A neural mechanism of social responsibility

Maria Gädeke[1], Tom Willems[1,2], Omar Salah Ahmed[1,3], 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: neural mechanisms, responsibility, social decisions, risky choices, anterior insula, superior temporal sulcus, computational reinforcement learning model, fMRI

Abstract

Many risky choices we make affect others in addition to ourselves, and choices made by others also affect us. To study the neural mechanisms underlying social responsibility, we used the following social decision paradigm. In each trial, participants or their game partner chose between a safe and a risky option in a gamble for money. If the risky option was chosen, the gamble was played out independently for both players, such that both could either win or lose the gamble. Participants reported their momentary happiness after experiencing the outcomes of the gambles. Responsibility influenced happiness: ratings were lower following negative outcomes resulting from participants' rather than their partner's choices. The findings of this first behavioural study were replicated in a separate participant sample in the second neuroimaging study. Insula activation was larger in response to negative social outcomes resulting from participants' rather than their partners' choices. A computational modelling-based analysis of these data revealed a cluster of voxels in left superior temporal sulcus whose activation fluctuated with reward prediction errors experienced by the game partner, but to a degree that varied depending on who made the choices leading to these prediction errors. These results suggest that the anterior insula and the superior temporal sulcus play complementary roles in the neural mechanisms of social responsibility.

Introduction

Imagine that it is your turn to choose a restaurant for dining with a friend. You can choose between two restaurants: one that you both know well, with very predictable food of average quality, and a new restaurant that neither you nor your friend has eaten in before. You decide to try the new one. Unfortunately, both your and your friend's dish turn out to be worse than the other restaurant's dishes. How would you feel as a result of your choice? Would you feel different if at least one of you had enjoyed the meal? How would you feel if it had been your friend's run to decide on the restaurant? Many decisions we take involve risks, not only for ourselves but also for other people. In the same way, decisions of others also affect us. This is the case in dyads as well as in large international political contexts, and in decisions ranging from deciding in which restaurant to have dinner to whether one should go to war. Surprisingly, the neural mechanisms underlying key aspects of making decisions in a social context are still unclear. Here, we aimed to study the neural mechanisms underlying responsibility in social decisions – i.e., decisions affecting other people as well as oneself.

Experiencing the outcome of a risky decision is accompanied by emotions that themselves influence future choices (Loewenstein et al., 2001). Negative outcomes lead to negative emotions, for example regret following choices for which one was responsible, or disappointment following choices for which one was not responsible (Bell, 1982, 1985; Coricelli & Rustichini, 2010; Frijda et al., 1989; Gilovich & Medvec, 1994; Giorgetta et al., 2013; Zeelenberg et al., 1998, 2000). Regret usually has a higher emotional impact than disappointment (Bault et al., 2016; Camille et al., 2004; Chua et al., 2009; Mellers et al., 1999), and when combined with a feeling of agency, has been thought to lead to the feeling of responsibility ever since Hellenistic philosophers (Frith, 2014). In social situations, responsibility for negative outcomes for others (interpersonal

harm) may even lead to additional negative emotions such as guilt, which are more likely to occur following interpersonal harm than harm to self only (Berndsen et al., 2004; Wagner et al., 2012; Zeelenberg & Breugelmans, 2008). Empathic responses to another's pain increase with the feeling of responsibility and agency for the pain (Lepron, Causse & Farrer, ProcRoySoc 2015). In this way, the feeling of responsibility influences subsequent behaviour.

What neural mechanisms are engaged when experiencing negative outcomes of risky choices for oneself and for others? Neuropsychological and neuroimaging data have implicated the orbitofrontal cortex, dorsomedial prefrontal cortex, anterior cingulate cortex, anterior insula and amygdala in mediating the experience of regret in single-person decisions (Camille et al., 2004; Chua et al., 2009; Coricelli et al., 2005). One study reported a regret-related amygdala response that increased when responsibility was higher, and this effect was largest in participants in whom responsibility most increased regret (Nicolle et al., 2011). Choosing for others compared to choosing for self is associated with higher activation in social brain regions and reduced activation in reward-related regions (Jung et al., 2013; Ogawa et al., 2018). In this study, activation associated with value computations was found in the amygdala for self-related decisions and in DMPFC for other-related decisions. In another study, VMPFC activity was modulated by activity in the inferior parietal lobule (IPL) when people made product purchase decisions for others, whereas no such effect was found when people made the same decisions for themselves (Janowski et al., 2013). However, the neural mechanisms involved in experiencing responsibility for negative social outcomes are still unclear.

Here, we built on previous work and used a social decision-making paradigm combined with computational reinforcement-learning modelling to study the neural mechanisms underlying social responsibility. Participants chose between a risky and a

safe monetary option, the outcome of which affected either only themselves or themselves and an interaction partner (Figure 1). In a third experimental condition, their partner chose for both themselves and the participant. As the relationship between partners of social decisions influences the utility of unequal outcomes (Loewenstein, Thompson & Bazerman, JPSP, 1989), we induced a non-competitive and sympathetic social atmosphere between participants by having them perform an ice-breaker game before the decision experiment. This game contributed to the feeling of making decisions with a real human being. Participants reported their momentary subjective well-being every two trials. Two experiments in separate sets of subjects were performed, the first outside the MRI scanner, and the second inside it.

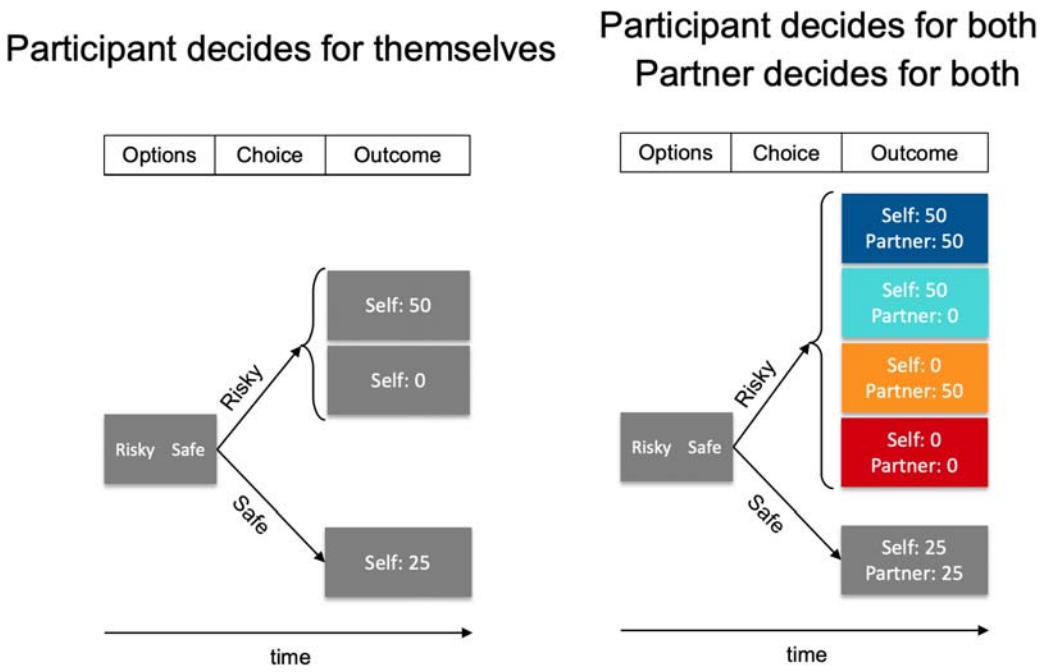

Figure 1*: Experimental design. In every trial, participants are presented with pairs of monetary options (a safe and a risky option; the risky option is a gamble between two equally probable alternatives). There are three conditions: in a non-social condition, the participant's choice leads to an outcome just for themselves (left panel); while in the social conditions (right panel), the decision is taken by the participant or by their partner, and leads in both cases to outcomes affecting both players. Importantly, when the*

risky option is selected in a social condition, the gamble is played out independently for both participants, such that both can receive the higher or lower outcome, independently from each other (coloured boxes). Selecting the safe option leads to equal outcomes for both participants.

Results

Study 1: Behaviour

Decisions

Participants made similar decisions when deciding for self only or for both partners (Figure 2A). Participants' risk attitudes were captured in the form of certainty equivalents, obtained by fitting exponential functions to each participant's proportion of "risky" choices as a function of the difference between the expected values of the safe and risky option (see Methods). Individual certainty equivalents did not significantly vary between the self-only and self-social conditions ($t(39)=0.34$, $p=0.74$, Cohen's $d=0.05$, paired-samples t-test; $BF_{10}=0.18$). Here and elsewhere, BF_{10} refers to the Bayes factor likelihood of the data under the experimental hypothesis H_1 relative to the null hypothesis H_0, calculated using the software package JASP (https://jasp-stats.org/).

Happiness

Participants rated their subjective happiness after the outcome of every second trial. First, we evaluated only the ratings participants gave after outcomes of risky choices (see Figure 2B). Unsurprisingly and consistent with previous findings (Rutledge et al., 2013 and 2016), participant reported greater average happiness after positive compared with negative outcomes in non-social trials ($t(39)=12.0$, $p=1.057e\text{-}14$, $d = 1.9$; $BF_{10}=6.532e^{11}$).

Happiness in social conditions (self-social and partner conditions) varied depending on the outcome for self, the outcome for the partner and depending on who

decided (respective statistical results: $F(1,39)=83.6$, $p=3.0e-11$, $\eta_p^2 = 0.68$; $F(1,39)=24.1$, $p=1.7e-5$, $\eta_p^2 = 0.38$; $F(1,39)=10.4$, $p=0.003$, $\eta_p^2 = 0.21$; 3-way repeated-measures ANOVA with factors decider, outcome for self and outcome for partner). Furthermore, there were significant interactions between outcome for self and outcome for partner, and between decider and outcome for partner ($F(1,39)=6.6$, $p=0.014$, $\eta_p^2 = 0.15$ and $F(1,39)=16.5$, $p=2.27e-4$, $\eta_p^2 = 0.30$). A Bayesian repeated-measures ANOVA analysis revealed that the best model included the factors decider, outcome for self, outcome for partner, interaction between decider and outcome for partner, and interaction between outcome for self and for partner ($BF_{10} > 1.7e^{37}$ compared to null model). Using post-hoc paired-samples t-tests, we then searched for the conditions in which decision-maker (self or partner) mattered and found higher happiness ratings after decisions taken by the partner when the outcome for the partner was negative (self-win/partner-lost: $t(39)=3.58$, $p=9.53e-4$, $d=0.56$, $BF_{10}=32$; self-lost/partner-lost: $t(39)=3.39$, $p=0.002$, $d=0.54$, $BF_{10}=19$), but not after positive outcomes for the partner (self-win/partner-win: $t(39)=0.28$, $p=0.78$, $d=0.05$; self-lose/partner-win: $t(39)=0.38$, $p=0.71$, $d=0.06$; $BF_{10}<0.2$). Participants thus felt worse following negative outcomes for the partner if those outcomes were consequences of decisions taken by themselves rather than by the partner.

Participants felt worse after unequal compared to equal outcomes ($F(1,39)=6.6$, $p=0.014$, $\eta_p^2 = 0.15$, 2-way repeated-measures ANOVA with factors outcome equality and decider; the best Bayesian ANOVA model included factors decider and equality, $BF_{10}=298$ compared to null model). Interestingly, inequality reduced happiness only when participants won ($t(39) > 4.2$, $p<0.001$, $d > 0.6$; $BF > 186$); when participants lost following their own decision, their happiness was instead further reduced when their partner lost as well ($t(39) = 3.2$, $p=0.003$, $d = 0.5$; $BF = 13.5$). Thus, we only observed evidence of advantageous inequality aversion, not disadvantageous inequality aversion.

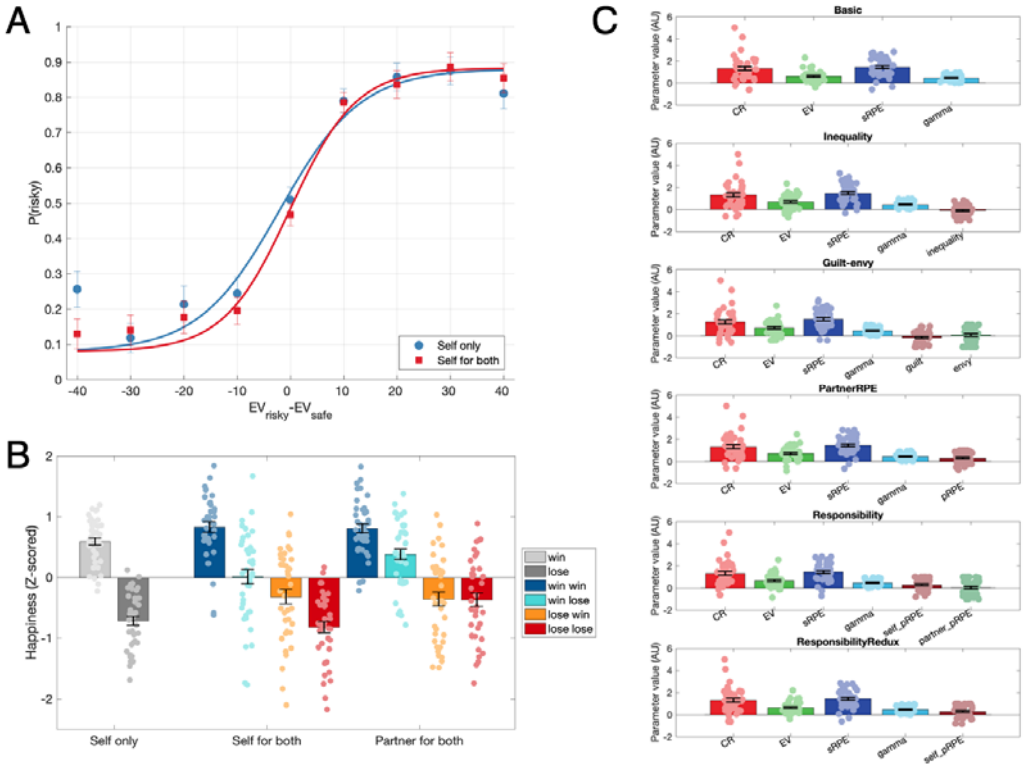

***Figure 2**: Behavioural results of Study 1. **A.** Group-aggregate of participants' decisions for self only and for both game partners. **B.** Happiness ratings following outcomes of risky decisions. Ratings are reported in different colours depending on the outcome (e.g. win-lose represents outcomes in which the participant won and the partner lost) and reported separately for the different experimental conditions (self only = participants decided only for themselves, self for both = participants decided for both themselves and their partner, partner for both = partner decided for both the participant and themselves). **C.** Parameters of the different reinforcement-learning models. In panels B and C, dots show data of individual participants, bars show the mean and error bars show standard errors of the mean.*

Reinforcement learning model of happiness variations

Next, we evaluated whether we could explain variations of happiness across all trials based on predicted and obtained monetary reward values. To this end, following a methodology established by Rutledge and colleagues (Rutledge et al., 2013 and 2016), we

built reinforcement learning models in which certain rewards (CR), the expected value (EV) of chosen gambles, reward prediction errors for the participant (sRPE, where s stands for "self") resulting from those gambles and additional parameters were all modelled separately, with influences that decayed exponentially with time (see Methods for equations and details). As Rutledge and colleagues (2016) did, we first assessed the fit of a basic model (termed "Basic") containing only the variables CR, EV and sRPE, to evaluate whether the variation of these fundamental reward variables explained our participants' variations in subjective well-being. This model explained the data reasonably well (goodness-of-fit measures of all models are shown in Table 1). We then evaluated additional models that included experimental factors identified in our non-computational analyses described above. As inequality reduced happiness, we created an "Inequality" model that included one term modeling the difference between the outcomes for the participant and their partner. As we only found signs of advantageous but not disadvantageous inequality, we created another model with separate terms for advantageous and disadvantageous inequality (the "Guilt-envy" model, which best explained the social decision data obtained by Rutledge et al., 2016). Because negative outcomes for the partner also reduced happiness ratings, we created a "Partner RPE" model, in which a new regressor represented the reward prediction errors for the partner (pRPE), in addition to the regressor for the participant's RPE (sRPE). As the outcome for the partner influenced happiness especially when these outcomes resulted from participant choices, we created a "Responsibility" model in which separate terms represented the partner's reward prediction errors resulting from choices made by the participant (self_pRPE) and by the partner (partner_pRPE). As we expected the latter regressor (partner_pRPE) to have a relatively small impact on participants' momentary happiness yet increase model complexity, we tested a final model, termed "Responsibility

Redux" and identical to the "Responsibility" model except for the omission of the partner_pRPE regressor.

These models all explained variations in happiness reasonably well, but their explanatory power varied (Table 1): R^2 and adjusted R^2 differed across models (R^2: $F(1.5,59.3) = 12.71$, $p < 0.001$, $\eta_p^2 = 0.25$; adjusted R^2: $F(1.5,58.5) = 7.03$, $p < 0.001$, $\eta_p^2 = 0.15$; repeated-measures ANOVA with Greenhouse-Geisser correction for non-sphericity in the data). Post-hoc tests revealed that the Responsibility model yielded significantly higher R^2 values than the other models (all $t>3.6$, $p<0.007$; Bonferroni-corrected t-tests) except for the Guilt-envy model ($t=2.19$, $p=0.17$), and that the adjusted R^2 of the Partner RPE, Responsibility and Responsibility Redux models were higher than that of the Basic model (all $t>3.6$, $p<0.01$). Three models – Basic, Partner RPE and Responsibility Redux – shared the lowest BIC, and the latter two shared the lowest AIC. Overall, the models that included pRPE (Partner RPE, Responsibility and Responsibility Redux) explained the data better than those modeling inequality (higher R^2 / adjusted R^2 or lower AIC / BIC; see Table 1).

Across models, we found that weights for CR, EV, sRPE and the forgetting factor γ were overall positive across participants [all $Z > 5.4$, $P < 0.0001$ (Figure 2D), Wilkoxon sign-rank tests were used because data were not normally distributed] with higher sRPE than EV weights (all $Z > 6.0$, $P < 0.001$) and higher weights for sRPE than for pRPE (all $Z > 6.0$, $P < 0.001$). The forgetting factor γ was constant across models, with medians ranging from 0.39 to 0.44. In the Responsibility model, the weights for the self_pRPE regressor (modeling partner RPEs resulting from choices made by the participant) were larger than 0 ($Z=3.52$, $p=0.0004$), but the weights for the partner_pRPE parameter were not ($Z=1.44$, $p=0.15$); there was no difference between these weights ($Z=0.91$, $p=0.37$).

Table 1. Model comparison in Study 1

Model	*N param*	*Mean R^2*	*Mean R^2adj*	*BIC*	*AIC*
Basic	3	0.328	0.305	**-999**	-1399
Inequality	4	0.346	0.316	-916	-1416
Guilt-envy	5	0.354	0.316	-785	-1385
Partner RPE	4	0.362	0.332	**-999**	**-1499**
Responsibility	5	**0.370**	**0.333**	-866	-1466
Responsibility Redux	4	0.361	0.331	**-999**	**-1499**

BIC, Bayesian Information Criterion. AIC, Akaike's Information Criterion. BIC and AIC values are summed across the 40 subjects. Model fits were performed with z-scored happiness ratings. All models contained separate terms for CR, gamble EV and sRPE, with influences that decayed exponentially over trials. The Inequality model included an additional parameter for the magnitude of the difference in outcomes between the two players. The Guilt–envy model included additional parameters for advantageous and disadvantageous inequality. The Responsibility model included additional parameters for partner RPE resulting from choices made by the participant (self_pRPE) and by the partner (partner_pRPE); the Responsibility Redux model was identical except for the omission of the partner_pRPE regressor. Best values of each variable are highlighted in bold font.

Study 2: Behaviour

We followed the same approach in the analysis of the behavioural data of Study 2 as we did in Study 1.

Decisions

Participants made similar choices when deciding just for themselves of for both partners (Figure 3A), and their individual certainty equivalents did not vary between these conditions ($t(39)=-1.09$, $p=0.28$, $d=-0.17$; $BF_{10}=0.297$).

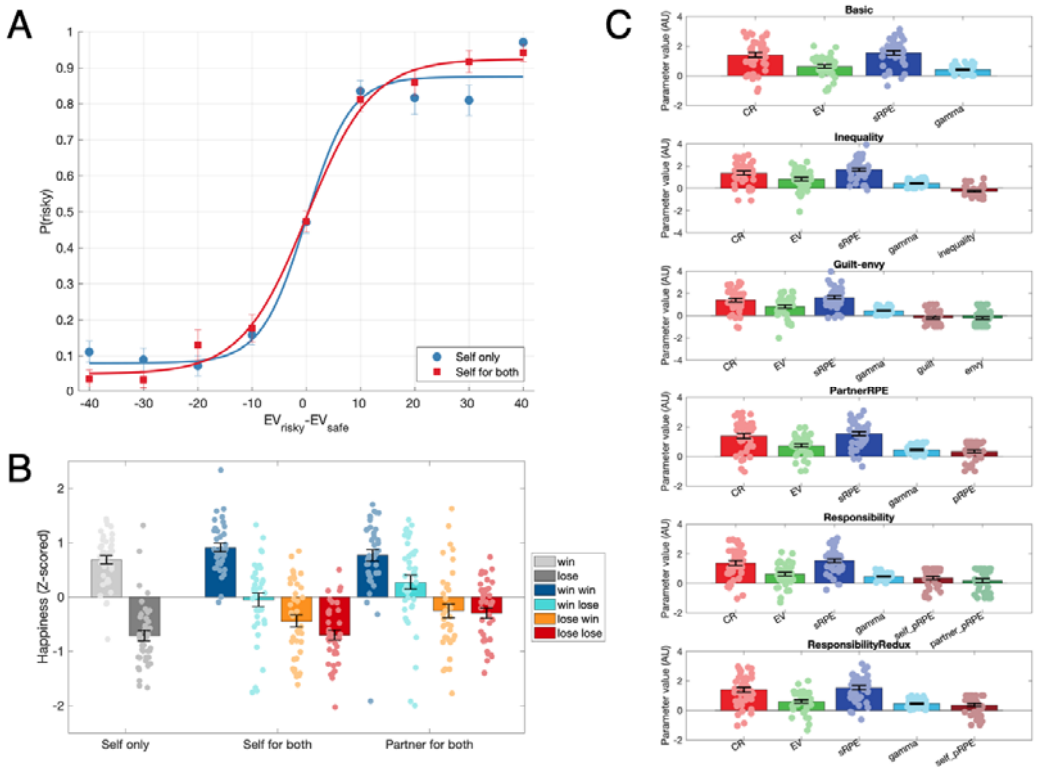

***Figure 3**: Behavioural results of Study 2. **A.** Group-aggregate of participants' choices for self only and for both game partners. **B.** Happiness ratings following outcomes of risky choices. As in Figure 2, ratings are reported in different colours depending on the outcome (e.g. win-lose represents outcomes in which the participant won and the partner lost) and reported separately for the different experimental conditions (self only = participants chose only for themselves, self for both = participants chose for both themselves and their partner, partner for both = partner chose for both the participant and themselves). **C.** Parameters of the different reinforcement-learning models. In panels B and C, dots show data of individual participants, bars show the mean and error bars show standard errors of the mean.*

Happiness

Happiness ratings in Study 2 were analysed as in study 1 and showed almost identical variations. Participants were happier following positive rather than negative outcomes in non-social trials ($t(39)=11.6$, $p=3.56e\text{-}14$, $d = 1.8$; $BF_{10}=2.03e^{11}$). Happiness in social conditions varied depending on the outcome for self, the outcome for the partner and

depending on who decided ($F(1,39)=110.8$, $p=5.86e-13$, $\eta_p^2 = 0.74$; $F(1,39)=26.2$, $p=8.64e-6$, $\eta_p^2 = 0.40$; $F(1,39)=7.8$, $p=0.008$, $\eta_p^2 = 0.17$). There were significant interactions between outcome for self and outcome for partner, and between decider and partner outcome ($F(1,39)=24.7$, $p=1.40e-5$, $\eta_p^2 = 0.39$ and $F(1,39)=6.1$, $p=0.018$, $\eta_p^2 = 0.13$). The best Bayesian repeated-measures ANOVA model again included the factors decider, outcome for self, outcome for partner, interaction between decider and outcome for partner, and interaction between outcome for self and for partner ($BF_{10} = 4.3e^{32}$ compared to null model). After negative outcomes for the partner, happiness was higher if the partner rather than the participant made the choice (self-win/partner-lost: $t(39)=2.13$, $p=0.04$, $d=0.34$, $BF_{10}=8.8$; self-lost/partner-lost: $t(39)=3.05$, $p=0.004$, $d=0.48$, $BF_{10}=1.3$), but this was not the case after positive outcomes for the partner (self-win/partner-win: $t(39)=1.25$, $p=0.22$, $d=0.20$, $BF_{10}=0.35$; self-lose/partner-win: $t(39)=1.26$, $p=0.22$, $d=0.20$, $BF_{10}=0.36$). Thus, participants again felt worse following negative outcomes for the partner if those outcomes resulted from their own rather than their partner's choices.

Unequal outcomes reduced happiness compared to equal outcomes ($F(1,39)=24.7$, $p=1.40e-5$, $\eta_p^2 = 0.39$; best Bayesian ANOVA model included factors decider and equality, $BF_{10}=28507$ compared to null model). As in Study 1, inequality reduced happiness only when participants won ($t(39) > 3.5$, $p<0.001$, Cohen's $d > 0.5$; $BF > 29$); when participants lost following their own choice, their happiness was marginally lower when their partner lost as well ($t(39) = 2.0$, $p=0.053$, Cohen's $d = 0.3$; $BF = 1.02$). Thus, we only observed evidence of advantageous inequality aversion, not disadvantageous inequality aversion, just as we observed in Study 1.

Reinforcement learning model of happiness variations

We fitted the same reinforcement learning models as in Study 1 to the variation of happiness ratings across trials. All models explained variations in happiness reasonably well, and their explanatory power varied (see Table 2). R^2 and adjusted R^2 differed across models (R^2: $F(1.4,53.3) = 16.1$, $p < 0.001$, $\eta_p^2 = 0.29$; adjusted R^2: $F(1.4,53.8) = 9.43$, $p < 0.001$, $\eta_p^2 = 0.19$). Post-hoc tests revealed that the Responsibility model yielded significantly higher R^2 values than all the other models (all $t>2.9$, $p<0.034$), and that the adjusted R^2 of the Partner RPE, Responsibility and Responsibility Redux models were higher than that of the Basic model ($t>3.7$, $p<0.01$). The Responsibility Redux model had the lowest BIC and AIC. Overall, the Responsibility and Responsibility Redux models explained the data best (higher R^2 / adjusted R^2 or lower AIC / BIC; see Table 2). While these differences highlight the cost of additional parameters for the AIC and BIC measures, the results overall demonstrate the impact of responsibility on participants' variations in momentary happiness: taking into account the identity of the decision-maker when modelling outcomes for the partner explains our data best.

Weights for CR, EV, sRPE and the forgetting factor γ were on average positive [all $Z > 4.94$, $p < 0.0001$ (Figure 3D), Wilkoxon sign-rank tests were used because data were not normally distributed] with EV weights lower than sRPE weights [all $Z > 3.7$, $p < 0.003$]. The forgetting factor γ was constant across models, ranging from 0.42 to 0.46. Parameters for the self_pRPE regressor were larger than 0 ($Z=3.64$, $p=0.0003$), parameters for the partner_pRPE regressor were not ($Z=0.85$, $p=0.40$), and parameter values for the self_pRPE regressor were not systematically higher than those for the partner_pRPE parameter ($Z=1.58$, $p=0.11$).

Table 2. Model comparison for Study 2

Model	*N param*	*Mean R^2*	*Mean R^2adj*	*BIC*	*AIC*

Basic	3	0.371	0.338	-617	-952
Inequality	4	0.394	0.350	-563	-981
Guilt-envy	5	0.405	0.350	-443	-944
Partner RPE	4	0.415	0.372	-629	-1047
Responsibility	5	**0.435**	**0.382**	-563	-1065
Responsibility Redux	4	0.422	0.380	**-664**	**-1082**

BIC and AIC values are summed across the 40 subjects. Methods and abbreviations are the same as in Table 1. The Responsibility model had the lowest AIC, highest R^2 and highest adjusted R^2 of all models, suggesting that it fitted the data best. Best values of each variable are highlighted in bold font.

In sum, in two studies, we found that participants felt worse following negative outcomes for the partner if those outcomes resulted their own rather their partner's choices. This influence of responsibility was confirmed when modelling the fluctuations in happiness using reinforcement learning models. On the basis of these findings, we investigated the neural correlates of social responsibility.

BOLD signal

To uncover the neural mechanisms of social responsibility, we investigated brain regions engaged during decision-making and at the time of receiving the outcomes of the choice. Subsequently, we used the "Responsibility" computational model to identify brain regions differentiating between partner reward prediction errors resulting from participant versus partner choices.

Brain regions engaged during decision-making

Bilateral ventral striatum was more active when risky rather than safe options were chosen (Right: x = 10, y = 12, z = -2, Z score = 4.56; Left: x = -12, y = 10, z = -8, Z

score = 4.37; $p<0.05$ corrected for multiple comparisons at the cluster level, Figure 4A). No brain regions showed significant activation in the reverse contrast. Clusters of voxels more active when the partner rather than the participant chose for both players were identified in the superior temporal sulcus, lateral and medial prefrontal as well as orbitofrontal cortex (Figure 4B, cold colours). Clusters of voxels more active when the participant rather than the partner chose for both players were identified in bilateral lingual gyrus, cuneus, caudate, superior temporal gyrus and prefrontal cortex (Figure 4B, warm colours).

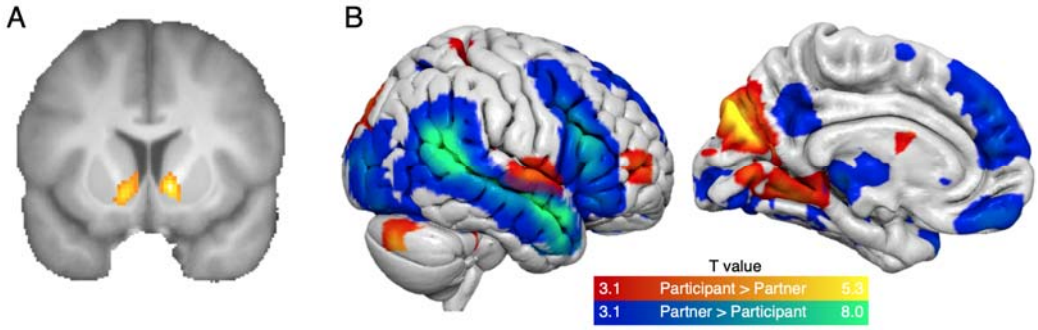

***Figure 4**: BOLD response during decision-making. **A.** Regions showing a greater response when participants chose the risky rather than the safe option, irrespective of social / non-social conditions. **B.** Regions showing a greater response when participants or their partner chose (social conditions only). Colourbar in B shows T values. Results shown survive thresholding at $p<0.05$ corrected for multiple comparisons at the cluster level, based on a voxelwise uncorrected threshold of $p<0.001$.*

Brain regions engaged during receipt of outcomes

Cluster of voxels more active during either participant's receipt of the outcomes of risky vs. safe choices were identified in the medial prefrontal cortex, bilateral anterior insula, bilateral dorsolateral prefrontal cortex, bilateral intraparietal sulcus region, right superior temporal sulcus region, and bilateral ventral striatum (Figure 5A). No significant results

were found in the opposite contrast. For a crucial test, we probed these clusters to identify those differentiating between negative and positive outcomes resulting from risky choices (i.e., prediction errors) that were also sensitive to the nature of the decision-maker. One region showed such a response: the left anterior insula (interaction between sign of the prediction error vs. decision-maker: $F(1,39) = 4.98$, $p = 0.03$; Figure 5B). This region was more active when negative prediction errors resulted from participant rather than partner choices. This difference was not found with positive prediction errors.

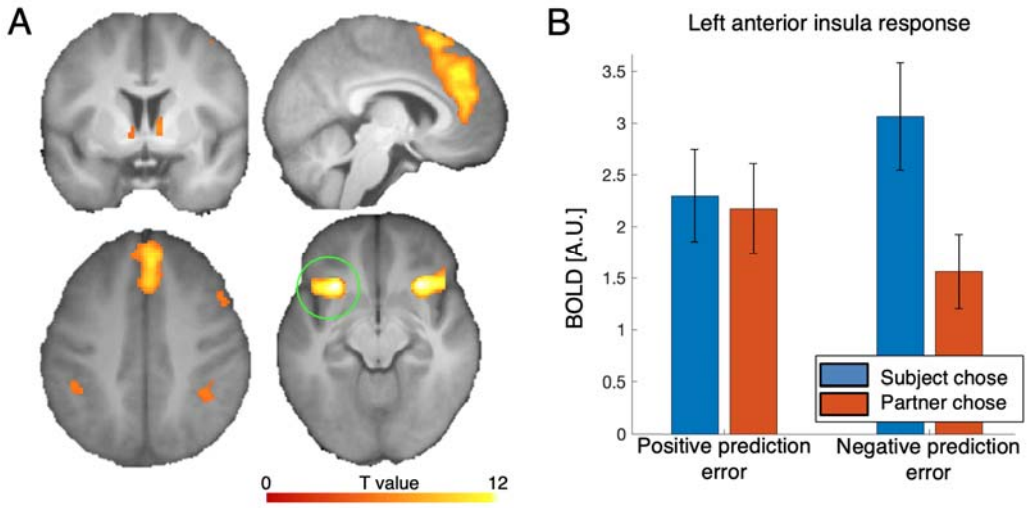

***Figure 5**: BOLD response during receipt of the outcome of choices. **A.** Brain regions more active during receipt of the outcomes of risky choices than safe choices. **B.** Activation in the left anterior insula (circled in green in A) showed a larger response to negative outcomes of risky decisions (=negative prediction errors) taken by the participant rather than by their partner, and no such difference was observed for positive prediction errors. Results shown survive thresholding at $p<0.05$ corrected for multiple comparisons at the cluster level, based on a voxelwise uncorrected threshold of $p<0.001$.*

Neural correlates of responsibility revealed by reinforcement learning model

Finally, within the brain regions sensitive to outcomes of risky choices, we searched for brain regions whose activation could be explained by BOLD responses predicted from

the fitted "Responsibility" reinforcement learning model, which we identified as the best model to explain the variations in momentary happiness during this experiment (see above). Many regions showed responses reflecting the prediction of one or more of these predictors (certain reward, expected value, participant reward prediction error, partner reward prediction error resulting from participant decisions, and partner reward prediction error resulting from partner decisions; Figure 6A). As a crucial test, we searched for regions showing a higher response to partner reward prediction errors resulting from participant rather than partner choices, and found one cluster in the left superior temporal sulcus (MNI: -52, -32, 0; Z = 4.6; Figure 6B).

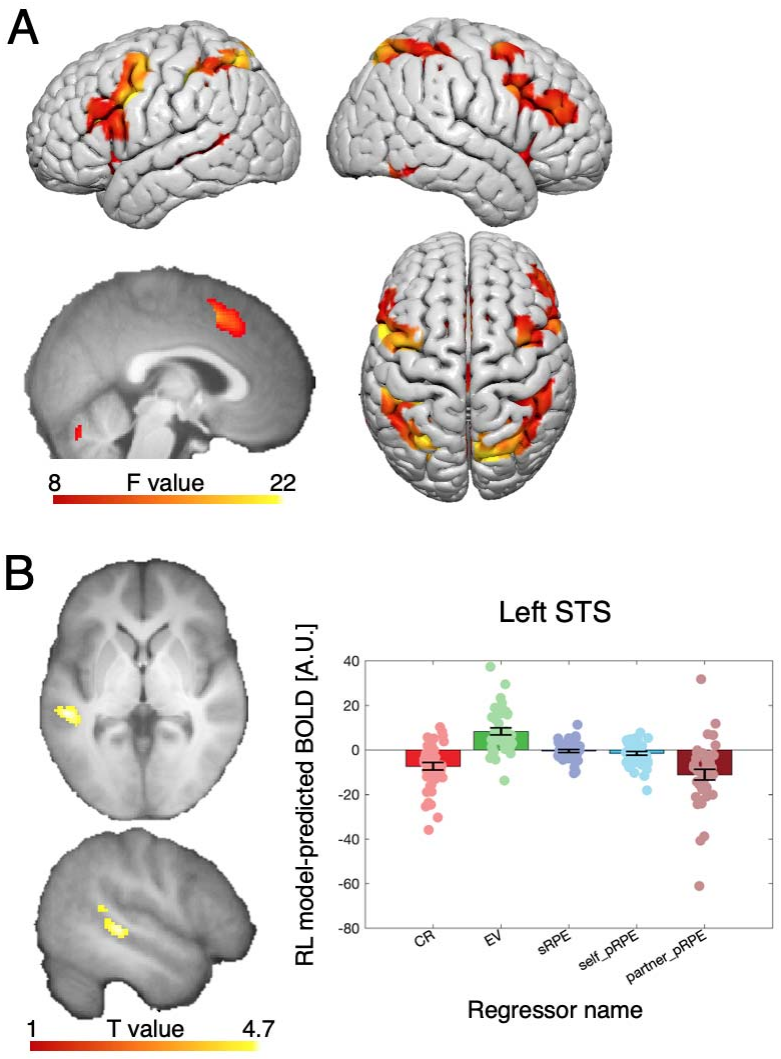

***Figure 6**: Brain regions sensitive to outcomes of risky choices whose response could be predicted based on the fitted reinforcement learning model that best explained the variations in momentary happiness ("Responsibility" model). **A.** Brain regions responding to one or more of the 5 computational regressors tested (Effects of interest; F test). **B.** 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. The response to each of the 5 regressors in this cluster is shown on the right (dots = individual participant data; bars = mean; error bars = S.E.M.). All activations shown survive thresholding at $p<0.05$ corrected for multiple comparisons at the cluster level, based on a voxelwise uncorrected threshold of $p<0.001$. Dots show data of individual participants, bars show the mean and error bars show standard errors of the mean.*

Discussion

This study reports an experimental measurement of social responsibility and its neural correlates. Specifically, we found that being responsible for a negative outcome affecting an interaction partner makes one feel worse than witnessing a negative outcome resulting from the partner's choice. The left anterior insula reflected this effect: Neural activation in this region was higher in response to negative partner outcomes resulting from participant choices rather than from partner choices. A computational model could explain trial-by-trial variations in momentary happiness well; the model best explaining the data differentiated between partner reward prediction errors resulting from participant and partner choices, confirming the importance of responsibility for the emotional consequences of risky social choices. A cluster of voxels in the left superior temporal sulcus region showed activation differentiating between the BOLD responses predicted from these computational regressors. Our findings provide an insight into the

neural mechanisms underlying responsibility for the outcomes of social decisions under risk.

Our study builds on previous work utilizing individual subjective reports of momentary well-being during decision-making tasks to quantify social phenomena. Our design is a modification of a study by Rutledge and colleagues that had revealed the impact of outcome inequality on momentary well-being during risky decision-making (Rutledge et al., 2016). The approach of monitoring emotional responses to experiment events allows to assess the emotional impact of responsibility for the game outcomes separately from economic strategies involved in decision-making. This is an advantage when one's aim is to identify neural mechanisms underlying social emotions independently from decision-making strategies, which is the case in the present study.

Two effects predicted from previous studies were not found in the present study. First, participants made similarly risky choices when deciding for themselves or for both themselves and an interaction partner. A recent meta-analysis comparing decisions for self and for another person concluded that overall, decisions for someone else were somewhat riskier than decisions for oneself, although effects varied strongly across studies (Polman & Wu, 2019). 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. In any case, the fact that participants made similarly risky decisions for themselves or for both does not impact the findings of our study. The second effect we did not find in the present study is a sign of disadvantageous inequality – we only found evidence of advantageous inequality. While asymmetrical outcomes in a social context are related to inequality and generally disliked (Loewenstein et al., 1989), the relation between the interacting people plays an important role in evoking those emotions. Here, we purposefully created a non-competitive social atmosphere in order to

maximize the emotional impact of responsibility for interpersonal outcomes. Our participants might have been biased towards being glad for the other's sake rather than begrudging their luck. Our findings are compatible with reports indicating that reduced anonymity can enhance generosity (Dufwenberg & Muren, 2006) and that affective empathy supports altruistic sharing (Edele et al., 2013). Because we have not tested our paradigm in a competitive social contest, we cannot ascertain whether our non-competitive context was the cause for the absence of disadvantageous inequality, but we consider this explanation likely. In any case, we do not believe that our findings are invalidated by the absence of signs of disadvantageous inequality.

We have not asked participants to specifically report the emotions they experienced during the experiment. As a result, we cannot ascertain whether people experienced guilt, regret, disappointment or another specific emotion during the experiment. However, the precise naming of emotions was not the aim of our study, nor was it to induce guilt or regret. Indeed, previous work on social emotions evoked by decisions under risk recommend to "study the experiential phenomenology of emotions instead of mere emotion labels" (Zeelenberg & Breugelmans, 2008). Our aim was to assess the impact of social responsibility, which we very simply implemented by asking either the participant or their interaction partner to make decisions affecting both of them. We consider the fact that the identity of the decision-maker had an impact on how negative outcomes for the partner affected momentary well-being as sufficient for our aim.

The left anterior insula and the left superior temporal sulcus (STS) region were found to have BOLD responses to negative outcomes for the partner that were larger when those outcomes resulted from participant rather partner choices. Those regions may thus be part of the neural mechanisms underlying the feeling of, or associated with

the negative emotions resulting from, responsibility for the social consequences of one's actions. We identified those regions using different regression models, which assessed somewhat different effects. The cluster in STS, a structure of which a part is thought to serve as a social brain network hub (Dasgupta et al., 2017; Deen et al., 2015; Lee & McCarthy, 2014), showed activation varying with the value of the partner outcome. Negative parameter estimates for the regressor coding those variations resulting from partner choices (Figure 6B) indicate that STS activation decreased with positive partner outcomes and increased with negative partner outcomes. This activation profile is compatible with empathic concern for the partner, which is itself congruent with previous findings linking the STS with cognitive perspective taking during charitable giving (Tusche et al., 2016) or representation of other people's interests during altruistic choice (Hutcherson et al., 2015). These findings concur with other data linking STS to mentalizing computations during strategic social choice (Carter et al., 2012; Hampton et al., 2008; Hill et al., 2017).

In contrast to the STS, anterior insula (AI) activation differentiated between deciders only for negative outcomes for the partner, with the highest response observed when such outcomes resulted from participant's choices. This is compatible with findings associating AI with empathy for negative emotions such as disgust (Wicker et al., 2003) or pain (Gu et al., 2012; Lamm et al., 2011), or with affective empathy during charitable giving (Tusche et al., 2016). Indeed, a previous study found left AI activation that increased with empathic responses to the suffering of others (Bird et al., 2010). AI is associated with emotion awareness (Bird et al., 2010; Gu et al., 2013), and AI lesions combined with poor emotion awareness (alexithymia) are associated with reduced altruistic attitudes (Chau et al., 2018). The involvement of AI in social responsibility may be related to affective empathy for the other person's misfortune resulting from one's own choices.

A previous study of responsibility in social decisions relied on the influence of responsibility on the feeling of regret (Nicolle et al., 2011). In this study, a participant and a variable number of other deciders chose between two lotteries. Consistent with previous work, witnessing that the outcome of a non-selected lottery was larger than the outcome of the selected lottery lead to the experience of regret (Camille et al., 2004). Regret increased with the degree of objective and subjective responsibility, as did regret-related amygdala responses to outcomes associated with regret (Nicolle et al., 2011). In contrast, orbitofrontal and cingulate cortex showed enhanced responses to regret-related outcomes when participants were not objectively responsible for their actions. The different neural correlates identified in this previous study and in ours may related to the dependent measure that was used to assess the impact of responsibility – our present approach searched for effects of responsibility on variations of momentary wellbeing. Our results presented here fit into the current successful trend of using computational models to investigate social cognition (Charpentier & O'Doherty, 2018; Hackel & Amodio, 2018; Konovalov et al., 2018).

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. All subjects gave written informed consent and the studies were conducted in accordance with the latest revision of the Declaration of Helsinki. Subjects 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). 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 data from 4 participants in Study 2 were excluded from the fMRI data analysis because they moved their head too much for reliable motion correction (>3mm or >3°).

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).

Confederate partner

Participants played with a friendly same-sex confederate 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 confederate throughout the task. This icebreaker game led to an agreeable social atmosphere and positive, non-competitive, sympathetic attitude between the participants and the confederate.

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 (Fig. 1*A* and *B*). There were 3 kinds of trials: decisions by the participant only for themselves, decisions by the participant for themselves and the partner, and decisions by the partner for both themselves and the participant. Importantly, when the risky option in a social condition was selected, the gamble was played out independently for the participant and the partner, such that both could receive the higher or lower outcome, 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. The amounts earned per task were determined as follows. There were with 20 mixed trials, 20 gain trials, and 20 loss trials per session. In the mixed trials, subjects chose between a safe option of 0 € or a gamble (=risky option) consisting of a gain and a loss amount. Gain amounts were selected randomly out of the following: 15, 25, 40, 55 or 75 cents. Loss options were determined by multiplying the gain amount with a randomly selected multiplier amount 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 (10, 15, 20, 25, or 30 cents) or a gamble with 0 € and a higher gain amount, calculated by multiplying the safe amount with one of a set of multiplier 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 gamble with 0 € or a larger loss. Certain loss amounts were -10, -15, -20, -25 or -30 cents, and the larger loss amount of the risky option was a multiple of the certain loss amount

calculated using one the multipliers above (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 and right) was determined randomly on every trial, and the different trial types (participant chooses for both, partner chooses for both, participant chooses for self; gain, loss and mixed trials) were presented in randomized order, with the constraint that there could not be more than 2 conditions of the same kind in a row. Subjects had unlimited time to choose between the safe or risky option. Decisions were displayed for 2s and gamble 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.

Study 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).

Computational modelling

1) Decisions: Certainty equivalents

The certainty equivalents of the risky options in the self-only and self-social conditions were calculated by fitting an exponential function to each participant's proportions of

risky choices. This function represents the assumption that the likelihood of choosing the risky option on a given trial is related to the difference between the expected value of the risky option (EV_R) and the safe option (EV_S) and is described by the following relationship:

$$P(R) = (1 + e^{-\lambda(EV_R - EV_S)})^{-1}$$

(equation 1)

where P(R) is the probability of choosing the risky option, EV_R and EV_S are the expected values of the risky and the safe options (respectively), and $\boldsymbol{\lambda}$ is indicative of the steepness of the decision curve, individually fitted to each participant's data using the Levenberg-Marquardt algorithm for solving non-linear least squares problems, implemented in Matlab's lsqcurvefit function and a range of = [0, 30]. After fitting, certainty equivalents were defined as the EV_R-EV_S difference that led to 50% risky choices, and certainty equivalents as well as values were compared between self-only and self-social conditions using paired t-tests.

2) Happiness: Reinforcement learning model

We then modelled the variations in momentary well-being (happiness ratings) using a computational model that has been shown to explain happiness variations in a similar task as a function of the combined influence of recent reward expectations, prediction errors resulting from those expectations, and differences in rewards obtained by the game partners (Rutledge et al., 2014, 2016). We ran 6 models: i) a "Basic" model (the winning model in Rutledge et al. 2014, in which there was no social condition; equation 2) with chosen certain rewards (CR), expected value of chosen gambles (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), with without reward expectations and reward differences between partners; ii)

an "Inequality" model, in which one regressor modelled both disadvantageous and advantageous inequality; 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; iv) a "Partner RPE" model in which a regressor modelled the interaction partner's reward prediction errors; v) a "Responsibility" model in which reward prediction errors of the partner resulting from participant and partner choices were modelled separately; and vi) a "Responsibility redux" model identical to the precedent except for the absence of the regressor modelling reward prediction errors of the partner resulting from partner choices. The Basic, Inequality and Guilt-envy models are identical to those used in Rutledge et al., 2016.

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 2)

where t is trial number, γ 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 gamble on trial *j*, EV_j is the average reward for the gamble 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 gamble. 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.

On the basis of this model, we built 4 other models 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 3)

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 4)

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 "Partner RPE" model was:

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

(equation 5)

Here, w_4 represents the influence of the partner's RPE (pRPE) on the participants' variation in happiness. The next two models included terms relating to the partner's RPE. The equation for the "Responsibility" model was:

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

(equation 6)

In this model, the influence of the partner's RPE resulting from participant choices (self_pRPE) is modelled separately from those resulting from the partner's RPE resulting from partner choices (partner_pRPE). The equation for the last model, "Responsibility redux", was:

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

(equation 7)

Statistics

Statistics on the behavioural data and neural activation data (see below) were performed using JASP software (JASP Version 0.9.2; RRID:SCR_015823; JASP Team 2018; jasp-stats.org). Whole-brain activation statistics were performed with SPM12 software (RRID:SCR_007037; Wellcome Trust Centre for Neuroimaging, London, UK; http://www.fil.ion.ucl.ac. uk/spm) running in MATLAB with corrections for multiple

comparisons (for details, see 'fMRI data analysis' below). All statistical tests were two-tailed. Bayes factors were calculated using default priors and express the probability of the data given H1 relative to H0 (BF10, values larger than one are in favour of H1). Effect sizes were calculated using standard approaches implemented in JASP software.

fMRI data acquisition and preprocessing

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 subjects 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 preprocessed and analysed 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

Functional data were analysed using a two-stage approach based on the general linear model (GLM) implemented in SPM12: individual participants' data were modelled with fixed effects models, and their summary data were entered in a random effects model for group statistics and inferences at the population level. The fixed effects models implemented a mass univariate analysis applied to the preprocessed data. The first model included the following event types for each session: certain reward (CR, relevant when a safe option was chosen), expected value (EV, relevant when a risky decision was chosen) and prediction error for the participant (sRPE, determined at the outcome of a risky decision), all modelled separately for social trials with participant decisions, social trials with partner decisions, and non-social trials; then, another set of event types coding the difference between outcome for self and partner, modelled separately for social trials with participant decisions and social trials with partner decisions. All event types were modelled as stick functions at the time of occurrence and convolved by SPM with its canonical hemodynamic function, with monetary values serving as parametric modulators. The second model was designed to identify brain regions whose activation reflected the variables of the computational model: receipt of certain rewards, expected value of the chosen option, participant's reward prediction error (sRPE), partner RPE resulting from decisions made by the participant (self_pRPE) and 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 of the 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. All models 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. Linear contrasts were applied to the individual parameter estimates of the response to the experimental conditions, resulting in contrast images. These were subjected to a group-wise random effects ANOVA in order to identify brain regions showing a greater response when participants chose the risky rather than the safe option, regions showing different responses depending on who took the decision, regions more active when participants received the outcome of risky rather than safe decisions, and regions with BOLD signal reflecting variations predicted from the computational model. The results from the computational model were constrained to regions that showed a stronger response during risky than safe outcomes in the non-computational model, thresholded at $p=0.001$ uncorrected, using the inclusive mask option in SPM. All clusters shown survive a significance threshold of $p<0.05$ with family-wise error correction for multiple comparisons (FWE) across the whole brain, based on an uncorrected voxel-wise threshold of $p<0.001$.

Acknowledgements

We would like to thank Dr. Robb Rutledge for discussions and for providing the reinforcement learning code using in their previous publication (Rutledge et al., 2016).

References

Bault, N., Wydoodt, P., & Coricelli, G. (2016). Different Attentional Patterns for Regret and Disappointment: An Eye-tracking Study. *Journal of Behavioral Decision Making*, *29*(2–3), 194–205. https://doi.org/10.1002/bdm.1938

Bell, D. E. (1982). Regret in Decision Making under Uncertainty. *Operations Research*, *30*(5), 961–981.

Bell, D. E. (1985). Disappointment in Decision Making Under Uncertainty. *Operations Research*, *33*(1), 1–27. https://doi.org/10.1287/opre.33.1.1

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

Camille, N., Coricelli, G., Sallet, J., Pradat-Diehl, P., Duhamel, J. R., & Sirigu, A. (2004). The involvement of the orbitofrontal cortex in the experience of regret. *Science*, *304*(5674), 1167–1170. https://doi.org/10.1126/science.1094550

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

Charpentier, C. J., & O'Doherty, J. P. (2018). The application of computational models to social neuroscience: Promises and pitfalls. *Social Neuroscience*, *13*(6), 637–647. 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

Chua, H. F., Gonzalez, R., Taylor, S. F., Welsh, R. C., & Liberzon, I. (2009). Decision-related loss: Regret and disappointment. *NeuroImage*, *47*(4), 2031–2040. https://doi.org/10.1016/j.neuroimage.2009.06.006

Coricelli, G., Critchley, H. D., Joffily, M., O'Doherty, J. P., Sirigu, A., & Dolan, R. J. (2005). Regret and its avoidance: A neuroimaging study of choice behavior. *Nature Neuroscience*, *8*(9), 1255–1262. https://doi.org/10.1038/nn1514

Coricelli, G., & Rustichini, A. (2010). Counterfactual thinking and emotions: Regret and envy learning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *365*(1538), 241–247. https://doi.org/10.1098/rstb.2009.0159

Dasgupta, S., Tyler, S. C., Wicks, J., Srinivasan, R., & Grossman, E. D. (2017). Network Connectivity of the Right STS in Three Social Perception Localizers. *Journal of Cognitive Neuroscience*, *29*(2), 221–234. https://doi.org/10.1162/jocn_a_01054

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. *Cerebral Cortex*.

Dufwenberg, M., & Muren, A. (2006). Generosity, anonymity, gender. *Journal of Economic Behavior & Organization*, *61*(1), 42–49. https://doi.org/10.1016/j.jebo.2004.11.007

Edele, A., Dziobek, I., & Keller, M. (2013). Explaining altruistic sharing in the dictator game: The role of affective empathy, cognitive empathy, and justice sensitivity. *Learning and Individual Differences*, *24*, 96–102. https://doi.org/10.1016/j.lindif.2012.12.020

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

Frijda, N. H., Kuipers, P., & Schure, E. (1989). Relations Among Emotion, Appraisal, and Emotional Action Readiness. *Journal of Personality and Social Psychology*, *57*(2), 212–228.

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

Gilovich, T., & Medvec, V. H. (1994). The temporal pattern to the experience of regret. *Journal of Personality and Social Psychology*, *67*(3), 357–365. https://doi.org/10.1037//0022-3514.67.3.357

Giorgetta, C., Grecucci, A., Bonini, N., Coricelli, G., Demarchi, G., Braun, C., & Sanfey, A. G. (2013). Waves of regret: A meg study of emotion and decision-making. *Neuropsychologia*, *51*(1), 38–51. https://doi.org/10.1016/j.neuropsychologia.2012.10.015

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

Hackel, L. M., & Amodio, D. M. (2018). Computational neuroscience approaches to social cognition. *Current Opinion in Psychology*, *24*, 92–97. https://doi.org/10.1016/j.copsyc.2018.09.001

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

Janowski, V., Camerer, C., & Rangel, A. (2013). Empathic choice involves vmPFC value signals that are modulated by social processing implemented in IPL. *Social Cognitive and Affective Neuroscience*, *8*(2), 201–208. https://doi.org/10.1093/scan/nsr086

Jung, D., Sul, S., & Kim, H. (2013). Dissociable neural processes underlying risky decisions for self versus other. *Frontiers in Neuroscience*, *7*(7 MAR), 1–12. https://doi.org/10.3389/fnins.2013.00015

Konovalov, A., Hu, J., & Ruff, C. C. (2018). Neurocomputational approaches to social behavior. *Current Opinion in Psychology*, *24*, 41–47. https://doi.org/10.1016/j.copsyc.2018.04.009

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

Lee, S. M., & McCarthy, G. (2014). Functional Heterogeneity and Convergence in the Right Temporoparietal Junction. *Cerebral Cortex*.

Loewenstein, G. F., Hsee, C. K., Weber, E. U., & Welch, N. (2001). Risk as Feelings. *Psychological Bulletin*, *127*(2), 267–286. https://doi.org/10.1037/0033-2909.127.2.267

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

Mellers, B., Schwartz, A., Ritpv, D., Tunac, N., Lu, T., Lande, E., Compian, L., & Ritov, I. (1999). Emotion-based choice. *Journal of Experimental Psychology: General*, *128*(3), 332–345. https://doi.org/10.1037/0096-3445.128.3.332

Nicolle, A., Bach, D. R., Frith, C., & Dolan, R. J. (2011). Amygdala involvement in self-blame regret. *Social Neuroscience*, *6*(2), 178–189. https://doi.org/10.1080/17470919.2010.506128

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), 1–2. https://doi.org/10.1038/s41598-018-31308-6

Polman, E., & Wu, K. (2019). Decision making for others involving risk: A review and meta-analysis. *Journal of Economic Psychology*, *72*, 200–218. https://doi.org/10.1016/j.joep.2019.03.007

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. *PNAS*, *111*(33), 12252–12257.

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

Wagner, U., Handke, L., Dörfel, D., & Walter, H. (2012). An experimental decision-making paradigm to distinguish guilt and regret and their self-regulating function via loss averse choice behavior. *Frontiers in Psychology*, *3*(OCT), 1–12. https://doi.org/10.3389/fpsyg.2012.00431

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.

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

Zeelenberg, M., Van Dijk, W. W., Manstead, A. S. R., & Van Der Pligt, J. (1998). The Experience of Regret and Disappointment. *Cognition and Emotion*, *12*(2), 221–230. https://doi.org/10.1080/026999398379727

Zeelenberg, M., Van Dijk, W. W., Manstead, A. S. R., & Van Der Pligt, J. (2000). On bad decisions and disconfirmed expectancies: The psychology of regret and disappointment. *Cognition and Emotion*, *14*(4), 521–541. https://doi.org/10.1080/026999300402781