JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Neural mechanisms underlying individual differences in control-averse behavior

Sarah Rudorf[1,2], Katrin Schmelz[3,4], Thomas Baumgartner[1,2], Roland Wiest[5], Urs Fischbacher[3,4] and Daria Knoch[1,2]

[1]*Department of Social Psychology and Social Neuroscience, Institute of Psychology, University of Bern, 3012 Bern, Switzerland*
[2]*Center for Cognition, Learning and Memory, University of Bern, 3012 Bern, Switzerland*
[3]*Department of Economics, University of Konstanz, 78464 Konstanz, Germany*
[4]*Thurgau Institute of Economics, 8280 Kreuzlingen, Switzerland*
[5]*Department of Neuroradiology, Inselspital, 3010 Bern, Switzerland*

DOI: 10.1523/JNEUROSCI.0047-18.2018

Received: 9 January 2018

Revised: 20 March 2018

Accepted: 25 April 2018

Published: 14 May 2018

Author contributions: S.R. wrote the first draft of the paper; S.R., K.S., T.B., U.F., and D.K. edited the paper; S.R., K.S., T.B., U.F., and D.K. designed research; S.R. and R.W. performed research; S.R. analyzed data; S.R., K.S., T.B., U.F., and D.K. wrote the paper.

Conflict of Interest: The authors declare no competing financial interests.

This project was supported by a grant to DK by the Mens Sana Foundation.

Correspondence to: Sarah Rudorf, Department of Social Psychology and Social Neuroscience, Institute of Psychology, University of Bern, Fabrikstrasse 8, 3012 Bern, Switzerland, sarah.rudorf@psy.unibe.ch, or Daria Knoch, Department of Social Psychology and Social Neuroscience, Institute of Psychology, University of Bern, Fabrikstrasse 8, 3012 Bern, Switzerland, daria.knoch@psy.unibe.ch

Cite as: J. Neurosci ; 10.1523/JNEUROSCI.0047-18.2018

Neural mechanisms underlying individual differences in control-averse behavior

Running title: Neural mechanisms of control-averse behavior

Sarah Rudorf[1,2], Katrin Schmelz[3,4], Thomas Baumgartner[1,2], Roland Wiest[5], Urs Fischbacher[3,4], Daria Knoch[1,2]

[1] Department of Social Psychology and Social Neuroscience, Institute of Psychology, University of Bern, 3012 Bern, Switzerland
[2] Center for Cognition, Learning and Memory, University of Bern, 3012 Bern, Switzerland
[3] Department of Economics, University of Konstanz, 78464 Konstanz, Germany
[4] Thurgau Institute of Economics, 8280 Kreuzlingen, Switzerland
[5] Department of Neuroradiology, Inselspital, 3010 Bern, Switzerland

Correspondence to:
Sarah Rudorf
Department of Social Psychology and Social Neuroscience, Institute of Psychology, University of Bern
Fabrikstrasse 8, 3012 Bern, Switzerland
sarah.rudorf@psy.unibe.ch
or
Daria Knoch
Department of Social Psychology and Social Neuroscience, Institute of Psychology, University of Bern
Fabrikstrasse 8, 3012 Bern, Switzerland
daria.knoch@psy.unibe.ch

Number of pages: 47
Number of figures: 8
Number of tables: 2
Number of words in
Abstract: 206
Introduction: 648
Discussion: 1475

Conflict of Interest: The authors declare no competing financial interests.
Acknowledgements: This project was supported by a grant to DK by the Mens Sana Foundation.

ABSTRACT

When another person tries to control one's decisions, some people might comply, but many will feel the urge to act against that control. This control aversion can lead to suboptimal decisions and it affects social interactions in many societal domains. To date, however, it has been unclear what drives individual differences in control-averse behavior. Here, we address this issue by measuring brain activity with fMRI while healthy female and male human participants make choices that are either free or controlled by another person, with real consequences to both interaction partners. In addition, we assessed the participants' affects, social cognitions and motivations via self-reports. Our results indicate that the social cognitions perceived distrust and lack of understanding for the other person play a key role in explaining control aversion at the behavioral level. At the neural level, we find that control-averse behavior can be explained by functional connectivity between the inferior parietal lobule and the dorsolateral prefrontal cortex, brain regions commonly associated with attention reorientation and cognitive control. Further analyses reveal that the individual strength of functional connectivity complements and partially mediates the self-reported social cognitions in explaining individual differences in control-averse behavior. These findings therefore provide valuable contributions to a more comprehensive model of control aversion.

SIGNIFICANCE STATEMENT

Control aversion is a prevalent phenomenon in our society. When someone tries to control their decisions, many people tend to act against the control. This can lead to suboptimal decisions, like noncompliance to medical treatments or disobeying the law. The degree to which individuals engage in control-averse behavior, however, varies significantly. Understanding the proximal mechanisms that underlie individual differences in control-averse behavior has potential policy implications, for example when designing policies aimed at increasing compliance with vaccination recommendations, and is therefore a highly relevant research goal. Here, we identify a neural mechanism between parietal and prefrontal brain regions that can explain individual differences in control-averse behavior. This mechanism provides novel insights into control aversion beyond what is accessible through self-reports.

INTRODUCTION

When others try to control our decisions, many of us will feel the urge to counteract and thereby reestablish our valued freedom of choice. This aversive reaction to the exogenous control of one's freedom of choice, or in short control aversion, puts a strain on many societal domains, for example in the form of patient noncompliance to psychiatric therapy (De las Cuevas et al., 2014), adolescent defiance against parents (Van Petegem et al., 2015), or employees' reduced work performance when faced with a restrictive employer (Falk and Kosfeld, 2006). Critically, the degree to which individuals engage in control-averse behavior varies largely, which has been documented in numerous studies (Falk and Kosfeld, 2006; Ziegelmeyer et al., 2012; Schmelz and Ziegelmeyer, 2015). What drives these individual differences in control-averse behavior, however, has remained an open question.

Previous work has shown that individuals whose decisions are controlled by another person often report thoughts about the other person's motives, such as distrust, and lack of understanding for the other person's decision to control (Falk and Kosfeld, 2006). For example, when an employer requests a minimum effort from her employee, the employee may perceive this as a signal of distrust in her intrinsic work motivation. A separate line of work has highlighted the motivation to restore one's freedom of choice, termed reactance, as the key player in driving control-averse behavior (Brehm, 1966; Miron and Brehm, 2006). For example, the elimination of a choice option can lead to an increased desire for that option, which is interpreted as an indirect strategy of freedom restoration (Miron and Brehm, 2006). Moreover, reactance is assumed to be accompanied by negative affects, such as anger (Dillard and Shen, 2005). Therefore, negative affects and individual tendencies to express one's anger outward might contribute to the display of control-averse behavior. The literature thus delivers several

plausible variables that might drive individual control-averse behavior. Much of the support to date, however, comes from post-hoc self-reports or measures of behavioral intentions in hypothetical scenarios. Here, we use a neurophysiological measure of the decision processes during real restrictions of the subjects' freedom of choice. By doing so we aim to identify the proximal mechanisms that give rise to individual differences in control-averse behavior. Specifically, we test whether activation in and functional connectivity with the brain regions that are differentially activated during the restriction of the freedom of choice can explain individual differences in control-averse behavior. Moreover, we investigate to what degree this neurophysiological measure complements and mediates self-report data in predicting individual control-averse behavior.

To this end, we combine functional magnetic resonance imaging (fMRI) with a Control aversion task (Fig. **1**), in which subjects make decisions that are either free or controlled by another person (Falk and Kosfeld, 2006; Schmelz and Ziegelmeyer, 2015). For each decision, subjects allocate money between themselves and another person by choosing between options that increase in fairness and generosity, called generosity levels. Crucially, the options were designed to establish an intrinsic motivation to choose a high level when subjects can decide freely. When the other person requests a minimum level and thereby tries to control the subject's choice, control-averse behavior is defined as choosing a lower level (Falk and Kosfeld, 2006; Schmelz and Ziegelmeyer, 2015). Therefore, the decrease of average chosen levels when the other person tries to control the subject's decision as opposed to the free decisions serves as a measure of individual control-averse behavior. Critically, the decisions in the task are not hypothetical, but have real consequences for both interaction partners and thus share an important quality with control-averse behavior outside the laboratory. This setup allows us not only to measure control-averse behavior in an ecologically valid fashion, but to

investigate the neural responses during the actual decision making process. We find that a neural mechanism involving parietal and prefrontal brain regions complements and partially mediates self-reported social cognition in explaining individual differences in control-averse behavior.

MATERIALS AND METHODS

Participants

We recruited a total of 61 students from the University of Bern for participation in this study. Students of economics, psychology and social sciences were excluded from participation to reduce the possibility of prior knowledge of the concept of control aversion. All participants were right-handed, nonsmokers and reported no history of psychological disorders, neurological or cardiovascular diseases. After data acquisition, ten participants were excluded due to excessive movements during fMRI scan (> 5 mm in translation or > 5 degrees in rotation), noncompliance to instructions or technical problems. The remaining 51 participants (23 female; mean 22 ± 3 SD years) were included in the analysis. All participants received a compensation of CHF 50 (≈ USD 50) for participation in the study in addition to the payoff from the Control aversion task described in the next section. The study was approved by the Bern Cantonal Ethics Commission and all participants gave informed, written consent.

Experimental design

Control aversion task

The Control aversion task (Fig. **1**) is designed to confront subjects with real restrictions of their freedom of choice by another person and is based on previous work in behavioral economics (Falk and Kosfeld, 2006; Schmelz and Ziegelmeyer, 2015). The gist of the task is that the subject is asked to allocate money between herself and another person, called player A. But before the subject makes a decision, player A can decide to let the subject choose freely (Free condition) or request a minimum amount of money (Controlled condition).

For the purpose of this study, subjects were presented with 16 anonymous other persons' (players A's) decisions from a pilot study, in random order. The small number of trials was chosen to increase credibility and reduce possible habituation effects. To ensure equal estimation power of the blood oxygen level-dependent (BOLD) signal across conditions, the players A's decisions were preselected such that the subjects engaged in the same number of trials in the Free and in the Controlled condition, i.e. eight trials per condition. All subjects were informed that the players A's decisions had been prerecorded for logistic reasons, and they were asked to decide as if the respective person was present. To remind subjects of this instruction, we presented the line "A new player A is deciding" for a jittered interval of 2.4-8.6 s at the beginning of each trial. Subjects were also informed that their choices had real consequences in the sense that one trial would be randomly selected and paid out to themselves and the corresponding player A. None of the subjects voiced suspicions about the existence of the players A. After a jittered fixation display of 2-6 s, subjects learned whether the player A let them choose freely (Free condition) or whether the player A requested a minimum amount of monetary units (MUs) (Controlled condition). After a delay of three seconds, subjects

made a choice between sets of monetary allocations, called generosity levels, ranging from a selfish (subject : player A, 99:1 MUs) to a more generous, equal allocation (80:80 MUs) (all possible generosity levels are depicted in Fig. **1**). Subjects made their choice by moving a red selection frame from a random position to their desired option and pressing an OK button. Response times were not constrained to motivate deliberate decisions; subjects were asked, however, to respond as soon as they had come to a decision (response times, mean 5 ± SD 4.3 s). Note, that for the fMRI analysis we separated the times before and after subjects started to move the selection frame to capture the decision window and the motor responses separately. The durations as used in the fMRI analysis are shown in Figure **1**. In the Free condition, subjects had the choice between generosity levels one to five (from left to right). In the Controlled condition, subjects' choice was restricted to generosity levels two (97:30 MUs) to five. A central feature of the task is that the player A's payoff increases as a concave function of the generosity levels with relatively small and convex costs for the subject. Moreover, the most generous level (level five) also represented the fairest and equal option and the highest sum of payoffs. These features were added to ensure that subjects are intrinsically motivated to choose a high level, which is a prerequisite for control aversion in this task (Schmelz and Ziegelmeyer, 2015). Lastly, the subject's payoff remains constant for levels two to three. This was done to motivate subjects to choose level three over level two in the Free condition, and to provide space for the choice of a lower level in the Controlled condition that is independent of economic self-interest. The difference between a subject's mean chosen level in the Free condition minus the subject's mean chosen level in the Controlled condition served as the measure of the individual level of control-averse behavior.

After another jittered fixation display of 5-8 s, subjects were asked to indicate how they had felt during the decision by rating their unhappiness and anger on 5-point pictorial Self-Assessment Manikins (SAMs) (Bradley and Lang, 1994), each separated by a jittered fixation display of 1-4 s. The unhappiness scale ranged from 1 = “happy” to 5 = “unhappy”, and the anger scale ranged from 1 = “calm” to 5 = “angry”. As a manipulation check we implemented a third scale, the having control scale, which ranged from 1 = “being controlled” to 5 = “having control”. Finally, a fixation cross was displayed for 1.2-6.4 s before the next trial began.

Prior to scanning, subjects read the instructions and were quizzed to ensure they had understood the task and its payoff scheme. Subjects then practiced four simulated trials of the Control aversion task outside of the scanner to familiarize themselves with the task timing and the response buttons. Then subjects completed the scanning task in one continuous session of approximately 12 minutes. At the end of the task, one trial was randomly selected for payoff to the subject and the matched player A. Therefore, all trials were incentive compatible to motivate subjects to decide according to their true preferences. The profits in the selected trial were converted into CHF (with 1 MU = CHF 0.20 ≈ USD 0.20). Based on the task, the subjects received a mean CHF 18.30 ± 1.40 SD, and the players A received a mean CHF 11.10 ± 3.80 SD.

Ratings of perceived distrust, understanding, freedom restoration and fairness

Directly after scanning, we assessed subjects’ thoughts during the Control aversion task with a list of items. For each item, subjects were asked to rate how strongly the described thought had influenced their decisions on a 7-point Likert scale ranging from 1 = “not at all” to 7 = “absolutely”. Based on the seminal study by Falk and Kosfeld (2006), we assessed subject’s perceived distrust and understanding with the items

"When player A requests a minimum of generosity, he distrusts me and I dislike that." ('perceived distrust'), and "I understand when player A requests a minimum of generosity." ('understanding'). Based on reactance theory (Brehm, 1966; Miron and Brehm, 2006), we assessed subjects' motivation to restore their freedom of choice in the Controlled condition with the item "When player A restricts the generosity levels, I want to use my remaining freedom of choice all the more." ('freedom restoration'). In addition, we asked subjects whether fairness had played a role in their own decisions with the item, "I think that my payoff and player A's payoff should not be too far apart." ('fairness').

Assessment of outward directed anger expression

To assess subjects' general tendency to direct their anger outward, we asked subjects to fill in the German version of the State-Trait Anger Expression Inventory (STAXI) (Spielberger, 1988; Schwenkmezger et al., 1992). The STAXI is comprised of the five subscales state anger, trait anger, inward directed anger expression, outward directed anger expression and controlling one's anger expression. Here, we focus on the subscale for outward directed anger expression (AO). The AO subscale consists of 8 items that describe ways of expressing one's anger, e.g. "I fly off the handle". Subjects rated these items on a 4-point Likert scale ranging from 1 = "almost never" to 4 = "almost always". Based on the subjects' ratings, the sum scores were computed. In our sample, the AO subscale had an acceptable internal consistency (Cronbach's alpha = 0.73). On average, subjects had an AO score of mean 12.24 ± SD 3.02 (range 8-22), which is similar to the norm student sample reported in Schwenkmezger et al. (1992).

MRI data acquisition and preprocessing

All MRI data were acquired on a Siemens Trio 3.0 Tesla whole-body scanner (Siemens, Erlangen) using a 12-channel head coil. The functional session started off with a localizer scan followed by the Control aversion task implemented in E-Prime 3.0 (Psychology Software Tools). The task was projected onto a screen that the subjects viewed through an angled mirror mounted to the head coil. Subjects made their responses on a two-button response box in each hand. While subjects were playing the task, we acquired gradient echo T2*-weighted echo-planar images (EPIs) with BOLD contrast (approx. 400 volumes per subject, 32 slices per volume, ascending order, Field of View 192 x 192 x 110 mm, slice thickness 3 mm, gap 0.45 mm, repetition time 2190 ms, echo time 30 ms, flip angle 90°). Volumes were acquired in axial orientation at a +15° tilt to the anterior commissure-posterior commissure line. After the functional session, T1-weighted 3D modified driven equilibrium Fourier transformation (MDEFT) images were acquired from each subject (176 slices, Field of View 256 x 224 x 176 mm, slice thickness 1 mm, no gap, repetition time 7.92 ms, echo time 2.48 ms, flip angle 16°). Preprocessing of the functional images was implemented in the MATLAB based software Statistical Parametric Mapping 12 (SPM12, version r6685; http://www.fil.ion.ucl.ac.uk/spm). Preprocessing included motion correction (realignment to the mean EPI), segmentation of the T1 image into six tissue classifications (gray matter, white matter, cerebro-spinal fluid, bone, soft tissue and air tissue), application of this segmentation to the mean EPI, coregistration of all EPIs to the mean EPI using the pullback procedure in the SPM12 deformation tool and normalization of all EPIs to MNI standard space (Montreal Neurological Institute, http://www.bic.mni.mcgill.ca) (Evans et al., 1993). Finally, we smoothed the EPIs with a 4 mm full width at half maximum Gaussian kernel.

Analysis aim and structure

The central aim of our analyses was to identify a neurophysiological mechanism that can explain individual differences in control-averse behavior in addition to or beyond self-report data. To this end, our analyses followed a hierarchical structure. First, we identified the best predictor of individual control-averse behavior based on self-report data. Second, we identified a neurophysiological mechanism that predicts individual control-averse behavior. Third, we identified the best combination of predictors based on both self-report and neural data. Fourth, we tested whether the neural predictor mediates the self-report data in predicting individual control-averse behavior.

Behavioral data analyses

All behavioral data were analyzed using the MATLAB Statistics and Machine Learning Toolbox (R2015b, MathWorks). Because the behavioral data did not follow normal distributions as assessed by Kolmogorov-Smirnov tests, nonparametric tests were applied. Paired samples were compared using the Wilcoxon signed rank test. Correlations were assessed using Spearman's rho as well as bisquare robust regressions. For all behavioral analyses, two-tailed *p* values are reported.

Identifying the best predictor of individual control-averse behavior based on self-report data

We first identified the best predictor of individual control-averse behavior based on self-report data. To this end, we ran a series of generalized linear models using the function fitglm as implemented in the MATLAB Statistics and Machine Learning Toolbox (R2015b, MathWorks). For each model, the dependent variable was the individual level

of control-averse behavior, as measured by the mean chosen level in the Free condition minus the mean chosen level in the Controlled condition. The self-report variables served as predictors. For conciseness, we report only models with predictors that showed a significant correlation with individual control-averse behavior. To reduce multicollinearity among the predictors, we computed two new variables using principal component analysis as implemented in the MATLAB function pca. The new variable 'social cognition' is the first principal component of the normalized ratings of the item 'perceived distrust' (coefficient 0.88) and the reversed item 'understanding' (coefficient 0.48). The second new variable 'negative affect' is the first principal component of the normalized mean unhappiness rating (coefficient 0.80) and the normalized mean anger rating in the Controlled minus the Free condition (coefficient 0.59). As predictors we used combinations of main effects and interactions of 'social cognition', 'negative affect', and the normalized ratings of the item 'freedom restoration'. The most relevant models are illustrated in Figure **4**. We compared the models using the Bayesian information criterion (BIC) and R^2 to identify the best model fit. Lower values in BIC and greater values in R^2 indicate better model fits.

fMRI data analyses

The statistical analysis of the fMRI data was also carried out in SPM12 (version r6685). We modeled each subject's BOLD response with a General Linear Model (GLM) that was estimated using SPM12's standard hemodynamic response function and a high pass filter of 128 Hz as well as correction for intrinsic autocorrelations. SPM12's internal masking threshold for the estimation of the beta parameters was set to 0.4 to ensure inclusion of subcortical brain regions. The GLM contained two regressors of interest as boxcar functions: (1) decisions in the Controlled condition, and (2) decisions in the Free

condition (each with a duration from the respective onset of the choice options until the first button press, illustrated as 'decision window' in Fig. **1**). Note that due to a high consistency in the subjects' choices and therefore in the subjects' and player A's payoff within each condition and subject (Fig. **2**B), it was not feasible to additionally control for the subjects' or player A's payoff in the GLM. As nuisance regressors, we modeled (3) the display of the text "A new player A is deciding…" (duration 2.4-8.6 s), (4) motor response (duration from the first button press until press of the OK button), (5) unhappiness rating (duration = reaction times), (6) anger rating (duration = reaction times), (7) manipulation check, i.e. feeling of being controlled rating (duration = reaction times), (8) six motion parameters. For every subject we created contrast images for the two regressors of interest.

At the group level, we used random effects analyses. For all random effects analyses, we applied whole-brain correction for multiple comparisons at the cluster level: We calculated the corrected cluster extent (k_E) for each *t* test using Gaussian Random-field theory as implemented in SPM12 with a cluster-defining individual voxel threshold of $t = 2.68$ ($p < 0.005$) to achieve a FWE-corrected statistical threshold of $p_{FWE} < 0.05$ (minimum $k_E > 40$, range 40-44).

The aim of the fMRI analysis was to identify a neurophysiological mechanism that can predict individual differences in control-averse behavior. Specifically, we investigate whether activations in and interactions with the brain regions that are differentially activated for decisions in the Controlled and the Free condition correlate with individual control-averse behavior. We did so in three fMRI analysis steps, which will be described in the following sections.

fMRI analysis step 1: Localization of brain regions differentially activated for decisions in the Controlled and the Free condition

To identify the brain regions that are differentially activated during decisions in the Controlled and the Free condition, we tested the corresponding contrast images in a paired *t* test at the group level. Because we had no strong anatomical hypotheses, we applied whole-brain corrected analysis. Based on the paired *t* test, we created two masks for all suprathreshold voxels within a 10-mm sphere around the group peak voxel in the right and left inferior parietal lobule (IPL), respectively, at a threshold of $p < 0.005$, uncorrected (peak MNI coordinates for right IPL: 39 −40 40; for left IPL: −42 −40 47, illustrated in Fig. **5**). The spheres were applied to isolate the activation in the IPL from more posterior activation. The masks were used to extract and illustrate the mean beta estimates as implemented in the MarsBaR toolbox (Brett et al., 2002), as well as for search volumes in the functional connectivity analyses and time course analyses that will be described in the fMRI analysis step 3.

fMRI analysis step 2: Covariate analysis of activation differences for decisions in the Controlled and the Free condition and control-averse behavior

The second step of the fMRI analysis was to investigate whether individual control-averse behavior could be predicted by activation differences for decisions in the Controlled and Free condition. To test this, we included the individual level of control-averse behavior as a covariate in the paired *t* test (random effects analysis), using a whole-brain analysis. The individual level of control-averse behavior was computed as the mean chosen level in the Free condition minus the mean chosen level in the Controlled condition, with the result that increasing values reflect increasing levels of control-averse behavior.

366

367 *fMRI analysis step 3: Covariate analysis of the functional connectivity seeded in the IPL*
368 *and control-averse behavior*

369 The third step of the fMRI analysis was to investigate whether individual control-averse
370 behavior could be explained by neural interactions with the brain regions that are
371 differentially active for decisions in the Controlled and Free condition. For this purpose
372 we conducted functional connectivity analyses seeded in the right and left IPL as
373 identified in the paired *t* test for decisions in the Controlled > Free condition. To assess
374 the functional connectivity, we used psychophysiological interaction (PPI) analysis with
375 two psychological factors of interest that were derived from the GLM: (1) decisions in the
376 Controlled condition; (2) decisions in the Free condition. We extracted single-subject
377 time courses in the right and the left IPL, respectively, as follows: using the search
378 volumes derived from the paired *t* test for decisions in the Controlled > Free condition at
379 the group level (illustrated in Fig. **6**), we identified, for each subject, the peak *Z* value for
380 the contrast of decisions in the Controlled > Free condition and extracted the first BOLD
381 signal eigenvariate from a 5-mm sphere around this individual peak. This approach was
382 chosen to account for between-subject variability in the spatial location of the peak
383 activation. The extracted BOLD signal eigenvariate was then deconvolved and multiplied
384 with the two psychological factors of interest to create the PPI terms (Controlled PPI,
385 Free PPI), which were then convolved with the standard SPM12 hemodynamic response
386 function. Lastly, for each seed, the two PPI terms, the BOLD signal eigenvariate, and all
387 regressors described in the GLM were entered into a new GLM (GLM-PPI). For all
388 subjects, we created contrast images for the two PPI terms. To identify brain regions that
389 show an increased functional connectivity with the right and left IPL, respectively, we
390 tested the associated contrast images Controlled PPI > Free PPI in two separate paired *t*

391 tests at the group level (random effects analyses). Finally, to test whether the functional
392 connectivity seeded in the IPL predicts control-averse behavior, we included the
393 individual level of control-averse behavior as a covariate in the paired *t* tests of
394 Controlled PPI > Free PPI (random effects analyses), using whole-brain analyses.
395 Based on the covariate analysis, we created two new masks for all suprathreshold
396 voxels in the right and left dorsolateral prefrontal cortex (dlPFC)/ middle frontal gyrus,
397 respectively, at a threshold of $p < 0.005$, uncorrected (Table **1**, Fig. **6**). These masks
398 were used to extract and illustrate the mean beta estimates as implemented in the
399 MarsBaR toolbox (Brett et al., 2002) (Fig. **5**), and as search volumes for additional time
400 course analyses (Fig. **6**) as follows.
401 To further examine individual differences in the temporal characteristics of the BOLD
402 signal underlying the decisions in the Controlled and Free condition in the seed (bilateral
403 IPL) and target regions (bilateral dlPFC/ middle frontal gyrus) of the functional
404 connectivity analysis, we performed *post hoc* time course analyses using the search
405 volumes as described above. For each subject and each search volume, we identified
406 the peak *Z* value for the contrast of decisions in the Controlled > Free condition and
407 extracted the raw event-related BOLD response from a 5-mm sphere around this
408 individual peak, which was identical to the procedure used in the PPI analysis. Event-
409 related BOLD responses were estimated by two Finite Impulse Response models for
410 decisions in the Controlled condition and decisions in the Free condition, respectively,
411 adjusted for nuisance effects of the motion regressors and resampled to time bins of
412 0.5 s as implemented in the rfxplot toolbox (Gläscher, 2009). We then divided the
413 subjects into groups of not control-averse subjects (with levels of control-averse
414 behavior ≤ 0, n = 10) and control-averse subjects (with levels of control-averse behavior
415 > 0, n = 41) and plotted the averaged time courses across subjects in each group

separately for decisions in the Controlled and the Free condition (Fig. **6**). Note, that the raw event-related BOLD signal is independent of any model assumptions. The time course analyses therefore provide additional insights into the temporal characteristics of the BOLD signal in the target regions. Due to the use of non-independent masks, however, it is important to note that the time course analyses were not used to infer the magnitude of the effect Controlled > Free condition.

Identifying the best combination of predictors of individual control-averse behavior based on self-report and neural data

Building upon the behavioral results and the result of the functional connectivity analysis, we next investigated whether models based on self-report data could be improved by including neural data. To this end, we ran a new series of generalized linear models using the function fitglm as implemented in the MATLAB Statistics and Machine Learning Toolbox (R2015b, MathWorks). For each model, the dependent variable was the individual level of control-averse behavior, as measured by the mean chosen level in the Free condition minus the mean chosen level in the Controlled condition.

We compared the best model based on self-report data with models based on the neural data and combinations of neural and self-report data. As neural predictor we used the difference between the subject-wise estimate of the connectivity between right IPL and right dlPFC during decisions in the Controlled and the Free condition (Controlled PPI − Free PPI). This neural predictor was combined with main effects of and interactions with the predictors ‘social cognition’, ‘negative affect’, and ‘freedom restoration’. The most relevant models are illustrated in Figure **7**. Again, we compared the models with regard to the BIC and R^2.

Mediation analysis of self-report and neural predictors of individual control-averse behavior

Building upon the result of the model comparisons, we next investigated the association between social cognition, the right IPL-dlPFC connectivity and control-averse behavior. To this end, we performed a mediation analysis using the MATLAB based mediation toolbox described by Wager et al. (2008, https://github.com/canlab/MediationToolbox). We based the test on three criteria, which are illustrated in the three-variable path model in Figure **8**. First, the predictor must be related to the mediating variable (path a). Second, the mediator must be related to the outcome after controlling for the predictor (path b). Third, the mediation effect defined as product of the a and b path coefficients (a*b) must be significant. A significant mediation effect indicates that the mediator significantly reduces and therefore explains the predictor-outcome relationship (difference between path c and c'). If the predictor still explains significant variance in the outcome after controlling for the mediator (path c'), we speak of a partial mediation. A mediation analysis is conceptually different from a moderation analysis (see model 10 in Fig. **7**), which tests whether the level of the moderating variable can predict the strength of the relationship between the predictor and the outcome (Baron and Kenny, 1986; Wager et al., 2008). In other words, a moderator indicates *when* a predictor-outcome association occurs, whereas a mediator explains *how or why* such an effect occurs (Baron and Kenny, 1986). We therefore ran the mediation analysis to test whether the right IPL-dlPFC connectivity represents the mechanism through which social cognition affects control-averse behavior.

As the predictor we used the subject-specific variable 'social cognition'. The mediator was the difference between the subject-wise estimate of the connectivity between right IPL and right dlPFC during decisions in the Controlled and the Free condition (Controlled

PPI − Free PPI). The outcome was the individual level of control-averse behavior, as measured by the mean chosen level in the Free condition minus the mean chosen level in the Controlled condition. Statistical significance was assessed using a bootstrap test with 1000 samples.

RESULTS

Behavioral results

Control-averse behavior and its association with negative affect, perceived distrust, understanding, and freedom restoration

While lying in the fMRI scanner, subjects made choices under two conditions (Fig. **1**): In the Free condition, subjects could choose freely among five allocation options, called generosity levels, ranging from selfish to more generous and equal monetary allocations between themselves and another person. In the Controlled condition, the other person requested a minimum of level two and thereby eliminated the most selfish and unequal option. A manipulation check showed that subjects indeed indicated having more control in the Free condition (mean 4.42 ± SD 0.73, median 4.75) than in the Controlled condition (mean 3.88 ± SD 0.88, median 4.00; Wilcoxon signed rank test, two-tailed, $Z = 4.69$, $p < 0.001$, Hodges-Lehmann estimator of differences 0.63, 95 % CI [0.38, 0.94], Fig. **2**A).

First, we tested whether the restriction of the freedom of choice had an effect on subjects' generosity as measured by the chosen generosity level. As expected, subjects chose, on average, lower generosity levels in the Controlled condition (mean 3.50 ± 0.78 SD, median 3.50) than in the Free condition (mean 4.34 ± 0.57 SD, median 4.50; Wilcoxon signed rank test, two-tailed, $Z = -5.64$, $p < 0.001$, Hodges-Lehmann estimator

of differences −1.00, 95 % CI [−1.19, −0.81], Fig. **2**B). Note that the statistical test was corrected for a bottom effect, following the procedure by Falk and Kosfeld (2006). Subjects demonstrated high consistency in their choice preferences: they showed a variance of mean 0.31 ± 0.33 SD, median 0.21 in the Controlled condition and a variance of mean 0.33 ± 0.37 SD, median 0.21 in the Free condition (Fig. **2**C). We therefore averaged each subjects' choices within each condition and used the difference between each subject's mean chosen level in the Free condition minus the subject's mean chosen level in the Controlled condition as the measure of the individual level of control-averse behavior. Notably, the individual levels of control-averse behavior varied from −0.25 to 2.13 (mean 0.82 ± 0.64 SD, median 0.88), a variation that stems mostly from the mean chosen levels in the Controlled condition rather than the Free condition as illustrated in in Figure **2**B-C. In other words, subjects chose similarly high levels in the Free condition, whereas choices are more heterogeneous in the Controlled condition. For two subjects, the level of control-averse behavior was −0.25, which did not result from systematic choices, but rather from a single outlier choice of a lower level in the Free condition. Because these subjects otherwise demonstrated zero difference in their choices between the two conditions, they were treated as not control-averse.

Second, we tested whether subjects' individual control-averse behavior was associated with negative affects (Dillard and Shen, 2005). To capture negative affects, we used trial-by-trial anger ratings of unhappiness and anger on pictorial 5-point SAM scales (Bradley and Lang, 1994). Indeed, we found a significant association of control-averse behavior with both negative affect ratings: the unhappier (Spearman's rho = 0.49, $p < 0.001$; robust $R^2 = 0.26$, $p < 0.001$) and the angrier (Spearman's rho = 0.46, $p = 0.001$; robust $R^2 = 0.23$, $p < 0.001$) subjects were in the Controlled compared with the Free condition, the greater was their individual level of control-averse behavior (Fig.

516 **3**A). To additionally assess trait anger expression, we used a task-independent anger
517 expression inventory (STAXI, Schwenkmezger et al., 1992). Subjects' general tendency
518 to direct anger expression outward, however, did not correlate significantly with the
519 individual level of control-averse behavior (Spearman's rho = −0.01, robust $R^2 < 0.01$,
520 both $p > 0.9$, Fig. **3**A). Other subscales of the STAXI also showed no significant
521 association with control-averse behavior.
522 Third, we tested the association between subjects' individual control-averse behavior
523 and their self-reported thoughts as assessed by ratings after scanning. For each rating,
524 subjects were asked to indicate how strongly the described thought had influenced their
525 decision in the Control aversion task. Consistent with previous work (Falk and Kosfeld,
526 2006), we found that subjects demonstrated more control-averse behavior the more they
527 perceived the choice restriction as a signal of distrust by the other person (Spearman's
528 rho = 0.60, robust $R^2 = 0.32$, both $p < 0.001$, Fig. **3**B). By contrast, subjects
529 demonstrated less control-averse behavior the higher they rated understanding the other
530 person's request in the Controlled condition (Spearman's rho = −0.66, robust $R^2 = 0.37$,
531 both $p < 0.001$). Then we tested whether the motivation for freedom restoration had
532 influenced the subjects' decisions. Consistent with reactance theory (Brehm, 1966;
533 Miron and Brehm, 2006), our subjects' self-reported motivation to use their remaining
534 freedom of choice correlated significantly and positively with their level of control-averse
535 behavior (Spearman's rho = 0.37, $p = 0.008$, robust $R^2 = 0.17$, $p = 0.003$, Fig. **3**B).
536 Lastly, we asked subjects whether fairness had played a role in their decisions, i.e. the
537 thought that their own payoff and the other person's payoff should not be too far apart.
538 Interestingly, fairness correlated positively with the average chosen level within both the
539 Controlled condition (Spearman's rho = 0.51, robust $R^2 = 0.28$, both $p < 0.001$) and the
540 Free condition (Spearman's rho = 0.48, robust $R^2 = 0.26$, both $p < 0.001$), but was not

significantly associated with control-averse behavior (Spearman's rho = −0.20, p = 0.163, robust R^2 = 0.04, p = 0.144, Fig. **3**B).

Social cognition is the best self-report predictor of individual control-averse behavior

Next, we aimed to identify the best predictor of individual control-averse behavior based on self-report data. To this end, we computed and compared a series of generalized linear models. As predictors we focused on the self-reported variables that showed a significant correlation with control-averse behavior (Fig. **3**). To reduce multicollinearity among the predictors, we applied principal component analyses and computed the new variables 'social cognition' and 'negative affect'. The normalized ratings of the item 'freedom restoration' served as a third predictor. Model comparisons revealed that, based on the self-report data, the following model had the best model fit (Fig. **4**, Table **2**):

$$y_i = \beta_0 + \beta_1 SocialCognition_i + \varepsilon_i$$

where y_i is the level of control-averse behavior for subject i, and $SocialCognition$ is the first principal component of the normalized ratings of the items 'perceived distrust' and the reversed item 'understanding'. This model performed better in predicting individual control-averse behavior than any model that included negative affect or the motivation for freedom restoration either as main effects or interaction terms.

Neuroimaging results

Control-averse behavior is predicted by neural interactions between the right IPL and the dorsolateral prefrontal cortex (dlPFC)

The aim of the fMRI analysis was to identify a neurophysiological mechanism that can predict control-averse behavior. Specifically, we aimed to test whether neural responses and their interactions could explain individual differences in control-averse behavior. To do this, we ran covariate analyses between the individual control-averse behavior and neural activity in the brain regions that are differentially activated during decisions in the Controlled and the Free condition, as well as the functional connectivity seeded in these brain regions.

In a first step, the brain regions that are more strongly activated during decisions in the Controlled than in the Free condition were localized. We estimated a GLM that models the BOLD responses for decisions in the Controlled and the Free condition, respectively. The respective single-subject contrast images were then compared in a paired *t* test. We found that the right IPL (peak MNI coordinates 39 −40 40, $t = 3.99$, $p_{FWE} < 0.001$, whole-brain family-wise error (FWE)-corrected at the cluster level), the left IPL (peak MNI coordinates −42 −40 47, $t = 3.76$, $p_{FWE} = 0.042$), clusters in the bilateral superior parietal lobule extending into the occipital cortex (peak MNI coordinates right 15 −73 57, $t = 4.42$, $p_{FWE} < 0.001$; left −21 −64 43, $t = 4.43$, $p_{FWE} < 0.001$) and the right occipital cortex (peak MNI coordinates 39 −79 33, $t = 4.01$, $p_{FWE} = 0.042$) were more strongly activated during decisions in the Controlled than in the Free condition.

In a second step, we tested whether these activation differences between decisions in the Controlled and in the Free condition could explain individual differences in control-averse behavior by including the individual level of control-averse behavior as a covariate in the paired *t* test of the contrast images for decisions in the Controlled and

the Free condition. This covariate analysis revealed no significant association between control-averse behavior and the activation differences between decisions in the Controlled and the Free condition, even at a more liberal statistical threshold of $p < 0.005$, uncorrected.

In a third step, we asked whether individual differences in control-averse behavior could instead be explained by functional connectivity patterns. As the seed region of the functional connectivity, we focused on the bilateral IPL due to its suggested role in subjective choice restrictions (Filevich et al., 2013) and attention reorientation (Corbetta et al., 2008). Accordingly, the above described peak activation clusters in the bilateral IPL were used as search volumes for individual subjects' seeds for the functional connectivity analyses (illustrated in Fig. **5** and **6**). To assess the functional connectivity, we performed two psychophysiological interaction (PPI) analyses that included separate interaction terms between the right and left IPL BOLD time series, respectively, and regressors indicating decisions in the Controlled and the Free condition (Controlled PPI, Free PPI). We searched for brain regions whose functional connectivity with the IPL predicted control-averse behavior by including the individual level of control-averse behavior as a covariate in the paired *t* test of the contrast images for Controlled PPI > Free PPI. Whereas the covariate analysis seeded in the left IPL revealed no significant results, we found that for Controlled PPI > Free PPI, the right IPL showed increased functional coupling with the right dorsolateral prefrontal cortex (dlPFC)/ middle frontal gyrus ($p_{FWE} < 0.001$), the left angular gyrus ($p_{FWE} < 0.001$), the right precuneus ($p_{FWE} = 0.047$), the left dlPFC ($p_{FWE} = 0.042$) and the left IPL ($p_{FWE} = 0.033$) as a function of control-averse behavior (Fig. **5**, Table **1**). No significant negative association was observed. Complementary PPI analyses seeded in the superior parietal lobule and the occipital cortex revealed no significant association with control-averse behavior. To

inspect whether the positive correlation was driven by either one of the conditions, we extracted the mean beta estimates across the functional clusters of the bilateral dlPFC for the Controlled PPI and the Free PPI regressor, separately, and plotted them against the individual level of control-averse behavior (Fig. **5**). This inspection revealed that right IPL-dlPFC connectivity during the decisions increased with control-averse behavior in the Controlled condition and decreased with control-averse behavior in the Free condition. Hence, the higher the individual level of control-averse behavior, the greater is the change in right IPL-dlPFC connectivity during decisions in the Controlled in contrast to the Free condition. In addition, time course analyses showed that activation in the bilateral IPL increases immediately after the onset of the choice options, irrespective of individual control-averse behavior (Fig. **6**). In contrast, activation in the bilateral dlPFC synchronizes with activation in the IPL only for control-averse subjects and only during decisions in the Controlled condition.

The connectivity between right IPL and dlPFC complements self-reported social cognition in predicting individual control-averse behavior

Next, we aimed to identify the best combination of predictors of control-averse behavior based on both self-report and neural data. Specifically, we tested whether the functional connectivity with the IPL complements or exceeds the self-reports in predicting control-averse behavior. To this end, we computed a set of new generalized linear models that included the neural data. As neural predictor, 'PPI', we used the subject-wise beta estimate of the Controlled PPI minus the Free PPI regressor between the right IPL and the right dlPFC. We focused on the connectivity of the right IPL with the dlPFC because of their frequent coactivation during attention reorientation (Corbetta et al., 2008) and context-dependent decision making (Daw et al., 2006; Boorman et al., 2009; Rudorf and

Hare, 2014). This neural predictor was combined with main effects of and interactions with the predictors 'social cognition', 'negative affect', and 'freedom restoration'. Model comparisons revealed that a model that combined main effects of 'social cognition' and 'PPI' had the best overall model fit (model 9, Fig. **7**, Table **2**):

$$y_i = \beta_0 + \beta_1 SocialCognition_i + \beta_2 PPI_i + \varepsilon_i$$

This model performed better than any combination of the neural predictor with any other predictors based on self-report data (Fig. **7**). Moreover, it performed slightly better than a model including the interaction of 'social cognition' and 'PPI' (BIC = 65, R^2 = 0.60, model 10 in Fig. **7**), which revealed no significant interaction and therefore no moderation effect (β = −0.40, $t_{(49)}$ = −0.38, p = 0.702, 95% CI [−2.52, 1.71]). When we added the other self-report predictors (model 13 in Fig. **7**), the main effects of 'social cognition' and 'PPI' remained robust, whereas the other predictors showed no significant effect. Accordingly, the increase of connectivity between the right IPL and right dlPFC in the Controlled compared with the Free condition explains variance in individual control-averse behavior that exceeds model predictions based on self-report data.

The connectivity between right IPL and dlPFC partially mediates the association of social cognition with control-averse behavior

After having identified social cognition and the right IPL-dlPFC connectivity as the best predictors of individual control-averse behavior, we asked whether the connectivity might reflect the mechanism through which these social cognitions affect control-averse behavior and therefore capture joint variance. To test this question, we ran a mediation analysis using a three-variable path model (Fig. 8, Baron and Kenny, 1986; Wager et al., 2008), in which the predictor is 'social cognition', the mediator is the subject-wise beta estimate of the Controlled PPI minus the Free PPI regressor between the right IPL and

659 the right dlPFC, and the outcome is the individual control-averse behavior. Following
660 convention (Baron and Kenny, 1986), we considered the mediation to be significant if
661 three conditions were met: the predictor must be related to the mediator (path a), the
662 mediator must be related to the outcome after controlling for the predictor (path b), and
663 the mediation effect, i.e. the product of the a and b path coefficients (a*b = c−c'), must
664 be significant. The mediation analysis revealed that the relationship between social
665 cognition and control-averse behavior is partially mediated by the connectivity between
666 right IPL and right dlPFC, i.e. the mediator significantly reduces the association between
667 predictor and outcome (total effect, path c), but the predictor still explains significant
668 variance of the outcome (direct effect, path c', Fig. **8**). In other words, the right IPL-
669 dlPFC connectivity explains a significant part of the relationship between social cognition
670 and control-averse behavior, but the predictor and mediator each also explain
671 independent variance.
672
673
674 **DISCUSSION**
675 People value their freedom of choice highly. Interestingly, though, if another person tries
676 to restrict one's choice, some people will comply, whereas others will act against the
677 restriction. These individual differences in control-averse behavior have been well
678 documented, but their driving factors have remained a puzzle. Previous work has
679 suggested several potential predictors of control-averse behavior based on self-reports.
680 To date, however, we know very little about the mechanisms that underlie control-averse
681 behavior at the neural level. Here, we identify a neural mechanism that complements
682 and exceeds self-reported social cognitions, affects and motivations in explaining
683 individual differences in control-averse behavior.

To do so, we combined fMRI with a Control aversion task (Falk and Kosfeld, 2006; Schmelz and Ziegelmeyer, 2015), in which subjects' freedom of choice is controlled by another person, and subjects' subsequent monetary allocation to that person serves as a measure of control-averse behavior. Specifically, we aimed to identify neural mechanisms that could explain individual differences in control-averse behavior. Our results both replicate prior behavioral studies and provide novel insights into the neurobiological basis of control-averse behavior. We replicated that control of one's freedom of choice by another person reduces the willingness to allocate money to that person (Falk and Kosfeld, 2006; Schmelz and Ziegelmeyer, 2015). This effect was augmented in subjects who had little understanding for the other person's behavior or who perceived the restriction of their freedom of choice as a signal of distrust in their intrinsic motivation to choose a generous and fair allocation (Falk and Kosfeld, 2006). We also found that control-averse behavior was accompanied by negative affects (Dillard and Shen, 2005) and the motivation to restore one's freedom of choice (Brehm, 1966; Miron and Brehm, 2006). This is in line with previous research on reactance that has focused on behavioral intentions in hypothetical scenarios (Sittenthaler et al., 2015) or behavior in non-social settings (Chartrand et al., 2007). Our study complements and extends this research by providing evidence of the motivation to act against the restriction of one's freedom of choice during social decisions with actual consequences. A direct comparison of the predictors based on the self-report data revealed that a combination of the social cognitions perceived distrust and understanding explained individual control-averse behavior best at the behavioral level.

At the neural level, we found that control-averse behavior could be predicted by functional connectivity between the right IPL and the bilateral dlPFC/ middle frontal gyrus. Notably, our finding is specific to the right IPL, which corroborates previous work

examining its role in subjective choice restrictions (Filevich et al., 2013). The involvement of both IPL and dlPFC in control-averse behavior could be attributed to their functions suggested in previous neuroimaging studies. The IPL has traditionally been associated with the reorienting of attention to both social and non-social stimuli (Corbetta et al., 2008) as well as number processing (Dehaene et al., 2003). Also, more recent work has linked the IPL to social distance encoding (Chiao et al., 2009; Parkinson et al., 2014), suggesting that the IPL might perform analogous operations in visuospatial and social contexts (Yamazaki et al., 2009; Parkinson et al., 2014). Therefore, it seems plausible that the differential IPL activation during decisions in the Controlled compared with the Free condition might reflect the encoding of or attention reorientation to the context (i.e. being controlled or not) that is relevant for the decision (i.e. to counteract or not). The differential IPL activation alone, however, did not explain individual differences in control-averse behavior, suggesting that the IPL encodes the difference between the Controlled and the Free condition irrespective of the subjects' individual control aversion. Instead, individual differences in control-averse behavior could be explained by the connectivity of right IPL with the dlPFC, two regions that are directly connected through fiber tracts (Mars et al., 2012). Moreover, the IPL and regions in the lateral PFC show robust intrinsic functional coupling (Mars et al., 2011) as well as increased task-based coupling during changes of choice strategy (Daw et al., 2006; Boorman et al., 2009). Follow-up studies could investigate whether individual differences in anatomical or resting state functional connectivity between the IPL and dlPFC might contribute to control-averse behavior.

The dlPFC has been commonly associated with cognitive control (MacDonald et al., 2000; Miller and Cohen, 2001) and overcoming conflicts in decisions that require self-control (Knoch et al., 2006; Hare et al., 2009; Figner et al., 2010; Baumgartner et al.,

2011). Correspondingly, the notion that control-averse behavior requires cognitive control is supported by our behavioral data: Although all subjects demonstrated an intrinsic motivation to choose a high level, control-averse subjects were more likely to dislike the restriction of their freedom of choice and to feel the urge to use their remaining freedom of choice. This suggests that control-averse subjects perceived the decisions in the Controlled condition as a conflict between the general motivation to choose a high level and the condition-specific motivation to act against the restriction. Given its suggested role in cognitive control, this could explain why the dlPFC was more strongly recruited by control-averse subjects during decisions in the Controlled condition as indicated by the connectivity analysis and illustrated in the time course plots. Furthermore, model comparisons indicate that the right IPL-dlPFC connectivity explains additional variance of the individual control-averse behavior that has remained unexplained by self-reports alone. More specifically, we find that the neural data complement the self-reports of social cognitions. Together, these two predictors explain a sizable amount of variance in the control-averse behavior and provide the best data fit among the tested models. Notably, the IPL cluster that we find lies in close proximity to the temporoparietal junction (Mars et al., 2012; Igelstrom et al., 2015), which is considered a key region in social cognition (Decety and Lamm, 2007; Cabeza et al., 2012; Carter and Huettel, 2013; Krall et al., 2015). It has been proposed that the IPL shares information with the temporoparietal junction via joint connections in the dlPFC/ middle frontal gyrus (Corbetta et al., 2008), matching the target region of our connectivity analysis. In line with this notion, we found that the right IPL-dlPFC connectivity partially mediates the association between social cognition and control-averse behavior. The partial mediation and model comparisons further suggest that the right IPL-dlPFC connectivity explains variance that could not be captured by self-reports. This

emphasizes once more that for a comprehensive understanding of a complex human behavior such as control-averse behavior, it is essential to incorporate neurophysiological factors. Although the IPL and dlPFC certainly have intricate roles in decision making, together our data provide evidence that the Controlled condition represents a socially salient event and that the right IPL-dlPFC connectivity might contribute to the integration of social cognition into control-averse behavior.

Lastly, it is important to acknowledge limitations of our study and provide suggestions on how to address them in future work. First, it would be interesting to see whether our results generalize to non-social scenarios. Falk and Kosfeld (2006) have demonstrated, however, that replacing the player A with a computer algorithm eliminates control-averse behavior, suggesting that the aversion to the choice restriction might be confounded with the social aspect in our task. Therefore, designing a study that analogously varies the degree of choice restrictions in both a social and non-social context could be an interesting future endeavor.

Furthermore, we opted for a small number of trials to increase credibility and limit possible habituation and attention biases. This means that, while our neuroimaging results survive whole-brain correction, some brain activation might have gone undetected. Using a greater number of trials, however, would have come at the risk of a less robust measure of control-averse behavior. In the current data, the robustness of our measure of control-averse behavior is supported by the consistent correlations with the affect and self-report ratings. Similar sanity checks should be incorporated in future neuroimaging studies on control-averse behavior.

To sum up, this study provides first insights into the neural drivers of individual differences in control-averse behavior, a social phenomenon that is ubiquitous in our society. The prevalence of control-averse behavior and its potential negative

784 consequences have become evident in previous behavioral studies. Advancing our
785 understanding of the mechanisms that give rise to individual differences in control-
786 averse behavior therefore represents an important research goal. Here, we have
787 approached this goal by identifying a neural mechanism that can explain individual
788 differences in control-averse behavior. Our results suggest that a key driver of control-
789 averse behavior is the connectivity between brain regions that are reliably, albeit not
790 exclusively, involved in attention reorientation and cognitive control. This connectivity
791 complements what could be measured by self-reports alone and thereby improves our
792 understanding of the mechanisms underlying control-averse behavior. While more work
793 is needed to investigate the exact neural computations and extend these findings to
794 more complex social interactions, this study has brought us a significant step forward in
795 unraveling the drivers of individual differences in control-averse behavior.
796
797

REFERENCES

Baron RM, Kenny DA (1986) The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. J Pers Soc Psychol 51:1173–1182.

Baumgartner T, Knoch D, Hotz P, Eisenegger C, Fehr E (2011) Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. Nat Neurosci 14:1468–1474.

Boorman ED, Behrens TEJ, Woolrich MW, Rushworth MFS (2009) How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. Neuron 62:733–743.

Bradley M, Lang PJ (1994) Measuring emotion: the Self-Assessment Manikin and the semantic differential. J Behav Ther Exp Psychiatry 25:49–59.

Brehm JW (1966) A theory of psychological reactance. New York: Academic Press.

Brett M, Anton J-L, Valabregue R, Poline J-B (2002) Region of interest analysis using an SPM toolbox. Neuroimage 16:CD-ROM.

Cabeza R, Ciaramelli E, Moscovitch M (2012) Cognitive contributions of the ventral parietal cortex: An integrative theoretical account. Trends Cogn Sci 16:338–352.

Carter RM, Huettel SA (2013) A nexus model of the temporal-parietal junction. Trends Cogn Sci 17:328–336.

Chartrand TL, Dalton AN, Fitzsimons GJ (2007) Nonconscious relationship reactance: When significant others prime opposing goals. J Exp Soc Psychol 43:719–726.

Chiao JY, Harada T, Oby ER, Li Z, Parrish T, Bridge DJ (2009) Neural representations of social status hierarchy in human inferior parietal cortex. Neuropsychologia 47:354–363.

Corbetta M, Patel G, Shulman GL (2008) The reorienting system of the human brain:

from environment to theory of mind. Neuron 58:306–324.

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876–879.

De las Cuevas C, Peñate W, Betancort M, de Rivera L (2014) Psychological reactance in psychiatric patients: Examining the dimensionality and correlates of the Hong Psychological Reactance Scale in a large clinical sample. Pers Individ Dif 70:85–91.

Decety J, Lamm C (2007) The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. Neuroscientist 13:580–593.

Dehaene S, Piazza M, Pinel P, Cohen L (2003) Three parietal circuits for number processing. Cogn Neuropsychol 20:487–506.

Dillard JP, Shen L (2005) On the nature of reactance and its role in persuasive health communication. Commun Monogr 72:144–168.

Evans AC, Collins DL, Mills SR, Brown ED, Kelly RL, Peters TM (1993) 3D statistical neuroanatomical models from 305 MRI volumes. In: IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference, pp 1813–1817. IEEE.

Falk A, Kosfeld M (2006) The hidden costs of control. Am Econ Rev 96:1611–1630.

Figner B, Knoch D, Johnson EJ, Krosch AR, Lisanby SH, Fehr E, Weber EU (2010) Lateral prefrontal cortex and self-control in intertemporal choice. Nat Neurosci 13:538–539.

Filevich E, Vanneste P, Brass M, Fias W, Haggard P, Kühn S (2013) Brain correlates of subjective freedom of choice. Conscious Cogn 22:1271–1284.

Gläscher J (2009) Visualization of group inference data in functional neuroimaging.

Neuroinformatics 7:73–82.

Hare TA, Camerer CF, Rangel A (2009) Self-control in decision-making involves modulation of the vmPFC valuation system. Science 324:646–648.

Igelstrom KM, Webb TW, Graziano MSA (2015) Neural Processes in the Human Temporoparietal Cortex Separated by Localized Independent Component Analysis. J Neurosci 35:9432–9445.

Knoch D, Pascual-Leone A, Meyer K, Treyer V, Fehr E (2006) Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science 314:829–832.

Krall SC, Rottschy C, Oberwelland E, Bzdok D, Fox PT, Eickhoff SB, Fink GR, Konrad K (2015) The role of the right temporoparietal junction in attention and social interaction as revealed by ALE meta-analysis. Brain Struct Funct 220:587–604.

MacDonald AW, Cohen JD, Stenger VA, Carter CS (2000) Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. Science 288:1835–1838.

Mars RB, Jbabdi S, Sallet J, O'Reilly JX, Croxson PL, Olivier E, Noonan MP, Bergmann C, Mitchell AS, Baxter MG, Behrens TEJ, Johansen-Berg H, Tomassini V, Miller KL, Rushworth MFS (2011) Diffusion-Weighted Imaging Tractography-Based Parcellation of the Human Parietal Cortex and Comparison with Human and Macaque Resting-State Functional Connectivity. J Neurosci 31:4087–4100.

Mars RB, Sallet J, Schüffelgen U, Jbabdi S, Toni I, Rushworth MFS (2012) Connectivity-based subdivisions of the human right "temporoparietal junction area": Evidence for different areas participating in different cortical networks. Cereb Cortex 22:1894–1903.

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202.

Miron AM, Brehm JW (2006) Reactance Theory-40 years later. Zeitschrift Für Sozialpsychologie 37:9–18.

Parkinson C, Liu S, Wheatley T (2014) A Common Cortical Metric for Spatial, Temporal, and Social Distance. J Neurosci 34:1979–1987.

Rudorf S, Hare TA (2014) Interactions between dorsolateral and ventromedial prefrontal cortex underlie context-dependent stimulus valuation in goal-directed choice. J Neurosci 34:15988–15996.

Schmelz K, Ziegelmeyer A (2015) Social distance and control aversion: Evidence from the internet and the laboratory. Thurgau Inst Econ Res Pap Ser 100:1–25.

Schwenkmezger P, Hodapp V, Spielberger C (1992) Das State-Trait-Ärgerausdrucks-Inventar (STAXI). Bern: Huber.

Sittenthaler S, Traut-Mattausch E, Steindl C, Jonas E (2015) Salzburger State Reactance Scale (SSR Scale): Validation of a scale measuring state reactance. Z Psychol 223:257–266.

Spielberger CD (1988) State-Trait Anger Expression Inventory (STAXI). Odessa, FL: Psychological Assessment Resources.

Van Petegem S, Soenens B, Vansteenkiste M, Beyers W (2015) Rebels with a cause? Adolescent defiance from the perspective of reactance theory and self-determination theory. Child Dev 86:903–918.

Wager TD, Davidson ML, Hughes BL, Lindquist M a, Ochsner KN (2008) Prefrontal-subcortical pathways mediating successful emotion regulation. Neuron 59:1037–1050.

Yamazaki Y, Hashimoto T, Iriki A (2009) The posterior parietal cortex and non-spatial cognition. F1000 Biol Rep 6:1:74.

Ziegelmeyer A, Schmelz K, Ploner M (2012) Hidden costs of control: Four repetitions

898 and an extension. Exp Econ 15:323–340.

899

FIGURE LEGENDS

Figure 1. Control aversion task. For every trial, the subject is presented with the decision from a new player A and the available generosity levels. Each generosity level represents an allocation of monetary units between the player A (top value) and the subject (bottom value). In the Free condition (blue frame), the player A lets the subject choose freely between level one to five (from left to right). In the Controlled condition (orange frame), the player A requests a minimum of level two and thereby restricts the subject's choice to the levels two to five. The decision window that is highlighted in the figure is defined as the time between the onset of the choice options and the initial movement of the red selection frame. Lastly, the subject is presented with three pictorial assessment scales, which range from unhappy to happy (left to right), from calm to angry, and from being controlled to having control. The durations of the fixation displays were jittered.

915 **Figure 2. Choice behavior. A-B**, boxplots of the ratings of having control and
916 chosen generosity levels, respectively, in the Controlled and the Free condition.
917 The central mark of each box shows the median, the box edges show the 25th and
918 75th percentiles, and the whiskers represent the limit beyond which a data point is
919 considered an outlier (denoted as cross). The connected data points in the center
920 show individual subject's means. **C**, The histograms show the distribution of
921 subjects' mean and variance of chosen levels in the Controlled and the Free
922 condition. Data from n = 51 subjects are shown.
923
924

925 **Figure 3. Correlation of control-averse behavior with negative affects,**
926 **perceived distrust, understanding, freedom restoration and fairness. A**, mean
927 unhappiness and anger ratings in the Controlled minus the Free condition and
928 individual tendencies for outward directed anger expression, respectively, plotted
929 against the individual control-averse behavior, computed as the difference between
930 the mean chosen level in the Free minus the Controlled condition. **B**, individual
931 ratings of perceived distrust, understanding, freedom restoration and fairness
932 plotted against individual control-averse behavior. Observations are jittered along
933 the x-axis to reduce overlap for visualization. Regression lines were fitted with
934 bisquare robust regressions. Data from n = 51 subjects are shown.
935

936 **Figure 4. Models based on self-report data.** These diagrams show seven models
937 predicting individual control-averse behavior (y), based on self-reports of social
938 cognition (S), freedom restoration (F), and negative affect (A). Arrows indicate main
939 effects. The bar graphs show the Bayesian Information Criterion (BIC) and R^2 for
940 each model, with the winning model highlighted in black.
941

Figure 5. Connectivity between right IPL and dlPFC predicts individual differences in control-averse behavior. The figure illustrates that the functional connectivity during decisions in the Controlled as opposed to the Free condition (Controlled PPI − Free PPI) between the right IPL (seed) and regions in the dlPFC/ middle frontal gyrus and the posterior parietal cortex increases as a function of individual control-averse behavior. Left, Statistical parametric maps of the covariate analysis are shown, color coded for the *t* values as indicated by the color bar, thresholded at $p_{FWE} < 0.05$, and projected on a template brain in MNI space. Right, Graphs show the individual level of control-averse behavior (x-axes) plotted against the single-subject means of the beta estimates extracted from the functional clusters in the right and left dlPFC (circled on the left) for the Controlled PPI − Free PPI effect, the Controlled PPI effect and the Free PPI effect seeded in the right IPL (y-axes). Observations are jittered along the x-axis to reduce overlap for visualization. Regression lines were fitted with bisquare robust regressions. Data from n = 51 subjects are shown. R, right; L, left; IPL, inferior parietal lobule; dlPFC, dorsolateral prefrontal cortex.

Figure 6. BOLD time courses of decisions in the Controlled and Free condition. The IPL shows a similar pattern for not control-averse subjects (with levels of control-averse behavior ≤ 0, n = 10) and control-averse subjects (with levels of control-averse behavior > 0, n = 41), whereas the dlPFC shows a distinct pattern for control-averse subjects. The graphs show averaged time courses of BOLD activation in the bilateral IPL (top row) and the bilateral dlPFC/ middle frontal gyrus (bottom row) for decisions in the Controlled (orange) and the Free condition (blue). The brain maps in the center depict the search volumes used for the time course extractions. The horizontal lines at the top of the graphs indicate time points at which the conditions differ significantly (Wilcoxon signed rank test, two-tailed, $p < 0.05$). The dashed vertical lines mark the onset of the decision window, at which the time courses were mean-corrected. The transparent areas show standard errors of the mean. Note that these plots were not used to infer the main effect of Controlled > Free condition. L, left; R, right; IPL, inferior parietal lobule; dlPFC, dorsolateral prefrontal cortex.

Figure 7. Models based on self-report and neural data. These diagrams show seven models predicting individual control-averse behavior (y), based on self-reports of social cognition (S), freedom restoration (F), negative affect (A), and subject-wise estimates of right IPL-dlPFC connectivity in the Controlled minus the Free condition (PPI). Arrows indicate main effects, and the line with a circular endpoint in model 10 indicates an interaction effect. The bar graphs show the Bayesian Information Criterion (BIC) and R^2 for each model, with the winning model highlighted in black.

Figure 8. Results of the mediation analysis testing the relationship between social cognition, right IPL-dlPFC connectivity and control-averse behavior. Left, the path model shows the path coefficients with standard errors of the mean in parentheses, significant at * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$. Right, histogram of the bootstrapped distribution of the mediation effect (a*b = c-c'). The lighter grey portion of each histogram denotes the 95% confidence interval for the effect. Data from n = 51 subjects were included in this analysis.

991 **TABLES**

992

993 **Table 1. Regions in which the connectivity for decisions in the Controlled minus**
994 **the Free condition (Controlled PPI − Free PPI) seeded in the right IPL is positively**
995 **associated with individual control-averse behavior.**

Region	Side	MNI coordinates x	y	z	Cluster size k_E	Max stat t	p_{FWE}
dlPFC/ middle frontal gyrus	R	42	47	22	105	4.88	<0.001
		24	50	5		4.58	
		48	35	29		4.16	
Angular gyrus	L	−33	−55	36	411	4.80	<0.001
		6	−70	50		4.67	
		27	−73	50		4.40	
Precuneus	R	18	−67	29	40	4.80	0.047
		3	−67	29		3.10	
		21	−58	26		3.01	
dlPFC/ middle frontal gyrus	L	−45	29	29	41	4.54	0.042
		−39	38	26		3.10	
		−45	35	19		2.95	
IPL	L	−39	−52	57	43	4.22	0.033
		−33	−58	57		3.35	
		−24	−64	60		2.95	

996 Results from the covariate analysis are shown (sample size n = 51 subjects). Height
997 threshold $t_{(49)}$ = 2.68, extent threshold k_E > 40. All activations survive whole-brain

998 correction for multiple comparisons based on FWE-control at the cluster level. dlPFC,
999 dorsolateral prefrontal cortex; IPL, inferior parietal lobule.
1000
1001
1002 **Table 2. Model comparison. Individual differences in control-averse behavior**
1003 **predicted by social cognition and right IPL-dlPFC connectivity in the Controlled**
1004 **minus the Free condition (models 1 and 9 in Fig. 7).**

Dependent variable: control-averse behavior												
model 1							model 9					
					95% CI						95% CI	
	β	SE	*t*	*p*	Lower	Upper	*β*	SE	*t*	*p*	Lower	Upper
Cognition	1.36	0.19	7.19	< 0.001	0.98	1.74	1.06	0.20	5.39	< 0.001	0.66	1.45
IPL-dlPFC connectivity							0.92	0.28	3.28	0.002	0.36	1.49
(Intercept)	0.84	0.06	13.51	< 0.001	0.72	0.97	0.43	0.14	3.16	0.003	0.16	0.71
BIC	68.1						61.7					
R^2	0.51						0.60					
Observations	51											

1005