected value of the lottery and the value of the safe option (Study 1: **Figure 2A**, $t(4796) = 9.26$, $p < 3.1e^{-20}$, $\beta = 0.074$, *95% CI* = [0.059 0.090]; Study 2: **Figure 2D**, $t(3829) = 10.62$, $p < 5.3e^{-26}$, $\beta = 0.093$, *95% CI* = [0.075 0.110]; mixed-effects regressions, see Equation 1, detailed results are reported in **Supplementary Table 1**). Participants chose the lottery more often in the *Solo* condition than in the *Social* condition in Study 1 ($t(4796) = 2.54$, $p = 0.011$, $\beta = 0.164$, *95% CI* = [0.038 0.291]), but this difference was not found in Study 2 ($t(3829) = 0.23$, $p = 0.82$, $\beta = 0.015$, *95% CI* = [-0.109 0.138]). There was no significant interaction between the difference in expected values and experimental conditions in either study ($p > 0.52$).

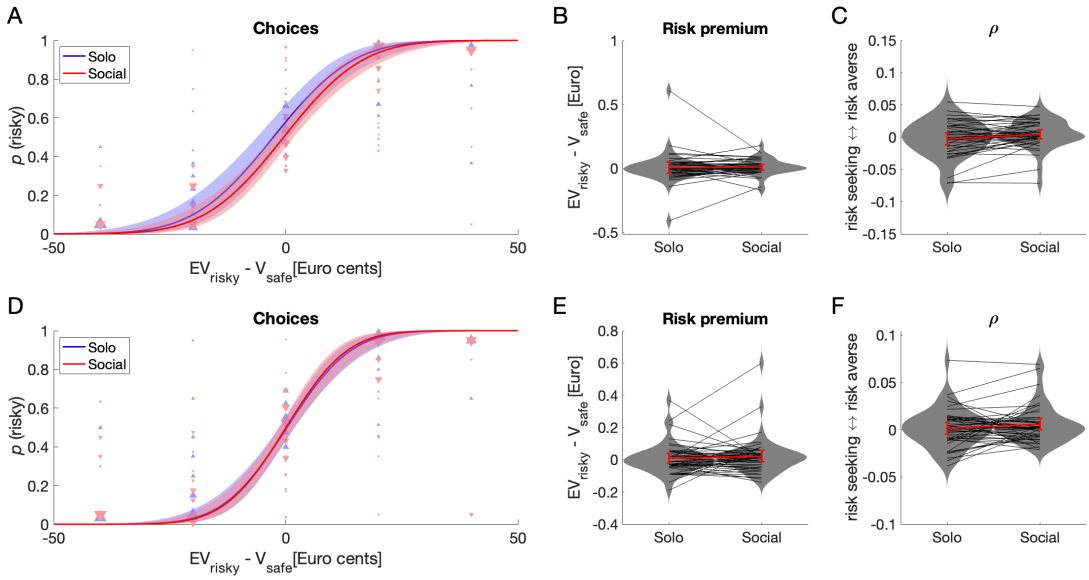


***Figure 2**: Participant choices in Studies 1 (outside fMRI, N = 40) and 2 (inside fMRI, N = 44). **A** and **D**. The probability of choosing the risky option (lottery) in both Solo and Social conditions is well explained by the difference in expected value of the risky and safe choice options ($EV_{risky} - V_{safe}$). Participants chose the risky option slightly more often in the Solo condition than in the Social condition in Study 1 (**A**) but not in Study 2 (**D**). Lines are predicted values of a logit linear mixed regression model fitted to the choice data (see Results). Error areas indicate 95% pointwise confidence intervals for the*

*predicted values. Triangles indicate individual average choice proportions binned by $EV_{risky} - V_{safe}$ value; size of triangle reflects the number of participants contributing to a datapoint. Blue upwards-pointing and red downwards-pointing triangles are data from the Solo and Social conditions, respectively.* ***B*** *and* ***E****. Risk premiums did not differ between Solo and Social conditions.* ***C*** *and* ***F****. Values of the risk aversion parameter ρ in the Solo and Social conditions were broadly consistent with Risk premium values, but showed that participants were slightly more risk averse in the Social than in the Solo condition in Study 1 only (see Results). In panels B, C, E and F, grey lines and markers show individual data, red lines show means with 95% confidence intervals about the means, and grey areas are kernel density plots representing the distribution of the data.*

*Choices: risk preferences*

To better assess whether people's risk preferences varied between the *Solo* and the *Social* condition, we evaluated two additional measures. First, we calculated for each participant a "risk premium", defined as the difference between the expected value of a lottery and its certainty equivalent (See Equation 2 in Methods; positive risk premiums indicate risk aversion). Second, we used an expected utility theory (EUT) approach to calculate a parameter $\rho$ that describes a decision-maker's risk attitude under the assumption of constant absolute risk aversion (see Equation 3 in Methods). We used two measures because fitting the EUT model, the more comprehensive measure that takes into account all choices of a participant, requires many trials to be fitted reliably, a condition that was not satisfied in many participants of both studies; in contrast, the risk premium's single-point measure of risk aversion could be estimated in all participants. Risk premiums did not differ between *Social* and *Solo* conditions (Study 1: **Figure 2B**, $t(39) = 1.53$, $p = 0.134$, *Cohen's d* = 0.24, $BF_{10} = 0.49$; Study 2: **Figure 2E**, $t(43) = -0.21$, $p = 0.84$, $d = -0.03$, $BF_{10} = 0.17$). Turning to the EUT approach, as differences in risk attitudes for gains and losses are well known (Kahneman & Tversky, 1979), we first estimated $\rho$ separately for gain and loss trials. As the number of these types of trials varied across participants, we only obtained reliable estimates for both types of trials in some participants (Study 1: 18 of 40; Study 2: 16 of 44). As $\rho$ did not vary between gain and loss trials (Study 1: $t(17) = 0.21$, $p = 0.84$, $d = 0.05$; Study 2: $t(15) = -0.61$, $p = 0.55$, $d = 0.15$; paired t-test), we then pooled across gain and loss trials, then estimated and compared $\rho$ for the *Solo* and *Social* conditions using paired t-tests. We found that participants were slightly more risk averse in the *Social* than in the *Solo* condition in Study 1 (**Figure 2C**, $t(39) = 2.27$, $p = 0.03$, $d = 0.36$, $BF_{10} = 1.69$) but not in Study 2 (**Figure 2F**, $t(43) = 1.40$, $p =$

0.17, $d$ = 0.21, $BF_{10}$ = 0.41). Risk premium and $\rho$ were highly consistent with each other across participants in both conditions for both studies (Study 1: *Solo* condition: $F(1,38) = 69.9$, $p < 0.001$, $R^2 = 0.65$; *Social* condition: $F(1,38) = 42.0$, $p < 0.001$, $R^2 = 0.55$; Study 2: *Solo* condition: $F(1,41) = 57.9$, $p < 0.001$, $R^2 = 0.59$; *Social* condition: $F(1,41) = 74.2$, $p < 0.001$, $R^2 = 0.64$; these and all subsequent regressions are robust).

In sum, participants showed very similar risk preferences when making decisions affecting only themselves (*Solo* condition) or themselves and their partner (*Social* condition), with a tendency towards higher risk aversion in the *Social* condition in Study 1.

*Momentary happiness: links to reward*

Momentary happiness was assessed every two trials in a similar manner to previous studies by Rutledge and colleagues (Rutledge et al., 2014, 2016). Across all trials, in both studies, participant momentary happiness correlated with rewards obtained in the current trial by the participant and by the partner (Correlations with participant reward, Study 1: $F(1,3598) = 691.5$, $p < 0.001$, $R^2 = 0.16$; partner reward, Study 1: $F(1,2426) = 128.6$, $p < 0.001$, $R^2 = 0.05$; participant reward, Study 2: $F(1,2630) = 650.8$, $p < 0.001$, $R^2 = 0.20$; partner reward, Study 2: $F(1,1735) = 111.8$, $p < 0.001$, $R^2 = 0.06$; **Figure 3A, 3B, 3E** and **3F**). Based on previous findings (Rutledge et al., 2016), we reasoned that participants' momentary happiness would be influenced not only by rewards obtained in the current trial, but also by expected rewards, reward prediction errors, rewards received in previous trials, and differences between rewards obtained by the participant and the partner. Crucial for our research question, we aimed to assess whether responsibility for these rewards, i.e., taking into account whether the rewards occurred following choices made by the participant or the partner, would also influence variations in happiness.

Following Rutledge and colleagues' methodology, which considers that changes in momentary happiness in response to outcomes of a probabilistic reward task are explained by the combined influence of recent reward expectations and prediction errors arising from those expectations, we fitted computational models to each participant's happiness data. In these models, certain rewards, the expected value of chosen lotteries, reward prediction errors and additional outcome parameters are modelled separately with influences that decay exponentially with time (see Methods for equations and details). We first fitted a '*Basic*' model (*Equation 4*) containing regressors coding for certain reward, expected value of chosen lotteries and participant reward prediction errors to evaluate whether these fundamental reward

variables explained variations in momentary happiness. This model explained the data reasonably well (see **Table 1**). Next, based on the notion that inequality between rewards received by the participant and the partner would influence happiness (e.g., Loewenstein et al., 1989), we assessed the fit of an '*Inequality*' model that included one term modelling the difference between the participant and partner outcomes (*Equation 5*). We also assessed the fit of Rutledge and colleagues' (2016) '*Guilt-envy*' model in which advantageous and disadvantageous inequality were modelled separately (*Equation 6*). Next, we tested a '*Responsibility*' model in which separate terms represented the partner's reward prediction errors resulting from choices made by the participant and by the partner (*Equation 7*). Finally, we optimised the Responsibility model by removing the regressor modelling partner reward prediction errors resulting from choices made by the partner, yielding a '*Responsibility Redux*' model (*Equation 8*).

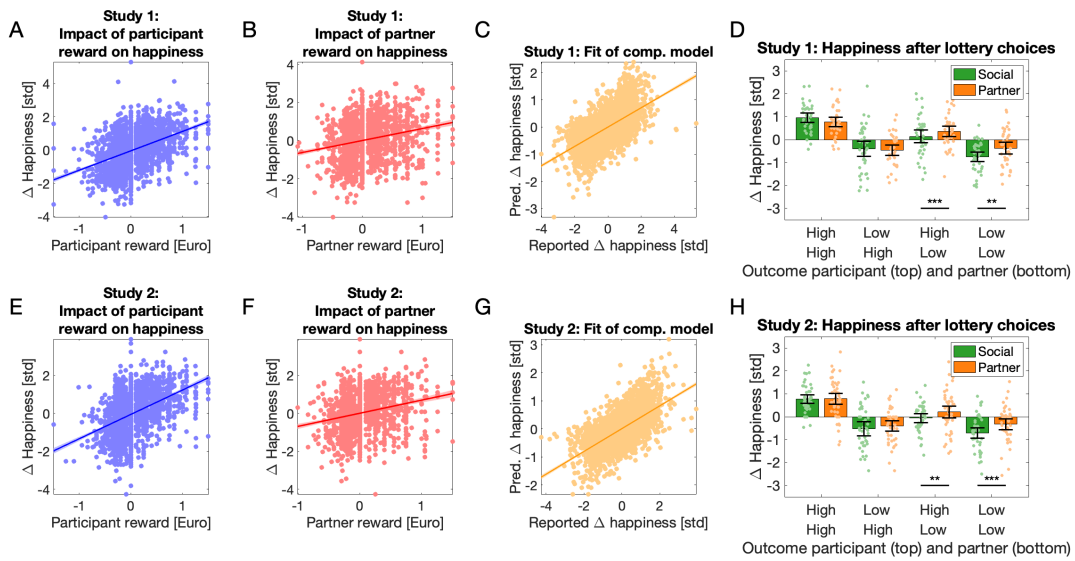


***Figure 3**: Participant momentary happiness in Studies 1(**A-D**) and 2 (**E-H**). Happiness varied with rewards received by the participant (**A** and **E**) and by the partner (**B** and **F**). Each dot is one trial; data are pooled across participants. Lines are fitted regression lines. A computational model taking into account expected, previous and current rewards, reward prediction errors for both participant and partner, and decision-maker (Responsibility Redux model, see Results) predicted the variations in participants' momentary happiness well (**C** and **G**, and Table 1). Changes in momentary happiness after lottery choices in Social and Partner conditions varied with lottery outcome and decision-maker (**D** and **H**). Data were binned according to outcome for each participant and decision-maker (Social = participant chose the lottery, Partner = partner chose the lottery). Crucially, responsibility for low lottery outcomes for the partner decreased participant happiness*

*more than the same outcomes following partner choices (see Results), which fits the definition of interpersonal guilt. In D and H, dots are individual datapoints, the bar indicates the mean, the error bars are 95% confidence intervals about the mean, and the stars indicate the significance of the "guilt effect" (see text): *** p<0.001; ** p<0.01.*

Table 1. Fits of computational models to momentary happiness data

| *Model* | *N param* | *Mean $R^2$* | *Mean $R^2$adj* | *BIC* | *AIC* |
|---|---|---|---|---|---|
| Study 1 | | | | | |
| Basic | 3 | 0.328 | 0.305 | **-1000** | -1399 |
| Inequality | 4 | 0.346 | 0.316 | -916 | -1416 |
| Guilt-envy | 5 | 0.354 | 0.316 | -785 | -1385 |
| Responsibility | 5 | **0.370** | **0.333** | -866 | -1466 |
| Responsibility Redux | 4 | 0.361 | 0.331 | -999 | **-1499** |
| Study 2 | | | | | |
| Basic | 3 | 0.374 | 0.340 | -693 | -1062 |
| Inequality | 4 | 0.394 | 0.350 | -620 | -1080 |
| Guilt-envy | 5 | 0.405 | 0.350 | -491 | -1043 |
| Responsibility | 5 | **0.433** | **0.380** | -616 | -1168 |
| Responsibility Redux | 4 | 0.422 | 0.379 | **-735** | **-1195** |

*BIC, Bayesian Information Criterion. AIC, Akaike's Information Criterion. BIC and AIC values are summed across participants. As in previous work by Rutledge and colleagues (Rutledge et al., 2014, 2016), model fits were performed with individually z-scored happiness ratings. Best values of each variable for each study are highlighted in bold font. For model details, see Results and Methods (Equations 4 to 8).*

All models explained variations in happiness reasonably well (**Table 1** and **Supplementary Materials**). Overall, the *Responsibility* and *Responsibility Redux* (**Figure 3C** and **3G**) models explained the data best (highest $R^2$ / adjusted $R^2$ or lowest AIC / BIC); a likelihood ratio test (*Equation 9*) revealed that the *Responsibility* model fitted better than all the other models, including the *Responsibility Redux* model (Study 1: all $LR \geq 47.36$, $p < 0.0001$; Study 2: all $LR \geq 77.83$, $p < 0.0001$). Weights for certain rewards, expected value of the lotteries, participant reward prediction errors and the forgetting factor $\gamma$ were overall positive across participants, models and studies (Study 1: all $Z > 5.4$, $p < 0.0001$; Study 2: all $Z > 4.9$, $p < 0.0001$, Wilcoxon sign-rank tests were used because data were not normally distributed). Medians of the forgetting factor $\gamma$ varied little across models, ranging from 0.39 to 0.44 in

Study 1 and from 0.42 to 0.46 in Study 2, indicating stable influences of previous trials on happiness across models. Participants' reward prediction errors (sRPE) influenced happiness more than partner's reward prediction errors, whether the latter resulted from the participant or the partner's choices (respectively *social_pRPE* and *partner_pRPE*): weights for *sRPE* were higher than for *social_pRPE* or *partner_pRPE* (Study 1: all $Z > 6.0$, $p < 0.001$; Study 2: all $Z > 3.7$, $p < 0.003$). The stability of these estimated parameters was verified using a parameter recovery procedure (see Methods and **Supplementary Materials**). These results thus replicate the finding that outcomes of risky social decisions influence momentary happiness (Rutledge et al., 2016). Crucially, we find here that the partner's reward prediction errors (*social_pRPE* and *partner_pRPE*) contributed to explaining changes in participants' momentary happiness: the *Responsibility* and *ResponsibilityRedux* models explained the data better than the models without these parameters (see **Table 1**). In particular, the partner's reward prediction errors resulting from the participants' decisions (*social_pRPE*), i.e. those pRPE for which participants were responsible, contributed to explaining our data (weights for *social_pRPE* were greater than 0: *Responsibility model*: Study 1: $Z = 2.85$, $p = 0.004$, Study 2: $Z = 3.26$, $p = 0.001$; *ResponsibilityRedux model*: Study 1: $Z = 2.93$, $p = 0.003$, Study 2: $Z = 3.30$, $p = 0.001$; weights for *social_pRPE* tended to be higher than weights for *partner_pRPE*: *Responsibility model*: Study 1: $Z = 2.14$, $p = 0.033$; Study 2: $Z = 1.41$, $p = 0.16$).

*Momentary happiness: effects of agency, responsibility and guilt*

Next, we assessed whether happiness varied depending on the participant's agency (*Social* + *Solo* vs. *Partner*), and found happiness to be lower when the participant chose, independent of the outcome (Study 1: $t(3600) = -3.92$, $p < 0.0001$, $\beta = -0.14$, 95% CI = [-0.20 -0.07]; Study 2: $t(2870) = -6.07$, $p < 0.0001$, $\beta = -0.24$, 95% CI = [-0.31 -0.16]). This is interesting in itself and may reflect the drive behind responsibility aversion reported by Edelson et al.'s 2018 study: being assigned the role of the decider in a social setting may make people slightly unhappy, perhaps due to "weight of the responsibility". To specifically search for a sign of interpersonal guilt, we analysed happiness values reported after outcomes of lottery choices in the *Social* and *Partner* conditions using conventional linear mixed models (*Equation* 10). Fitting several models revealed that lottery outcomes influenced happiness, and that the identity of the decision-maker played a role (**Supplementary Table 2**). Crucially, the interaction between partner outcome and decision-maker was significant (Study 1: $t(1180) = 3.52$, $p = 0.0004$, $\beta = 0.37$, 95% CI = [0.16 0.58]; Study 2: $t(937) = 2.85$, $p = 0.0045$, $\beta = 0.33$, 95% CI = [0.10 0.56]).

When the partner received the low lottery outcome, participant happiness was lower when they rather than the partner had chosen the lottery (**Figure 3D** and **3H**). When the partner received the low lottery outcome of a participant-chosen lottery, they would presumably feel let down because the participant could have chosen the better safe option. As participants knew this, they were likely to feel "simple guilt" (Battigalli & Dufwenberg, 2007). This behavioural effect (difference in happiness obtained when the partner received low lottery outcomes after participant rather than partner choices) is thus compatible with "simple guilt", and we will thus refer to it as "guilt effect". The "guilt effect" occurred whether the participant received the high lottery outcome (Study 1: $t(39) = -3.58$, $p < 0.001$, $d = 0.56$, $BF_{10} = 32$; Study 2: $t(43) = -2.68$, $p = 0.01$, $d = 0.4$, $BF_{10} = 3.8$) or the low lottery outcome (Study 1: $t(39) = -3.39$, $p = 0.002$, $d = 0.54$, $BF_{10} = 19$; Study 2: $t(43) = -3.58$, $p < 0.001$, $d = 0.54$, $BF_{10} = 33.5$). In Study 2, we will compare individual "guilt effect" values to individual guilt-related brain activation patterns (see end of the section on BOLD signal results). Responsibility for choices did not influence happiness following positive lottery outcomes for the partner (both studies, all $|t| < 1.3$, $p > 0.2$, $BF_{10} < 0.2$).

In sum, in both studies, we found evidence that participants' happiness was influenced by the outcomes of choices for their partner, especially by outcomes resulting from participant decisions, i.e., partner outcomes for which participants were responsible. Within these outcomes, participants felt worse following low lottery outcomes for the partner if those outcomes were consequences of their own choice rather than the partner's, which we interpret as interpersonal guilt.

## BOLD signal

Next we sought to uncover the neural mechanisms associated with our guilt effect and those involved in tracking consequences of participants' decisions on their partner. We analysed the BOLD responses of brain regions engaged during decision-making and at the time of receiving the outcomes of the choice using conventional as well as computational model-based analyses, using the fMRI data collected in Study 2.

### *Brain regions engaged during social decision-making*

Given that our task involved several conditions and successive trial components, we aimed to first replicate previous results related to the neural correlates of decisions under economic risk. We searched for brain regions engaged more when participants chose the risky instead of the safe option,

and found such responses in the bilateral ventral striatum (*Cohen's d* = 0.72 and 0.85 in the left and right clusters, respectively; **Figure 4A** and **Supplementary Table 3**), which replicates previous findings (Cui et al., 2022; Preuschoff et al., 2006). Next, we aimed to identify the brain regions associated with social decision-making under risk, and searched for regions more engaged during decisions in the *Social* condition compared to the *Solo* condition. Three significant clusters of voxels were identified (**Figure 4B** and **Supplementary Table 3**), in the precuneus (*d* = 0.79), the left temporo-parietal junction (TPJ; *d* = 0.59) and the medial prefrontal cortex (mPFC; *d* = 0.54). This also replicates previous findings associating this region with social decisions (e.g., Fareri et al., 2012; Jung et al., 2013; Nicolle et al., 2012; Ogawa et al., 2018; Piva et al., 2019).

To assess whether activations in these five regions were sensitive to the interaction between choice and condition using a sensitive method, we analysed the responses in all the voxels in these regions using linear mixed models (LMMs). Our model included factors *Choice* (*Safe* or *Risky*), *Condition* (*Social*, *Partner* or *Solo*), *Run* (run one or two) and *ROI* (left and right ventral striatum, mPFC, precuneus and TPJ), with all interactions. We also tested the same model without the factor *Run*, but this fitted less well (higher BIC) and was thus discarded. As all factors and interactions were significant (data not shown for sake of brevity), we applied the same models to each region separately (**Supplementary Table 4**). All regions showed a significant interaction between *Condition* and *Choice*, prompting us to examine these responses in detail. As we were particularly interested in the response to *Risky* decisions, we subtracted from it the responses to *Safe* decisions (labelled as "(R-S)" in **Figure 4C**) and compared this difference between (i) *Social* and *Solo* conditions, and (ii) between the *Social* and *Partner* conditions using LMMs. As above, we tested these models both with and without the factor *Run* and associated interaction, and we report the best-fitting model (see **Supplementary Table 4**). Only Precuneus and TPJ showed positive differences in both comparisons (**Figure 4C)**, indicating that these regions were most active when participants chose the lottery in the *Social* condition, the critical situation in which participants assume responsibility over others.

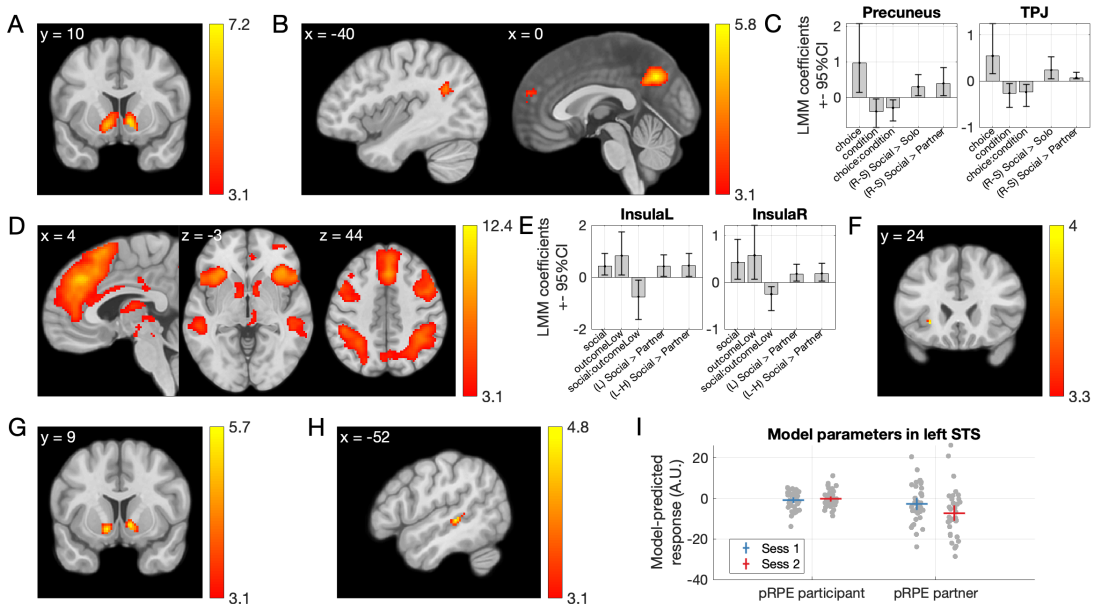


***Figure 4**: BOLD responses. **A.** Regions showing a greater response when participants chose the risky (lottery) rather than the safe option, irrespective of Social or Solo condition. **B.** Regions showing a greater response when participants chose for both themselves and their partner rather than just for themselves (Social > Solo). **C.** Coefficients of linear mixed models indicate that two of these regions, precuneus and TPJ, were most active when participants chose the lottery in the Social condition. R-S indicates results of LMMs based on the Risky-Safe response difference. All coefficients and differences reported are significantly different from 0 (see **Supplementary Table 4**). **D.** Brain regions more active during receipt of the outcomes of lotteries than safe choices (all conditions). **E.** Coefficients of linear mixed models indicated that insula ROIs (see D) mirrored the guilt effect observed in our behavioural data: voxels here responded more to low lottery outcomes (L) for the partner when these resulted from participant's rather than the partner's choices, even when responses to high outcomes were subtracted (L-H). All coefficients and differences reported are significantly different from 0 (see **Supplementary Table 6**). **F.** A mass-univariate, voxel-wise analysis showed a compatible result: A cluster of voxels within the left insula ROI showed higher responses to low lottery outcomes for the partner if these resulted from participant rather than partner choices. **G.** Activation in bilateral ventral striatum explained by a computational model-based regressor coding participant rewards. **H**. Within brain regions sensitive to outcomes of risky choices, one cluster in the left superior temporal sulcus region showed a higher response to partner reward prediction errors resulting from participant rather than partner choices. **I.** Response in this cluster to the computational-model-based regressors coding participant reward prediction resulting from participant and partner choices, for both sessions of the experiment. All results shown survive thresholding at p<0.05 corrected for multiple*

*comparisons at the cluster level, based on a voxel-wise uncorrected threshold of p<0.001. Colours in panels A-I indicate T values. In C and E, bars indicate the estimated coefficients. In C, E and I, error bars are 95% confidence intervals.*

*Brain regions engaged during receipt of outcomes*

Next, we focused on responses during choice outcomes. Cluster of voxels more active during receipt of lottery outcomes than outcomes of safe choices were identified in the bilateral anterior insula, dorsal mPFC (dmPFC), right superior temporal sulcus (STS), bilateral ventral striatum, right dorsolateral prefrontal cortex, and bilateral inferior parietal lobe (**Figure 4D**; for details including effect sizes, see **Supplementary Table 5**). Except for the STS, all these regions have been previously associated with processing of risk and/or ambiguity (Wu et al., 2021).

In our definition, guilt occurs due to responsibility for low lottery outcomes for the partner. To search for regions involved in this situation, we again analysed voxel responses using linear mixed models (LMMs). Our first model included factors *Lottery outcome* (*High* or *Low*), *Condition* (*Social* or *Partner*), *Run* (run one or two) and *ROI* (left and right ventral striatum, left and right insula, left and right parietal cortex, dmPFC, right prefrontal and right middle temporal regions), with all interactions. We also tested the same model without the factor *Run*, which fitted less well and was discarded. As in the analysis of responses to decisions above, all factors and interactions were significant (data not shown for sake of brevity), and we applied the same models to each region separately (**Supplementary Table 6**). To identify regions likely to be involved in the guilt effect, we selected those satisfying two conditions: higher activity in the *Social* compared to the *Partner* condition, and a significant *Social:LowOutcome* interaction. This procedure revealed the insulae (**Figure 4E**) and the right middle temporal cortex (trend significant *Social* vs *Partner* difference in the latter region). To test if these regions responded more to *LowOutcomes* in the *Social* condition, we ran additional models on the responses to *Low lottery outcomes* only (labelled as "(L) Social > Partner" in **Figure 4E**), and on this response minus the response to *High lottery outcomes* (labelled as "(L-H) Social > Partner" in **Figure 4E**); and indeed, the insulae showed the sought-after effect (see **Supplementary Table 6** and **Figure 4E**). Thus, activation in our insula ROIs increased in situations during which participants experienced guilt for low outcomes impacting their partner, compared to similar outcomes resulting from the partner's choices.

To attempt to confirm these results with a classic fine-grained mass-univariate voxel-wise analysis, we searched, within regions responding more to outcomes of risky compared to safe choices, for higher responses to low lottery outcomes for the partner following participant choices compared to the same outcomes resulting from partner choices. We found a weak response in a small cluster within the left anterior insula (peak $T$ = 3.95, $d$ = 0.59, 22 voxels, peak intensity at [-28 24 -4]; **Figure 4F**). Given the documented association between anterior insula and guilt (see Introduction), we proceeded to test whether this result survived correction for family-wise errors due to multiple comparisons restricted to the left anterior insula gray matter [defined anatomically and thus independently from our findings, as the anterior short gyrus, middle short gyrus, and anterior inferior cortex in an anatomical maximum probability map (Faillenot et al., 2017)]. This correction resulted in a $p$ value of 0.024. This result, although it is only a small effect in a small cluster, is consistent with the mixed model analysis reported earlier.

*Neural correlates of responsibility for partner reward prediction errors revealed by computational model-based analysis*

Next, we attempted to explain BOLD responses using predictions of the "*Responsibility*" computational model (see Behaviour / Computational modelling of happiness data, above). We used the model to create expected BOLD responses for each participant (see Methods) and as a manipulation check searched for responses in ventral striatum evoked by participant rewards (O'Doherty et al., 2004, 2007). We found that activation in bilateral ventral striatum indeed increased with the amount of expected certain rewards and the expected values of chosen lotteries (left: $p_{FWE}$ = 0.002, $T$ = 5.63, $d$ = 0.75, $Z$ = 5.41, 110 voxels, peak at MNI [-14 8 -8], right: $p_{FWE}$ = 0.005, $T$ = 5.46, $d$ = 0.70, $Z$ = 5.26, 80 voxels, peak at MNI [10 10 -4], correction for multiple tests applied across the whole brain; **Figure 4G**). We thus used this model to search for voxels responding more to partner reward prediction errors resulting from participant rather than partner choices, within the regions sensitive to outcomes of risky choices. We found this effect in one cluster within the left STS ($p_{FWE}$ = 0.022, $T$ = 4.70, $d$ = 0.53, $Z$ = 4.57, 100 voxels, peak at MNI [-52 -32 0]; **Figure 4H**). Responses to the computational-model-based regressors coding for partner reward prediction errors resulting from participant and from partner choices in this cluster are shown in **Figure 4I**. This finding suggests that this region of the left STS tracks a partner's unexpected outcomes less when they do not follow from the participant's decisions.

*Functional connectivity*

Functional connectivity analyses have revealed differences in networks engaged by social and self-only choices (Jung et al., 2013; Ogawa et al., 2018), interactions between midbrain and anterior cingulate during compensation for guilt (Yu et al., 2014), and links between insula connectivity and responsibility aversion (Edelson et al., 2018). We hypothesized that connectivity with regions that showed guilt- and responsibility-related responses during the outcome phase (see previous paragraph) might change depending on whether participants made decisions for themselves only or for themselves and their partner, and depending on the type of choice (*Safe* or *Risky*). To this end, we used the left insula and left STS as seed regions for whole-brain seed-to-voxel psychophysiological interaction analyses (see Methods), to search for connectivity changes, during the choice phase of the trial, as a function of *Condition* and *Choice* (i.e., voxels with a significant interaction to *Condition* by *Choice*). The first analysis revealed a cluster in the right inferior frontal gyrus (IFG) whose connectivity to the insula (the seed region) was highest when participants made *Risky* choices for themselves and *Safe* choices for both players ($p_{FWE}$ = 0.020, $T$ = 4.34, $d$ = 0.80, $Z$ = 4.21, 115 voxels, peak at MNI [46 16 22]; **Figure 5**). A smaller cluster in the left IFG showed the same effect but did not survive correction for multiple tests ($p_{uncorrected}$ = 0.001, $T$ = 4.08, $Z$ = 3.97, 38 voxels, peak at MNI [-38 -2 26]). The second analysis revealed a smaller cluster in the left inferior frontal gyrus that did not survive corrections for multiple tests, where connectivity with the left STS (the seed region) showed the opposite pattern: connectivity was highest when participants made *Safe* choices for themselves and *Risky* choices for both players ($p_{uncorrected}$ = 0.001, $T$ = 4.44, $Z$ = 4.30, 35 voxels, peak at MNI [-48 14 6]; **Supplementary Figure 2**).

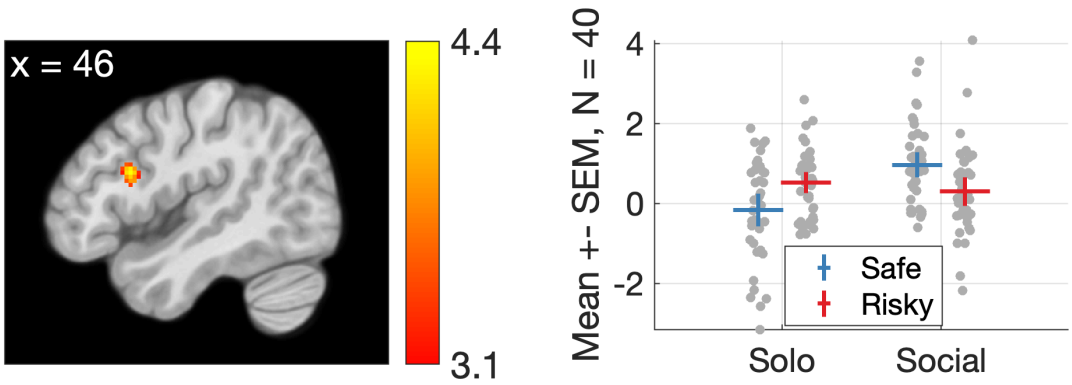


***Figure 5:*** *Changes in functional connectivity between the left anterior insula (seed) and a cluster in the right inferior frontal gyrus at the time of the choice as a function of condition (Social vs. Solo) and choice (Risky or Safe). In the righthand panel, dots are data of individual participants, the markers*

*represent means, and error bars indicate 95 confidence intervals about the mean.*

*Comparison with multivariate neural guilt signature (Yu et al., 2020)*

A recent study by Yu and colleagues re-analysed the results of two previous neuroimaging studies of guilt and obtained a neural multivariate guilt-related brain signature (guilt-related brain signature, GRBS) (Yu et al., 2020). As the GRBS and code to compare neural responses to it is freely available (https://github.com/canlab/Neuroimaging_Pattern_Masks/tree/master/Multivariate_signature_patterns/2019_Yu_Koban_Guilt), we compared this brain signature to the neural responses obtained in our task (response to low partner outcomes resulting from participant vs partner responses). The dot products between individual responses and the GRBS varied between -40.1 and 36.7, but overall these values were positive (mean = 5.22; median = 6.97; sign test: $p$ = 0.017; *Cliff's Delta* = 0.4 = medium effect size; data are not normally distributed). We assessed whether inter-individual differences in these dot product values correlated with the behavioural guilt responses, but did not find a significant association [*Spearman's Rho* = -0.058, $p$ = 0.725].

## Discussion

We report findings from an experiment on social responsibility and guilt in risky economic decisions, and their neural correlates. Being responsible for choosing a lottery that yielded a low outcome for a partner made our participants feel worse than witnessing the same outcome resulting from their partner's choice, which we interpret as interpersonal guilt; although we note that we have not asked participants specifically about which emotion they felt in these situations. Activation in the left anterior insula (aIns) reflected this effect, replicating previous associations between this region and feelings of guilt. Whole-brain activation patterns also resembled a neurometric marker of guilt (Yu et al., 2020). Connectivity between aIns and the right inferior frontal gyrus varied depending on whether participants chose the risky or safe option and whether only the participant or both themselves and the partner were affected by the outcome of this choice, suggesting that this part of prefrontal cortex is sensitive to guilt-related information during social choices. Computational models explained trial-by-trial variations in momentary happiness during the task; the best-fitting model differentiated between partner reward prediction errors resulting from participant and partner choices, indicating that the impact of outcomes for the partner on participants' happiness varied depending on who chose. This confirms the importance of responsibility in determining the emotional

consequences of risky social choices. fMRI analyses based on this computational model identified a left superior temporal sulcus (STS) region responding more to partner reward prediction errors resulting from participant rather than partner choices. This suggests a critical role of the STS in monitoring the consequences of one's risky decisions on others, an essential social cognitive function. A last analysis showed that connectivity between this STS region and the right Inferior Frontal Gyrus that varied depending on condition and choice, suggesting that the information processed in the STS is relayed to the IFG during social decisions. Our findings add to current understanding of the neural mechanisms underlying responsibility for, and guilt evoked by, the outcomes on others of social decisions under economic risk.

We used several approaches to compare choices made for self only or for both participant and partner. Participants made slightly more risk-seeking choices when deciding for themselves than for both themselves and the partner in Study 1, but this difference disappeared in Study 2. The $\rho$ parameter on which this finding in Study 1 is based could only be estimated in a minority of participants due to a relatively low number of trials, which suggests that this finding may not be very reliable. The simpler and more robust method (evaluation of a risk premium) showed no difference in risk aversion across conditions in either study. Overall, we believe that we do not have strong evidence of differences in risk preferences across conditions. Our findings are not unexpected given previous work. At least two studies reported that being responsible for somebody else's payoffs increases risk aversion (Fareri et al., 2022; Pahlke et al., 2015). However, participants in a recent neuroimaging study were more loss averse when choosing for themselves rather than known other people (Arioli et al., 2023). Indeed, recent large meta-analyses comparing risky choices for oneself vs. for others report either no difference in risk preferences (Batteux et al., 2019) or a small shift towards more risky decisions for others (Polman & Wu, 2020), with large variations across studies. Personal closeness to the others for which we decide seems to reduce differences in risk preferences (Fareri et al., 2022; Zhang et al., 2017), as does making decisions for self before making decisions for others (Ifcher & Zarghamee, 2020). In our study, decisions never only affected the partner, which most likely "watered down" any differences in risk attitude in decisions for oneself vs. for someone else. We also created a friendly prosocial environment in which people felt closer to each other than two strangers would (see Experiment Partner in Methods), and interleaved decisions for self only vs. for self and other. Similar risk preferences in decisions for self vs. self and other are thus unsurprising. The fact that participants made similarly risky decisions for themselves or for both

during the fMRI study was to our advantage, because it made the neural signals evoked in these situations more comparable.

Anterior insula (aIns) response, particularly in the left hemisphere, was highest when partners received low lottery outcomes resulting from participants' risky choices. This replicates a large body of evidence associating aIns with feelings of guilt evoked during social decisions (see Introduction). Because we have neither asked our participants specifically what they felt in these situations, nor specifically whether they experienced guilt, we cannot exclude the possibility that they have instead or in addition felt empathy for their partner, a feeling of failure or bad luck, or some other emotion. The aIns region has also been associated with empathy for negative emotions such as disgust (Wicker et al., 2003) or pain (Gu et al., 2012; Lamm et al., 2011), with affective empathy during charitable giving (Tusche et al., 2016), and more generally with emotion awareness (Bird et al., 2010; Gu et al., 2013). These functions might be supported by interoception (Craig, 2002): posterior insula is thought to receive interoceptive information from the body and to pass it on to the anterior insula for integration with sensory, emotional, cognitive, and motivational signals from other regions (Craig, 2009; Rogers-Carter & Christianson, 2019). This integration would allow to represent one's own, as well as estimates of other people's, feelings and bodily states and allow error-based learning based on these (Lamm & Singer, 2010; Rogers-Carter & Christianson, 2019; Singer et al., 2009). Our finding of changing functional connectivity between aIns and prefrontal cortex as a function of decision taken and experimental condition during choice may reflect changing emotional information transfer between these structures depending on choice and social context. How we feel when we witness our decisions' consequences on others is an important signal to consider when attempting to make good social decisions. Unsurprisingly, lesions of aIns are associated with reduced altruistic attitudes (Chau et al., 2018), individuals with higher levels of psychopathic traits show reduced modulation of aIns response to anticipated guilt (Seara-Cardoso et al., 2016), and show reduced guilt aversion (Gong et al., 2019).

Deciding for both self and partner rather than just for oneself evoked increased activations in precuneus, left TPJ and dmPFC, areas classically associated with social cognition (Frith & Frith, 2006; Schurz et al., 2014; Van Overwalle, 2009), and also with feelings of guilt (for meta-analyses, see Gifuni et al., 2017; Piretti et al., 2023). This replicates previous findings of engagement of dmPFC, STS and/or TPJ when making decisions involving others (Jung et al., 2013; Nicolle et al., 2012; Ogawa et al., 2018; Piva et al., 2019). Interestingly, our computational model-based analysis revealed a cluster of voxels in the left STS that responded positively to partner

prediction errors, but only when these resulted from decisions made by the participant instead of the partner. This is congruent with associations between STS and cognitive perspective taking during charitable giving (Tusche et al., 2016), representation of other people's interests during altruistic choice (Hutcherson et al., 2015), or more generally with mentalizing computations during strategic social choice (Carter et al., 2012; Hampton et al., 2008; Hill et al., 2017). The features of our experimental design (direct contrast of similar decisions made by the participant and partner; independent lotteries outcomes for self and other; quantification of the consequences of decisions through variations of momentary happiness) allowed to identify this very specific neural signal indicative of a neural sensitivity to the consequences of one's own actions on others, whether positive or negative. A neural signal coding partner reward prediction errors resulting from one's decisions seems essential to guide social decisions and as a basis for empathic concern for the people influenced by our actions. Anecdotally, we found a weak functional connection between this left STS cluster and left aIns that varied as a function of choice and experimental condition; this is interesting because connections between a region in the left TPJ and aIns have been shown to vary depending on people's tendency to seek or avoid responsibility for others (Edelson et al., 2018). Given the roles of the left STS/TPJ and aIns in guilt and social decisions, it is not surprising to find information exchanges between these structures. More experiments investigating specific computations in which these regions are involved (as in e.g., Charpentier & O'Doherty, 2018; Konovalov et al., 2021; Konovalov & Ruff, 2021) are likely to help understand the neural mechanisms by which guilt and responsibility influence social decision-making.

There are several limitations to our study. The first limitation is that the partner's decisions were in fact taken by an algorithm. This was not communicated to the participants and was thus a case of deception, which is inadmissible in behavioural economics. To our defense, this study was planned as a social neuroscience study and was executed before we had taken into account the practices of behavioural economists. However, this approach did have the advantage to eliminate potentially complex iterative reciprocal influences of decisions and outcomes between the players, which would have led to much more complicated emotional states and decisions that would have been difficult to understand and model. Thus, from an analytical point of view, our approach might have actually allowed a cleaner comparison between decisions and outcomes than if the partner had really made the decisions. The fact that partner outcomes also influenced participants' momentary happiness demonstrates that participants were not emotionally detached from the consequences of their actions on their partner. This finding is of course essential for the validity of our results, and

suggests that the effects could get stronger with more direct interactions. Therefore, although we will abstain from using deceptive practices when pursuing this research, we believe our findings to be valid.

Another limitation is that we have not asked participants to specifically name emotions as they proceeded through the experiment. As a result, we cannot ascertain whether people experienced guilt, shame, regret, disappointment or another specific emotion during the experiment. However, asking about specific emotions is itself associated with several drawbacks: emotion labels might have coloured participant's spontaneous emotional state; participant responses might be influenced by social desirability; thinking about which emotion best characterizes one's mental state takes time and distracts from the decision task; and the list of emotions to choose from is necessarily limited. In any case, the precise naming of emotions was not the aim of our study, instead we relied on a definition of guilt from psychological game theory (Battigalli & Dufwenberg, 2007). Indeed, previous work on social emotions recommends to "study the experiential phenomenology of emotions instead of mere emotion labels", because "as a psychological explanation of human behaviour, the phenomenological experience of an emotion is much more important than the label attached to this experience" (Zeelenberg & Breugelmans, 2008). The "neurometrics" analysis of our guilt-related activation maps showed a significant similarity to a published guilt-related brain signature (Yu et al., 2020). We thus believe that our participants did experience guilt when their choices led to low outcomes for their partner.

Several open questions remain at the end of this study. As discussed above, asking participants directly about which emotions they have felt during the different stages of this task would allow us to link subjective experience with our analytical measures. Testing more participants would allow us to assess the impact of inter-individual variations in personality traits on the experience as well as the behavioural and neural correlates of guilt and responsibility. Using more trials in the experiment would allow separate modelling of risk preferences in gain and loss trials in each experimental condition using expected utility models, and could allow testing whether changes in momentary happiness affect subsequent choices. Varying partner identities (friends, stranger, artificial agent) could reveal the impact of social discounting on guilt and responsibility. Functional connectivity analyses could also be performed for the data obtained in the feedback part of the paradigm. In sum, we believe that this experimental approach lends itself very well to the study of several aspects of social emotions.

## Methods

### Ethics

The studies fulfilled all relevant ethical regulations and were approved by the local ethics committee of the Medical Faculty of the University of Bonn, Germany (approval number 098/18). All participants gave written informed consent and the studies were conducted in accordance with the latest revision of the Declaration of Helsinki. Participants were remunerated for their time (10 Euros/hour) and received game earnings (0-10 Euros).

### Participants

40 healthy participants (14 male, mean age 26.1, range 22 to 31) participated in Study 1 (behaviour only study), and 44 healthy participants (19 male, mean (SD) age = 30.6 (6.5), range 23 to 50) participated in Study 2 (fMRI study). All participants provided written informed consent. Participants were recruited from the local population through advertisements on online blackboards at the University of Bonn and on local community websites, and through flyers posted in libraries, university cafeterias and sports facilities. The number of participants recruited in Study 2 corresponds to the sample size estimated using G*Power 3.1 software (Faul et al., 2007) for a two-way t-test assessing the difference between two means (matched pairs) based on the results of Study 1 (Cohen's d = 0.56), with alpha error = 0.05 and power (1-beta) = 0.95, the required sample size was 44. The fMRI data from 4 participants in Study 2 were excluded from the fMRI data analysis because of excessive head motion (>3mm or >3º).

### Experimental procedure

The design of the experiment was inspired by previous studies investigating how risky choices and their consequences influence momentary happiness (Rutledge et al., 2014, 2016). It was implemented in MATLAB (Version R2016b; RRID:SCR_001622; The MathWorks, Inc., Natick, MA) using the Psychtoolbox extensions (RRID:SCR_002881; http://psychtoolbox.org).

*Experiment partner*

Participants played with a friendly same-sex experiment partner (pairs of participants in Study 1, authors TW and MG in study 2), whom they met before the scan for an introduction session. In this session lasting about 15 minutes, participants played an ice breaker game with their experiment partner, in which both participant and partner took turns in drawing one half of a simple picture while being blindfolded and following verbal instructions given by their game

partner. Encouragements and other positive feedback were given by the partner throughout the task. This icebreaker game led to an agreeable social atmosphere and positive, non-competitive, sympathetic attitude between the participants and the partner. Results of a brief questionnaire indicate that this approach was successful: participants' average ratings of their partners in terms of sympathy, cooperativity, honesty, openness and sociability were all above 8 on a scale of 1 to 10, in both studies (**Supplementary Table 7**).

*Decision task*

Following the icebreaker game, participants in Study 1 performed 3 sessions of a task in which they decided on each trial between a safe and a risky monetary option (**Figure 1**). The risky monetary option was a lottery with two equiprobable outcomes (lottery). There were 3 kinds of trials: decisions by the participant only for themselves (*Solo* condition), decisions by the participant for themselves and the partner (*Social* condition), and decisions by the partner for both themselves and the participant (*Partner* condition). Importantly, when the risky option was selected in the *Social* or *Partner* condition, the lottery was played out independently for the participant and the partner, such that both could receive the higher (*HighOutcome*) or lower outcome (*LowOutcome*), independently from each other. In order to ascertain constant decisions by the partner, the partner's decisions were simulated using a simple algorithm that always selected the option with the highest expected value; i.e., it selected the lottery if EV diff > 0; *EVdiff* = (*high lottery outcome* + *low lottery outcome*) / 2 – *safe outcome*. The amounts earned per trial were determined as in a previous study (Rutledge et al., 2014), as follows. There were 20 mixed trials, 20 gain trials, and 20 loss trials per session. In the mixed trials, participants chose between a safe option of 0 € and a lottery consisting of a gain and a loss amount. Gain amounts were selected randomly out of the following: [15, 25, 40, 55, 75] cents. Loss options were determined by multiplying the gain amount with one randomly selected value out of the following: [-0.2, -0.34, -0.5, -0.64, -0.77, -0.89, -1, -1.1, -1.35, -2]. In gain trials, participants chose between a safe gain out of the following: [10, 15, 20, 25, 30] cents and a lottery with 0 € and a higher gain amount, calculated by multiplying the safe amount with one of the following values: [1.68, 1.82, 2, 2.22, 2.48, 2.8, 3.16, 3.6, 4.2, 5]. In loss trials, participants chose between a certain loss or a lottery with 0 € or a larger loss. Certain loss amounts were one of [-10, -15, -20, -25, -30] cents, and the larger loss amount of the lottery was a multiple of the certain loss amount and one of the following values: [1.68, 1.82, 2, 2.22, 2.48, 2.8, 3.16, 3.6, 4.2, 5]. The position of the safe and risky option on the screen (left or right) was determined randomly on every trial, and the different trial types (*Solo*, *Social* or *Partner* condition; gain, loss and mixed trials) were presented in randomized order, with the constraint that there could not be more than 2

trials of the same condition in a row. Participants had unlimited time to choose between the safe or risky option. Decisions were displayed for 2s and lottery outcomes for 2.5s. Trials were separated in time by an inter-stimulus interval (ISI) of 1 to 2s drawn randomly from a gamma distribution.

*Momentary happiness*

Every two trials, one ISI after the outcome of the previous trial, participants were asked "How happy are you right now?". They could respond by selecting a value on a scale from "very happy" (right) to "very unhappy" (left) by moving a cursor with a button press. The start position of the cursor was the midpoint of the scale, and the scale had 100 selectable options. For analysis, happiness ratings were Z-scored to cancel out effects of different rating variabilities across participants.

*Difference between Studies 1 (behaviour only) and 2 (fMRI study)*

In Study 2, participants performed 2 sessions of the experiment described above inside the fMRI scanner. All parameters were identical except that ISIs varied from 3 to 11s (drawn randomly from a gamma distribution). In Study 2, instead of playing against another same-sex participant, participants played either against experimenter MG or TW depending on sex (participant and partner were always of same sex) who were outside the scanner.

## Statistical Analysis

### *1) General information*

Statistical analysis was performed using Matlab R2024A (RRID:SCR_001622) and the lme4 package (Bates et al., 2022) (RRID:SCR_015654, https://cran.r-project.org/web/packages/lme4/index.html) in R (version 4.2.1, RRID:SCR_001905), and JASP (version 0.16.1, RRID:SCR_015823, https://jasp-stats.org/, JASP Team 2021). fMRI data were analysed using SPM12 software (RRID:SCR_007037; Wellcome Trust Centre for Neuroimaging, London, UK; http://www.fil.ion.ucl.ac.uk/spm). All statistical tests were two-tailed. Bayes factors were calculated using default priors in JASP and express the probability of the data given $H_1$ relative to $H_0$ ($BF_{10}$, values larger than one are in favour of H1). Effect sizes were calculated using standard approaches implemented in Matlab, including the Measures of Effect Size Toolbox (RRID:SCR_014703, https://www.mathworks.com/matlabcentral/fileexchange/32398-measures-of-effect-size-toolbox).

### *2) Decisions*

### a. *Generalized linear mixed logistic regression*

First, we assessed whether choices between the safe and risky options were reasonable and whether these choices varied between *Solo* and *Social* conditions. To do so, we modelled choices with a mixed-effects logistic regression with subject-specific random intercepts $u_{0j} \sim N(0, \delta_u^2)$ and slopes $[u_{1j}, u_{2j}] \sim N(0, \delta_u^2)$ and residual $\varepsilon_{ij} \sim L(0, 1)$, assumed to follow the standard logistic distribution. In this regression model, the choice of the risky option (Chose risky: 1 = Yes, 0 = No) was entered as the dependent variable, and explanatory variables were the difference between the expected value of the lottery minus the value of the safe option [ *EVdiff* = (*high lottery outcome* + *low lottery outcome*) / 2 – *safe outcome* ] and condition (*Cond*: 1 = *Solo*, 2 = *Social*; a categorical variable), as well as the interaction between them (*EVdiff X Cond*). All explanatory variables were mean-centered. The random slopes are designed to account for between-subject heterogeneity:

$$Chose\ risky_{ij} = 1 \text{ if } \beta 0 + (\beta 1 + u_{1j}) * EVdiff_{ij} + (\beta 2 + u_{2j}) * Cond_{ij} + \beta 3 * EVdiffXcond_{ij} + u_{0j} + \varepsilon_{ij} > 0, \text{ and } Chose\ safe_{ij} = 0 \text{ otherwise.}$$

This gives rise to the regression equation:

$$Chose\ risky_{ij} = F(\beta 0 + (\beta 1 + u_{1j}) * EVdiff_{ij} + (\beta 2 + u_{2j}) * Cond_{ij} + \beta 3 * EVdiffXcond_{ij} + u_j)$$

(*Equation 1*)

where $F(x) = 1 / [1 + \exp(-x)]$ is the cumulative distribution function of the standard logistic distribution. The subscript *j* indexes participants, while *i* indexes observations per participant. An observation corresponds to one choice (*Chose risky*: 1 = Yes, 0 = No).

This model was estimated for Study 1 and Study 2 separately, using the package *lme4* in R.

### b. *Estimating risk attitude 1: risk premium*

Next, we estimated risk attitudes by calculating a *risk premium* for each participant and condition, defined as the difference between the expected value of a lottery and its certainty equivalent (positive values indicate risk aversion, negative values risk seeking). The certainty equivalent is the smallest amount an individual would be indifferent to spending on a lottery. We estimated the certainty equivalent of the lotteries by fitting to each participants' decisions in each condition the logistic regression model with residual $\varepsilon_{ij} \sim L(0, 1)$, assumed to follow the standard logistic distribution:

$$Chose\ risky_i = F(\beta 0 + \beta 1 * EVdiff_i)$$

(*Equation 2*)

The *risk premium* for each condition was defined as the EVdiff value corresponding to 50% *risky* choices indicated by the fitted logistic regression model. In both Studies, these *risk premium* values were compared between *Solo* and *Social* conditions using paired *t*-tests.

### c. *Estimating risk attitude 2: CARA model*

Next, we estimated risk attitudes using expected utility theory. In this approach, an agent chooses between two options by comparing their expected utility values. Utility values reflect the agent's subjective values or preferences for a set of choice options, revealed by the agent's choices. Expected utility theory uses utility functions to describe the agents' preferences, including that agent's attitude towards risk. One simple model of risk aversion termed Constant Absolute Risk Aversion (CARA) assumes an exponential utility function:

$$U(x,\rho) = 1 - e(-\rho x)$$

Where $x$ is a monetary amount and $\rho$ is a constant that represents the degree of risk preference. $U$ has an upward slope for perceived gains, indicating higher utility for a larger amount received. For a risk averse individual facing a choice in the domain of gains, $\rho$ is positive and $U$ is concave: a sure amount is preferred over a lottery with the same expected value. For a risk seeking individual facing the same choices, $\rho$ is negative.

Under CARA, the expected utility of a lottery is:

$$EU_{\text{lottery}} = p * U(x_1,\rho) + (1-p) * U(x_2,\rho)$$

Where $x_1$ and $x_2$ are the two possible outcomes of the lottery and $p$ is the probability of obtaining outcome $x_1$.

The certainty equivalent of a lottery under CARA is:

$$CE_{\text{lottery}} = \log(1 - \rho * EU_{\text{lottery}}) / - \rho$$

For $\rho \neq 0$.

*(Equation 3)*

Following an econometric method (von Gaudecker et al., 2011), we estimated $\rho$ by fitting a probit regression with a cumulative Gaussian link function $\Phi$ to choices obtained in the *Solo* and *Social* conditions of each participant:

$$\Phi = 0.5 * (1 + \text{erf}((CE - VS) / (\sqrt{2} * \sigma)));$$

where $CE$ is the certainty equivalent of the lottery, $VS$ the value of the safe alternative option, $\sigma$ is Fechner noise and erf is the error function. The regression was fitted using nonlinear least-squares implemented in Matlab's

nlinfit.m. In both Studies, these $\rho$ values were compared between *Solo* and *Social* conditions using paired *t*-tests.

*3) Momentary happiness*

*a. Computational models of happiness variations*

Following previous work (Rutledge et al., 2014, 2016), we modelled the variations in momentary well-being (happiness ratings) using computational models that considered the influence of recent reward expectations, prediction errors resulting from those expectations, and differences in rewards obtained by the game partners. All models contained separate terms for certain rewards, expected value for lotteries and reward prediction errors, with influences that decayed exponentially over trials. We ran five models: i) a "Basic" model (*Equation 4*; the winning model in Rutledge et al. 2014 in which there was no *Social* condition) with certain rewards (*CR*), expected value of chosen lotteries (*EV*), participant's reward prediction errors (*sRPE*, where s stands for "self") and a forgetting factor that makes events in more recent trials more influential than those in earlier trials (gamma); ii) an "Inequality" model, in which one regressor modelled both disadvantageous and advantageous inequality (*Equation 5*); iii) a "Guilt-envy" model, the winning model in the social task in Rutledge et al., 2016, which contains separate additional regressors for disadvantageous and advantageous inequality (*Equation 6*); iv) a "Responsibility" model in which in addition to the participant's reward prediction errors (*sRPE*) we also modelled the partner's reward prediction errors (*pRPE*, where p stands for partner), with a distinction between *pRPE*s resulting from participant choices (*social_pRPE*) and partner choices (*partner_pRPE*) (*Equation 7*); and v) a "Responsibility redux" model identical to the precedent except for the absence of *partner_pRPE*s, as we realised that this regressor explained little variance (*Equation 8*). The Basic, Inequality and Guilt-envy models are identical to those used in Rutledge et al., 2016.

The equations of these models were as follows.

The equation for the Basic model was:

$$Happiness(t) = w_1 \sum_{j=1}^{t} \gamma^{t-j} CR_j + w_2 \sum_{j=1}^{t} \gamma^{t-j} EV_j + w_3 \sum_{j=1}^{t} \gamma^{t-j} sRPE_j$$

(*Equation 4*)

where t is trial number, $\gamma$ is a forgetting factor ($0 \leq \gamma \leq 1$) that weighs events in more recent trials more heavily than events in earlier trials (exponential decay over time), and weights $w_1$ to $w_3$ capture the influence of the following

different event types: $CR_j$ is the certain reward received when the safe option was chosen instead of the lottery on trial *j*, $EV_j$ is the average reward for the lottery if chosen on trial *j* and $sRPE_j$ is the participant's reward prediction error on trial *j* obtained as a result of choosing the lottery. Terms for unchosen options were set to zero. No constant term was used as ratings were z-scored, resulting in a mean of the data of 0.

As described above, we complemented the Basic model by adding terms representing additional influences on variations in participant happiness. Two models included regressors accounting for inequality, and two models included regressors accounting for partner reward prediction errors. The equation for the first of the inequality models, the "Inequality" model, was as follows:

$$Happiness(t) = w_1 \sum_{j=1}^{t} \gamma^{t-j} CR_j + w_2 \sum_{j=1}^{t} \gamma^{t-j} EV_j + w_3 \sum_{j=1}^{t} \gamma^{t-j} sRPE_j + w_4 \sum_{j=1}^{t} \gamma^{t-j} \max\left(|S_j - P_j|, 0\right)_j$$

(*Equation 5*)

Here, $S_j$ and $P_j$ are the rewards received by the participant and by their experiment partner, respectively, and thus $w_4$ relates to the influence of inequality of any kind on the variation in happiness. The equation of the "Guilt-envy" model was as follows:

$$Happiness(t) = w_1 \sum_{j=1}^{t} \gamma^{t-j} CR_j + w_2 \sum_{j=1}^{t} \gamma^{t-j} EV_j + w_3 \sum_{j=1}^{t} \gamma^{t-j} sRPE_j + w_4 \sum_{j=1}^{t} \gamma^{t-j} \max\left(S_j - P_j, 0\right)_j + w_5 \sum_{j=1}^{t} \gamma^{t-j} \max\left(P_j - S_j, 0\right)_j$$

(*Equation 6*)

Here, $w_4$ relates to advantageous inequality and $w_5$ relates to disadvantageous inequality, modelled separately. The next two models included terms relating to the partner's RPE.

The equation for the "Responsibility" model was:

$$Happiness(t) = w_1 \sum_{j=1}^{t} \gamma^{t-j} CR_j + w_2 \sum_{j=1}^{t} \gamma^{t-j} EV_j + w_3 \sum_{j=1}^{t} \gamma^{t-j} sRPE_j + w_4 \sum_{j=1}^{t} \gamma^{t-j} social_pRPE_j + w_5 \sum_{j=1}^{t} \gamma^{t-j} partner_pRPE_j$$

(*Equation 7*)

In this model, the influence of the partner's RPE resulting from participant choices in the *Social* condition (*social_pRPE*) is modelled separately from the influence of the partner's RPE resulting from partner choices, in the *Partner* condition (*partner_pRPE*). The equation for the last model, "Responsibility redux", was the same as for the Responsibility model (*Equation 7*), except that we omitted the regressor coding partner's RPE resulting from partner choices, as they had a smaller impact on participant happiness. This led to the following equation:

$$Happiness(t) = w_1 \sum_{j=1}^{t} \gamma^{t-j} CR_j + w_2 \sum_{j=1}^{t} \gamma^{t-j} EV_j + w_3 \sum_{j=1}^{t} \gamma^{t-j} sRPE_j + w_4 \sum_{j=1}^{t} \gamma^{t-j} social_pRPE_j$$

(*Equation 8*)

*Likelihood ratio test*

To formally assess which of our models fitted the data best, we supplemented the AIC, BIC, $R^2$ and adjusted $R^2$ values reported in Table 1 with a series of *likelihood ratio tests*: we compared pair-wise the likelihoods of the *Responsibility* model given the data to the likelihoods of all the other models. The equation for this test was the following:

$$LR = 2 * \log (L_0 / L_1);$$

where $L_0$ and $L_1$ are respectively the likelihoods of the simpler and more complex model. The likelihood ratio follows a Chi-square distribution, with degrees of freedom equal to the number of extra parameters in the more complex model. The equation simplifies to:

$$LR = -2 * (\log (L_1) - \log (L_0)$$

*(Equation 9)*

*Parameter recovery*

Stability of the estimated parameters of these models was evaluated by attempting to recover parameters from synthetic data created using each participant's real estimated parameters. After fitting each participants' momentary happiness data (see above), we synthesized new momentary happiness data based on each participant's estimated parameters, added 1SD of noise to the happiness data, fitted the model to these synthetic data,

and repeated this procedure 10 times, for both studies. We then compared these new estimated parameters to the actual parameters from which the synthetic data were generated, as follows: For each parameter, we calculated the mean of each participant's recovered parameters, and regressed these means on the participants' actual parameters (see **Supplementary Figure 1**). The results show that the estimated parameters could be reliably recovered from noisy synthetic data.

*b. Linear mixed model regressions of happiness variations*

To better understand the impact of responsibility for outcomes of risky social choices on happiness, we focused on the happiness reported after outcomes of lottery choices in *Social* and *Partner* trials and evaluated the impact of who chose the lottery. We performed this analysis using a mixed-effects linear regression analysis with residual $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ and subject-specific random intercepts $u_j \sim N(0, \sigma_u^2)$. Specifically, we regressed Z-scored momentary happiness ratings reported after these outcomes on dummy variables representing occurrences (coded as 1 if event present, 0 otherwise) of the participant receiving the high lottery outcome, the partner receiving the high lottery outcome, the participant having chosen the lottery, and interactions between these terms.

$$\text{Happiness(t)} = 1 + \beta_1 sHigh_{ij} + \beta_2 pHigh_{ij} + \beta_3 sDec_{ij} + \beta_4 sHigh^* pHigh_{ij} + \beta_5 sHigh^* sDec_{ij} + \beta_6 pHigh^* sDec_{ij} + u_j + \varepsilon_{ij}.$$

(*Equation 10*)

*sHigh* and *pHigh* had a value of 1 if the participant or the partner (respectively) received the high lottery outcome and 0 otherwise, *sDec* had a value of 1 if the participant chose the lottery. All models were estimated for Study 1 and Study 2 separately, using the *lme4 package in R*. In both studies, the model reported in *Equation 10* fitted the data significantly better ($p<2e-5$) than the simpler models without interactions (higher total and adjusted $R^2$, see **Supplementary Table 2**), but not significantly worse than the model with all interactions ($p>0.5$; tested with the ANOVA function in R). The regressors of the different models tested and their relative effects are reported in **Supplementary Table 2**.

**fMRI data acquisition and pre-processing**

Imaging data were collected on a 3T Siemens TRIO MRI system (Siemens AG, Erlangen, Germany) with a Siemens 32-channel head coil. Functional

data were acquired using a T2* echo-planar imaging (EPI) BOLD sequence, with a repetition time (TR) of 2500 ms, an echo time (TE) of 30 ms, 37 slices with voxel sizes of 2 x 2 x 3 mm, a flip angle of 90˚, a field of view of 192 mm and PAT two acceleration. To exclude participants with apparent brain pathologies and facilitate normalisation of the functional data, a high-resolution T1-weighted structural image was acquired, with a TR of 1660 ms, a TE of 2540 ms, 208 slices with voxel sizes of 0.8 x 0.8 x 0.8 mm and a field of view of 256 mm. Data were then pre-processed and analysed as in previous studies (e.g., Schultz et al., 2019) using standard procedures in SPM12. The first five volumes of each functional time series were discarded to allow for T1 signal equilibration. The structural image of each participant was coregistered with the mean functional image of that participant. Functional images were corrected for head movement between scans by a 6-parameter affine realignment to the first image of the time-series and then re-realigned to the mean of all images. The structural scan of each participant was spatially normalised to the current Montreal Neurological Institute template (MNI305) by segmentation and non-linear warping to reference tissue probability maps in MNI space, and the resulting normalisation parameters were applied to all functional images which were then resampled at 2 x 2 x 2 mm voxel size, then smoothed using an 8 mm full width at half maximum Gaussian kernel. Time series were de-trended by the application of a high-pass filter (cut-off period, 128 s).

**fMRI data analysis**

The pre-processed functional data were analysed using a standard two-stage approach based on the general linear model (GLM) implemented in SPM12 running in MATLAB, with corrections for multiple comparisons. Individual participants' data were modelled with fixed effects models, and their summary data were entered in random effects models for group statistics and inferences at the population level.

*Fixed-effects models*

The fixed effects models implemented a mass univariate analysis applied to the pre-processed data. The first GLM (GLM1) included the following event types for each session: presentation of choice options, safe choice, risky choice (=lottery), receipt of outcome of safe option, all modelled separately for the *Social*, *Partner* and *Solo* conditions; then receipt of high lottery outcome by the partner, separately coded for the *Social* and *Partner* conditions, and receipt of high lottery outcome by the participant; then receipt of low lottery outcome by the partner, separately coded for the *Social*

and *Partner* conditions, and receipt of low lottery outcome by the participant; a final regressor modelled the button presses that occurred when reporting momentary happiness. All events were modelled as stick functions at the time of occurrence and convolved by SPM using the canonical hemodynamic function. In addition, we created another GLM (GLM2) with regressors designed to identify brain regions whose activation reflected the variables of the best-fitting computational model (see above). This model contained regressors coding the receipt of certain rewards (*CR*), the expected value of the chosen option (*EV*), the participant's reward prediction error (*sRPE*), the partner RPE resulting from decisions made by the participant (*social_pRPE*) and the partner RPE resulting from decisions made by the partner (*partner_pRPE*). Those regressors were applied as follows to each experimental session of each participant: 1) fitting the computational model to the momentary happiness data; 2) creation of a series of stick functions representing the time of occurrence of the modelled events during the scan, with stick magnitude determined by the fitted value calculated by the model; 3) convolution of the series of stick functions with the canonical BOLD function (as implemented in SPM). The 5 computational regressors were entered as covariates into each participant's model, which also contained one regressor modelling the button presses that occurred when reporting momentary happiness. All GLMs contained six motion-correction parameters included as regressors of no-interest to account for motion-related artifacts. Regression coefficients (parameter estimates) were estimated for each regressor in each voxel of each participant's brain.

*Psychophysiological Interaction (PPI) analysis*

We used the gPPI toolbox (McLaren et al., 2012) for SPM to run seed-to-voxel functional connectivity analyses. We ran two models for each participant, both based on GLM1, with the following seeds: the left insula cluster more sensitive to the outcomes of *Risky* vs. *Safe* choices, yielding GLM3 (see **Figure 4D** and **E**), and the left STS cluster responding more to *social_pRPE* than *partner_pRPE*, yielding GLM4 (**Figure 4H**). Seeds were identical for all participants. PPI models thus contained the same regressors as GLM1, plus the interaction between the first eigenvariate of the signal in the seed region and the regressors coding the experimental conditions of interest: *Safe* and *Risky* choice in *Social*, *Partner* and *Solo* conditions.

*Random-effects models*

Individual parameter estimate maps of the responses to the experimental conditions were assessed in random-effects tests to evaluate effects at the

level of the group of participants. These tests were second-level models based on individual parameter estimate maps. We tested six models: a full factorial model to investigate the response during decisions [factors *Choice* (*Safe* or *Risky*) and *Condition* (*Social* or *Solo*), based on GLM1 estimate maps]; a full factorial model to investigate the response during the receipt of outcomes [factors *Outcome* (*Safe*, *HighOutcome* or *LowOutcome*) and *Condition* (*Social*, *Partner* or *Solo*), based on GLM1 estimate maps]; a paired t-test contrasting negative lottery outcomes for the partner resulting from participant or partner choices (based on GLM1 estimate maps); a model to investigate the response to the computational regressors (*CR*, *EV*, *sRPE*, *social_pRPE* and *social_pRPE*, based on GLM2 estimate maps); and two full factorial models to investigate functional connectivity during decisions [factors *Choice* (*Safe* or *Risky*) and *Condition* (*Social* or *Solo*), based on estimate maps of GLM3 and GLM4].

*Linear Mixed models*

Similar to the behavioural data, we analysed GLM1 parameter estimates from voxels of specific regions of interest (see Results) using mixed-effects linear regressions with residual $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ and subject-specific random effects $u_j \sim N(0, \sigma_u^2)$. These models were implemented in Matlab and fitted with the fitlme.m function using restricted maximum likelihood estimation. We regressed parameter estimates on dummy variables representing variables of interest. For responses during the decision phase of the experiment, our models included factors *Choice* (*Safe* or *Risky*), *Condition* (*Social*, *Partner* or *Solo*), *Run* (run one or two) and *ROI*, with all interactions. For responses during the outcome phase, our models included the factors *Social* (value of 1 for the *Social* condition and 0 for the *Partner* condition), *LowOutcome* (value of 1 for *Low lottery outcome* and 0 for *High lottery outcome)*, *Run*, with all interactions. Models without the factor *Run* were tested as well, and the best-fitting models were tested further. In all models, *Subject* was the only random factor (random intercept). To better understand significant interaction effects, we ran subsequent analyses on contrasts of parameter estimates (see Results and Supplementary materials).

*Thresholding*

All clusters reported survive a significance threshold of $p<0.05$ with family-wise error correction for multiple comparisons (FWE) across the whole brain or a smaller volume when explicitly mentioned, based on an uncorrected voxel-wise (cluster-forming) threshold of $p<0.001$.

## Data Availability

All behavioural data and group fMRI results maps needed to reproduce all the results and the figures of this study are freely available at https://github.com/BonnSocialNeuroscienceUnit/ResponsibilityExperiment. The entire raw dataset (anonymous data) is publicly available on OpenNeuro (https://doi.org/10.18112/openneuro.ds005588.v1.0.0).

## Code Availability

All the code used to replicate the experiment, analyse the data and produce the figures of this study are freely available at https://github.com/BonnSocialNeuroscienceUnit/ResponsibilityExperiment.

## Author contributions

Maria Gädeke: Investigation, Data curation, Software, Methodology, Writing – original draft, Writing – review and editing; Tom Willems: Investigation, Data curation, Software, Methodology; Omar Salah Ahmed: Software, Methodology; Bernd Weber: Resources, Funding acquisition; René Hurlemann: Conceptualization, Resources, Funding acquisition; Johannes Schultz: Conceptualization, Methodology, Data curation, Formal analysis, Visualization, Software, Supervision, Project administration, Writing – original draft, Writing – review and editing.

## Acknowledgements

We would like to thank R. Rutledge for discussions and for providing the reinforcement learning code using in their previous publication (Rutledge et al., 2016), H. Gerhardt for advice and help using linear mixed models, logistic regressions and economic models, as well as M. Heinrichs, T. Kalenscher, and C. Ruff for discussions. This work was supported by the Open Access Publication Fund of the University of Bonn.

## References

Arioli, M., Basso, G., Baud-Bovy, G., Mattioni, L., Poggi, P., & Canessa, N. (2023). Neural bases of loss aversion when choosing for oneself versus known or unknown others. *Cerebral Cortex*, *33*(11), 7120–7135. https://doi.org/10.1093/cercor/bhad025

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 & Biobehavioral Reviews*, *71*, 455–471. https://doi.org/10.1016/j.neubiorev.2016.09.019

Bates, D., Maechler, M., Bolker [aut, B., cre, Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., & simulate.formula), P. N. K. (shared copyright on. (2022). *lme4: Linear Mixed-Effects Models using 'Eigen' and S4* (Version 1.1-29) [Computer software]. https://CRAN.R-project.org/package=lme4

Batteux, E., Ferguson, E., & Tunney, R. J. (2019). Do our risk preferences change when we make decisions for others? A meta-analysis of self-other differences in decisions involving risk. *PLOS ONE*, *14*(5), e0216566. https://doi.org/10.1371/journal.pone.0216566

Battigalli, P., & Dufwenberg, M. (2007). Guilt in Games. *American Economic Review*, *97*(2), Article 2. https://doi.org/10.1257/aer.97.2.170

Baumeister, R. F. (1998). Inducing guilt. In *Guilt and children.* (pp. 127–138). Academic Press. https://doi.org/10.1016/B978-012148610-5/50007-2

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An Interpersonal Approach. *Psychological Bulletin*, *11*(2), 243–267.

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

Bird, G., Silani, G., Brindley, R., White, S., Frith, U., & Singer, T. (2010). Empathic brain responses in insula are modulated by levels of alexithymia but not autism. *Brain*, *133*(5), 1515–1525. https://doi.org/10.1093/brain/awq060

(Bud) 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

Carter, R. M. K., Bowling, D. L., Reeck, C., & Huettel, S. A. (2012). A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science*, *336*(6090), 109–111. https://doi.org/10.1126/science.1219681

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

Charness, G., & Dufwenberg, M. (2006). Promises and Partnership. *Econometrica*, *74*(6), 1579–1601. https://doi.org/10.1111/j.1468-0262.2006.00719.x

Charpentier, C. J., & O'Doherty, J. P. (2018). The application of computational models to social neuroscience: Promises and pitfalls. *Social Neuroscience*, *13*(6), Article 6. https://doi.org/10.1080/17470919.2018.1518834

Chau, A., Zhong, W., Gordon, B., Krueger, F., & Grafman, J. (2018). Anterior insula lesions and alexithymia reduce the endorsements of everyday altruistic attitudes. *Neuropsychologia*, *117*, 428–439. https://doi.org/10.1016/j.neuropsychologia.2018.07.002

Craig, A. D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Reviews Neuroscience*, *3*(8), 655–666. https://doi.org/10.1038/nrn894

Cui, L., Ye, M., Sun, L., Zhang, S., & He, G. (2022). Common and distinct neural correlates of intertemporal and risky decision-making: Meta-analytical evidence for the dual-system theory. *Neuroscience & Biobehavioral Reviews*, *141*, 104851. https://doi.org/10.1016/j.neubiorev.2022.104851

Edelson, M. G., Polania, R., Ruff, C. C., Fehr, E., & Hare, T. A. (2018). Computational and neurobiological foundations of leadership decisions. *Science*, *361*(6401), Article 6401. https://doi.org/10.1126/science.aat0036

Faillenot, I., Heckemann, R. A., Frot, M., & Hammers, A. (2017). Macroanatomy and 3D probabilistic atlas of the human insula. *NeuroImage*, *150*, 88–98. https://doi.org/10.1016/j.neuroimage.2017.01.073

Fareri, D. S., Niznikiewicz, M. A., Lee, V. K., & Delgado, M. R. (2012). Social Network Modulation of Reward-Related Signals. *Journal of Neuroscience*, *32*(26), 9045–9052. https://doi.org/10.1523/JNEUROSCI.0610-12.2012

Fareri, D. S., Stasiak, J. E., & Sokol-Hessner, P. (2022). Choosing for others changes dissociable computational mechanisms underpinning risky decision-making. *Scientific Reports*, *12*(1), 14361. https://doi.org/10.1038/s41598-022-18437-9

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Frith, C. D. (2014). Action, agency and responsibility. *Neuropsychologia*, *55*(1), 137–142. https://doi.org/10.1016/j.neuropsychologia.2013.09.007

Frith, C. D., & Frith, U. (2006). How we predict what other people are going to do. *Brain Research*, *1079*(1), 36–46. https://doi.org/10.1016/j.brainres.2005.12.126

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). https://doi.org/10.1073/pnas.1802523115

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

Gong, X., Brazil, I. A., Chang, L. J., & Sanfey, A. G. (2019). Psychopathic traits are related to diminished guilt aversion and reduced trustworthiness during social decision-making. *Scientific Reports*, *9*(1), 7307. https://doi.org/10.1038/s41598-019-43727-0

Gu, X., Gao, Z., Wang, X., Liu, X., Knight, R. T., Hof, P. R., & Fan, J. (2012). Anterior insular cortex is necessary for empathetic pain perception. *Brain*, *135*(9), 2726–2735. https://doi.org/10.1093/brain/aws199

Gu, X., Hof, P. R., Friston, K. J., & Fan, J. (2013). Anterior Insular Cortex and Emotional Awareness. *The Journal of Comparative Neurology*, *521*(15), 3371–3388. https://doi.org/10.1002/cne.23368

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences*, *105*(18), 6741–6746. https://doi.org/10.1073/pnas.0711099105

Hill, C. A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J. P., & Ruff, C. C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience*, *20*(8), 1142–1149. https://doi.org/10.1038/nn.4602

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A Neurocomputational Model of Altruistic Choice and Its Implications. *Neuron*, *87*(2), 451–462. https://doi.org/10.1016/j.neuron.2015.06.031

Ifcher, J., & Zarghamee, H. (2020). Behavioral economic phenomena in decision-making for others. *Journal of Economic Psychology*, *77*, 102180. https://doi.org/10.1016/j.joep.2019.06.003

Julle-Danière, E., Whitehouse, J., Vrij, A., Gustafsson, E., & Waller, B. M. (2020). The social function of the feeling and expression of guilt. *Royal Society Open Science*, *7*(12), 200617. https://doi.org/10.1098/rsos.200617

Jung, D., Sul, S., & Kim, H. (2013). Dissociable Neural Processes Underlying Risky Decisions for Self Versus Other. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00015

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–291.

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

Konovalov, A., Hill, C., Daunizeau, J., & Ruff, C. C. (2021). Dissecting functional contributions of the social brain to strategic behavior. *Neuron*, *109*(20), Article 20. https://doi.org/10.1016/j.neuron.2021.07.025

Konovalov, A., & Ruff, C. C. (2021). Enhancing models of social and strategic decision making with process tracing and neural data. *WIREs Cognitive Science*, *13*(1), e1559. https://doi.org/10.1002/wcs.1559

Krajbich, I., Adolphs, R., Tranel, D., Denburg, N. L., & Camerer, C. F. (2009). Economic Games Quantify Diminished Sense of Guilt in Patients with Damage to the Prefrontal Cortex. *Journal of Neuroscience*, *29*(7), 2188–2192. https://doi.org/10.1523/JNEUROSCI.5086-08.2009

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, *54*(3), 2492–2502. https://doi.org/10.1016/j.neuroimage.2010.10.014

Lamm, C., & Singer, T. (2010). The role of anterior insular cortex in social emotions. *Brain Structure and Function*, *214*(5–6), 579–591. https://doi.org/10.1007/s00429-010-0251-3

Loewenstein, G. F., Thompson, L., & Bazerman, M. H. (1989). Social Utility and Decision Making in Interpersonal Contexts. *Journal of Personality and Social Psychology*, *57*(3), 426–441. https://doi.org/10.1037/0022-3514.57.3.426

McLaren, D. G., Ries, M. L., Xu, G., & Johnson, S. C. (2012). A generalized form of context-dependent psychophysiological interactions (gPPI): A comparison to standard approaches. *NeuroImage*, *61*(4), 1277–1286.

Nicolle, A., Klein-Flügge, M. C., Hunt, L. T., Vlaev, I., Dolan, R. J., & Behrens, T. E. J. (2012). An Agent Independent Axis for Executed and Modeled Choice in Medial Prefrontal Cortex. *Neuron*, *75*(6), Article 6. https://doi.org/10.1016/j.neuron.2012.07.023

O'Doherty, J. P., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science*, *304*(5669), Article 5669. https://doi.org/10.1126/science.1094285

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-Based fMRI and Its Application to Reward Learning and Decision Making. *Annals of the New York Academy of Sciences*, *1104*(1), 35–53. https://doi.org/10.1196/annals.1390.022

Ogawa, A., Ueshima, A., Inukai, K., & Kameda, T. (2018). Deciding for others as a neutral party recruits risk-neutral perspective-taking: Model-based behavioral and fMRI experiments. *Scientific Reports*, *8*(1), Article 1. https://doi.org/10.1038/s41598-018-31308-6

Pahlke, J., Strasser, S., & Vieider, F. M. (2015). Responsibility effects in decision making under risk. *Journal of Risk and Uncertainty*, *51*(2), 125–146. https://doi.org/10.1007/s11166-015-9223-6

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-Analysis on Functional Neuroimaging Studies. *Brain Sciences*, *13*(4), 559. https://doi.org/10.3390/brainsci13040559

Piva, M., Velnoskey, K., Jia, R., Nair, A., Levy, I., & Chang, S. W. (2019). The dorsomedial prefrontal cortex computes task-invariant relative subjective value for self and other. *eLife*, *8*, e44939. https://doi.org/10.7554/eLife.44939

Polman, E., & Wu, K. (2020). Decision making for others involving risk: A review and meta-analysis. *Journal of Economic Psychology*, *77*, 102184. https://doi.org/10.1016/j.joep.2019.06.007

Preuschoff, K., Bossaerts, P., & Quartz, S. R. (2006). Neural Differentiation of Expected Reward and Risk in Human Subcortical Structures. *Neuron*, *51*(3), Article 3. https://doi.org/10.1016/j.neuron.2006.06.024

Rogers-Carter, M. M., & Christianson, J. P. (2019). An insular view of the social decision-making network. *Neuroscience & Biobehavioral Reviews*, *103*, 119–132. https://doi.org/10.1016/j.neubiorev.2019.06.005

Rutledge, R. B., De Berker, A. O., Espenhahn, S., Dayan, P., & Dolan, R. J. (2016). The social contingency of momentary subjective well-being. *Nature Communications*, *7*(May), 11825. https://doi.org/10.1038/ncomms11825

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, *111*(33), 12252–12257. https://doi.org/10.1073/pnas.1407535111

Schultz, J., Willems, T., Gädeke, M., Chakkour, G., Franke, A., Weber, B., & Hurlemann, R. (2019). A human subcortical network underlying social avoidance revealed by risky economic choices. *eLife*, *8*, e45249. https://doi.org/10.7554/eLife.45249

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

Seara-Cardoso, A., Sebastian, C. L., McCrory, E., Foulkes, L., Buon, M., Roiser, J. P., & Viding, E. (2016). Anticipation of guilt for everyday moral transgressions: The role of the anterior insula and the influence of interpersonal psychopathic traits. *Scientific Reports*, *6*(1), 36273. https://doi.org/10.1038/srep36273

Singer, T., Critchley, H. D., & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences*, *13*(8), 334–340. https://doi.org/10.1016/j.tics.2009.05.001

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral Emotions and Moral Behavior. *Annual Review of Psychology*, *58*(1), 345–372. https://doi.org/10.1146/annurev.psych.56.091103.070145

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

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, *30*(3), 829–858. https://doi.org/10.1002/hbm.20547

von Gaudecker, H.-M., van Soest, A., & WengstrÖm, E. (2011). Heterogeneity in Risky Choice Behavior in a Broad Population. *American Economic Review*, *101*(2), 664–694. https://doi.org/10.1257/aer.101.2.664

Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V., & Rizzolatti, G. (2003). Both of Us Disgusted in My Insula: The Common Neural Basis of Seeing and Feeling Disgust. *Neuron*, *40*, 655–664.

Wu, S., Sun, S., Camilleri, J. A., Eickhoff, S. B., & Yu, R. (2021). Better the devil you know than the devil you don't: Neural processing of risk and ambiguity. *NeuroImage*, *236*, 118109. https://doi.org/10.1016/j.neuroimage.2021.118109

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

Zeelenberg, M., & Breugelmans, S. M. (2008). The Role of Interpersonal Harm in Distinguishing Regret From Guilt. *Emotion*, *8*(5), 589–596. https://doi.org/10.1037/a0012894

Zhang, X., Liu, Y., Chen, X., Shang, X., & Liu, Y. (2017). Decisions for Others Are Less Risk-Averse in the Gain Frame and Less Risk-Seeking in the

Loss Frame Than Decisions for the Self. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01601

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