


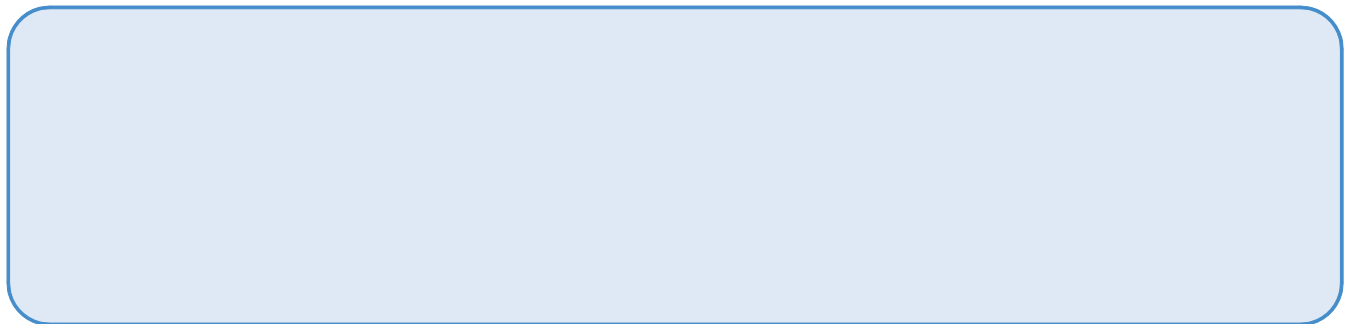


Right Temporoparietal Junction Underlies Avoidance of Moral Transgression in Autism Spectrum Disorder

Yang Hu,^{1,2,5*} Alessandra M. Pereira,^{3*}  Xiaoxue Gao,⁵ Bruno M. Campos,³ Edmund Derrington,^{2,4} Brice Corgnat,⁷  Xiaolin Zhou,^{1,5,6} Fernando Cendes,³ and  Jean-Claude Dreher^{2,4}

¹Key Laboratory of Applied Brain and Cognitive Sciences, School of Business and Management, Shanghai International Studies University, Shanghai 201620, People's Republic of China, ²Institut des Sciences Cognitives Marc Jeannerod, CNRS, Neuroeconomics Lab 69675 Bron, France, ³Neuroimaging Laboratory, School of Medical Sciences, The Brazilian Institute of Neuroscience and Neurotechnology, University of Campinas (UNICAMP), Campinas 13083-970, Brazil, ⁴Université Claude Bernard Lyon 1, 69100 Villeurbanne, France, ⁵School of Psychological and Cognitive Sciences, Peking University, Beijing 100871, People's Republic of China, ⁶PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, People's Republic of China, and ⁷EmLyon, 69130 Ecully, France

Autism spectrum disorder (ASD) is characterized by a core difference in theory-of-mind (ToM) ability, which extends to alterations in moral judgment and decision-making. Although the function of the right temporoparietal junction (rTPJ), a key neural marker of ToM and morality, is known to be atypical in autistic individuals, the neurocomputational mechanisms underlying its specific changes in moral decision-making remain unclear. Here, we addressed this question by using a novel fMRI task together with com-



Introduction

Autism spectrum disorder (ASD) is a complex neurodevelopment disorder with evident impairments in social interaction, communication, and interpersonal relationships (American Psychiatric Association, 2013), which are critically dependent on theory-of-mind (ToM) ability (Young et al., 2007; Margoni and Surian, 2016). These social problems are also associated with atypical moral cognition. Indeed, individuals with ASD have difficulties in evaluating the moral appropriateness of actions in terms of the intentions of the protagonists in hypothetical scenarios (Moran et al., 2011; Buon et al., 2013; Fadda et al., 2016). ASD individuals show differences in moral behaviors that lead to real consequences. For example, they are less sensitive to observation by others while making charitable decisions (Izuma et al., 2011).

Previous neuroimaging studies of healthy subjects lay a crucial foundation for improving our understanding of the neural basis underlying the atypical morality of ASD individuals. One of the key regions is the right temporoparietal junction (rTPJ), which is not only the hub of the ToM network (Schurz et al., 2014; Schaafsma et al., 2015), but is also well known for its crucial contribution to moral judgments (Young et al., 2007) and moral decisions involving tradeoffs between self-interest and other's welfare (Morishima et al., 2012; Tusche et al., 2016). Importantly, prior fMRI studies have also shown an atypical rTPJ activation in ASD cohorts compared with healthy control subjects (HCs) in a variety of social tasks that critically depend on ToM ability, such as processing naturalistic social situations (Pantelis et al., 2015), perceiving biological motion (Kana et al., 2009), or mentalizing about someone else (Lombardo et al., 2011). More relevantly, rTPJ in ASD individuals did not display reliable neural patterns that distinguish intentional harm from accidental harm (Koster-Hale et al., 2013). While these studies provide direct evidence of a ToM-related dysfunction of rTPJ in ASD individuals, the specific rTPJ dysfunction that drives atypical moral behaviors in ASD individuals remains largely unknown.

To address this question, we used a novel paradigm in an fMRI study where high-functioning ASD participants and HCs decided whether to accept or reject a series of offers. In particular, we independently manipulated two factors [i.e., Audience (whether decisions were made in public or private) and Moral Context (whether the offer involves a tradeoff between a personal financial loss and a charity donation, or between a personal financial gain and a donation to a morally bad cause)]. Moreover, the payoffs for participants and the associations varied across different trials in an orthogonal manner.

Combining computational modeling (Crockett, 2016; Kononov et al., 2018) and multivariate-based representational similarity analyses (RSAs; Kriegeskorte et al., 2008), the present design allowed us to directly test two predictions about different aspects of atypical moral behaviors in ASD individuals and their critical association with rTPJ dysfunction. The first prediction concerned social reputation (Frith and Frith, 2011), namely, how individuals care about their self-image in other's eyes. Evidence has shown that while making prosocial decisions, ASD participants show difficulties in sustaining a social reputation, which requires mentalizing ability (Izuma et al., 2011). Thus, compared with the HC group, ASD participants would show less distinction between their moral decisions made in public and in private. This would be associated with a reduced rTPJ engagement of representing information concerning social reputation, in the presence or absence of an audience.

Our second hypothesis was inspired by studies of moral judgments that reveal autistic individuals tend to judge moral culpability more often in terms of consequences (Moran et al., 2011; Fadda et al., 2016; Salvano-Pardieu et al., 2016), and often overevaluate the negative moral consequences (Moran et al., 2011; Bellesi et al., 2018). Hence, it was possible that compared with HCs, ASD participants would display increased aversion to the consequences of an immoral action and therefore reject more offers that earn themselves morally tainted profits. We further explored whether such behavioral differences could be explained by a reduced rTPJ representation of information concerning moral contexts in ASD participants.

Materials and Methods

Participants

A total of 48 participants were recruited for the present fMRI experiment. Specifically, 20 individuals with ASD (4 females; mean age, 17.0 ± 3.0 years; age range, 14–24 years; 3 left handed) were recruited via those who attended psychiatric and pediatric neurology clinics as outpatients and fulfilled the inclusion criteria. Twenty-eight HCs (10 females; mean age, 18.9 ± 3.0 years; age range, 14–25 years; 1 left handed) were recruited from the local community via fliers. Diagnoses of ASD were performed by a clinical pediatric neurologist according to the Autism Diagnostic Interview-revised (ADI-R; Table 1, all clinical tests describing the two samples). There were no significant between-group differences in gender ($\chi^2(1) = 1.395, p = 0.238$) and IQ (total: $t_{(43)} = -1.795, p = 0.080$; verbal: $t_{(43)} = -1.379, p = 0.175$; execution: $t_{(43)} = -1.421, p = 0.162$), except that ASD participants were slightly younger than HC participants ($t_{(46)} = -2.121, p = 0.039$).

The study was performed at the Imaging Center of the University of Campinas and approved by the local ethics committee (<https://plataformabrasil.saude.gov.br>; reference #CAAE 02388012.5.0000.5404; approved ethical statement #1904090). All experimental protocols and procedures were conducted in accordance with the institutional review board guidelines for experimental testing and complied with the latest revision of the Declaration of Helsinki.

Experimental design and task

We adopted a 2×2 within-subject design with a novel paradigm (but see Obeso et al., 2018; Qu et al., 2019). Specifically, participants decided wheth-365.719.27.8tie reviedpatlap

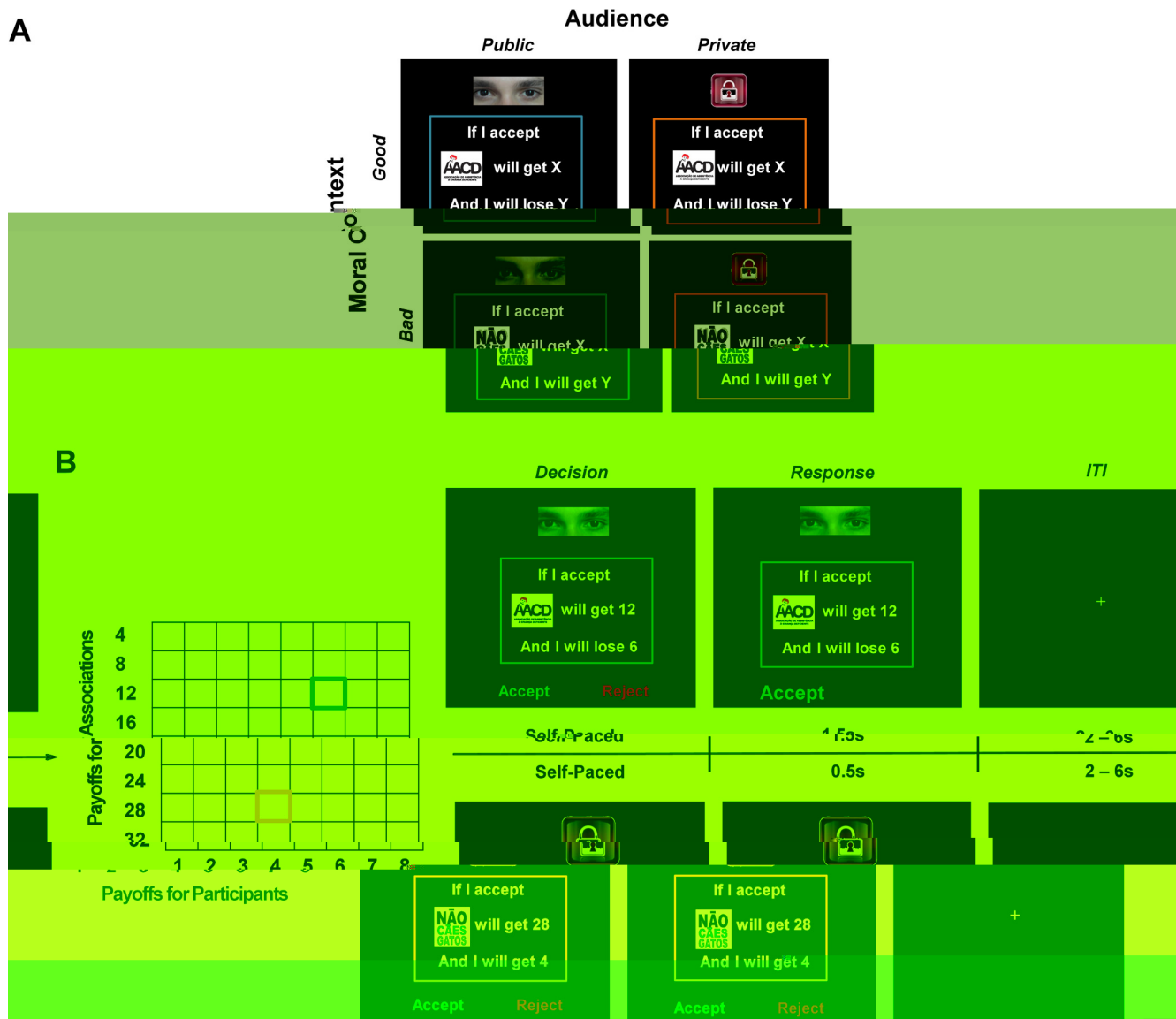


Figure 1. Illustration of experimental design and trial procedure. **A**, We used a 2×2 within-subject design by independently manipulating Audience (Private or Public) and Moral Context (Good or Bad), which yielded four experimental conditions (i.e., $Public_{Good}$, $Public_{Bad}$, $Private_{Good}$, and $Private_{Bad}$). The Public condition was indicated by the picture of “eyes,” and the Private condition was indicated by the picture of a “lock.” The Good Context involved a tradeoff between personal losses and benefits for a charity, whereas in the Bad Context participants traded personal benefits against benefits for a morally bad cause. **B**, Monetary payoffs (in Brazilian Real) for participants (8 levels: from 1 to 8, in steps of 1) and the association (8 levels: from 4 to 32, in steps of 4) were orthogonally varied, yielding 64 unique offers for each condition. In the example trial (one for the $Public_{Good}$ and the other for the $Private_{Bad}$ condition), participants were presented with an offer and decided whether to accept or reject the offer with no time limit. If they accepted the offer, both parties involved (i.e., the participant and the association) might undergo the financial consequences as proposed. If they rejected the offer, neither party would profit. In the Private condition, once a response was made, the screen was unchanged for 0.5 s to keep the chosen option private. In the Public condition, the chosen option was highlighted with a larger font and the nonchosen option disappeared, this lasted slightly longer (1.5 s) to further emphasize the presence of a witness. Each trial was ended with an intertrial interval (ITI) showing a jittered fixation (2.5 ~ 6.5 s).

could have real consequences. Thus, if participants accepted the offer in the Good context, they would lose a certain amount of money and the charity would be paid. If they do so in the Bad context, they would earn the money and the bad cause would also be paid. However, if participants rejected the offer, neither they nor the involved association would gain or lose any money (Fig. 1). Participants were also informed that all trials (decisions) were independent from each other so that the incentive consequences would not accumulate across the experiment. Only one trial would be randomly selected and paid at the end of the experiment.

One key aspect of the present design was that we varied the monetary stakes for the participants and the associations independently across trials within each condition. Personal payoffs (i.e., profits or losses) ranged from 1 to 8 in steps of 1 (unit, Brazilian Real; 1 Brazilian Real = ~0.2 US Dollars). Donations to both associations ranged from 4 to 32 in the steps of 4. The personal payoff and the donation were orthogonal, which led

to 64 different offers. Each offer appeared only once in each condition and thus summed up to 256 trials in total.

The functional scanning comprised four runs of 64 trials. Each moral context was assigned to either the first or the second of two runs. Each run consisted of two blocks, which included 32 trials presenting unique offers in either the Private or the Public condition. The order of runs involving Good/Bad moral contexts and Public/Private blocks were counterbalanced across participants. The trial order was randomized within each block. For each trial, participants were presented with the decision screen consisting of the payoff information for the participant (monetary gain or loss), and the association indicated by the corresponding symbol. The cue that signaled whether it was a Public (a picture of eyes) or a Private (i.e., a picture of a padlock) trial was also shown on the same screen. Here a cue of being watched was used as previous studies have consistently shown that it

axially oriented along the anterior commissure–posterior commissure plane and acquired in an ascending order. A high-resolution structural T1-weighted image was also collected for every participant using a 3D MRI sequence (TR = 7 ms; TE = 3.2 ms; flip angle = 8°; slice thickness = 1 mm; matrix = 240×240 ; FOV = $240 \times 240 \text{ mm}^2$).

Statistical analyses

One ASD participant was excluded from behavioral analyses because of the invariant response pattern (i.e., rejecting all trials in the task). After checking the preprocessed fMRI data, we excluded two more HC participants [one because half of the scanning data was lost because of a technical rea-

(dummy variable; reference level; same below), Audience (dummy variable; reference level: private; same below), Moral Context (dummy variable; reference level: good; same below), and their interactions (i.e., three two-way interactions and one three-way interaction) as the fixed-effect predictors. We also incorporated age as a covariate in the analyses to rule out its possible confounding effect. We included random-effects predictors that allowed varying intercepts across participants. For the statistical inference on each predictor, we performed the type II Wald χ^2 test on the model fits by using the Anova function in “car” package (Fox et al., 2016). Once the interactions were detected, we ran *post hoc* regressions on the subset of data given the different groups and then conditions. We reported the odds ratio (OR) as an index of effect size of each predictor on moral choices.

We also performed mixed-effects linear regression analyses on the log-transformed decision time (Anderson-Darling normality test: $A = 431.33$, $p < 0.001$) with the lmer function in the lme4 package, with the same fixed-effects predictors, random-effects predictors, and covariates as for the choice analyses. In addition, we also controlled the effect of specific decision (dummy variable; reference level: moral choice) in the regression model. We followed the procedure recommended by Luke (2017) to obtain the statistics for each predictor by applying the Satterthwaite approximations on the restricted maximum likelihood model (REML) fit via the “lmerTest” package (Luke, 2017). In addition, we reported the standardized coefficient (b_2) as an index of the effect size of each predictor on decision time together with other continuous dependent measures (e.g., rating, parameter estimates) using “EMAtools” (<https://cran.r-project.org/web/packages/EMAtools/>) and the “lm.beta” package (<https://cran.r-project.org/web/packages/lm.beta/>) for mixed effects and simple linear regression models, respectively.

Computational modeling

To examine how participants evaluated payoffs of each party and integrated them into a subjective value (SV), we compared the following eight models with different utility functions characterizing participants’ choices.

Model 1 was adapted from a recent study on moral decision-making by Crockett et al. (2014, 2017), which could be formally represented as follows:

$$SV(M_s, M_o) = \begin{cases} -(\alpha - q * \theta) * M_s + (1 - \alpha + q * \theta) * M_o & \text{if Good} \\ (\alpha - q * \theta) * M_s - (1 - \alpha + q * \theta) * M_o & \text{if Bad,} \end{cases}$$

where SV denotes the SV of the given trial if the participant chooses to accept. For rejection trials, SV is always 0 given the rule of the task (i.e., neither beneficiaries would gain the money; same for all models). M_s and M_o represent the payoff (gain or loss) for oneself and payoffs donated to the corresponding association. α ($0 < \alpha < 1$) is the unknown parameter of social preference that arbitrates the relative weight on the payoff for the participant in the decision. θ ($0 < \theta < 1$) is the unknown

six movement parameters were added to this GLM as covariates to account for artifacts of head motion. The canonical hemodynamic response function was used and a high-pass temporal filtering was performed with a default cutoff value of 128 s to remove low-

fMRI data preprocessing. Functional imaging data were analyzed using SPM12 (Wellcome Trust Center for Neuroimaging, University College London). The preprocessing procedure followed the pipeline recommended by SPM12. In particular, functional images (EPI) were first realigned to the first volume to correct motion artifacts, unwarped, and corrected for slice timing. Next, the structural T1 image was segmented into white matter, gray matter, and CSF with the skull removed, and coregistered to the mean functional images. Then all functional images were normalized to the Montreal Neurologic Institute (MNI) space, resampled with a $2 \times 2 \times 2$ mm³ resolution, in terms of parameters generated in the previous step. Last, the normalized functional images were smoothed using an 8 mm isotropic full-width at half-maximum based on a Gaussian kernel.

Within-subject representational similarity analyses. To clarify what information rTPJ exactly represents during the decision period that distinguished ASD participants from HC participants, we conducted a within-subject RSA in Python 3.6.8 using the *nltools* package (version 0.3.14; <https://github.com/cosanlab/nltools>). Some preparation was performed before implementing RSA. In particular, we established a trial-wise general linear model (GLM) for each participant, which included the onsets of the decision screen with the duration of decision time of each valid trial. Here, valid trials were those that conformed to neither the exclusion criterion for the behavioral data (trials with extremely fast or slow responses; see above for details) nor the fMRI data (trials in runs with excessive head motion). The onsets of button press and invalid trials were also modeled as separate regressors of no interest. In addition,

Supplementary univariate analyses

We also performed a traditional univariate GLM analysis to examine whether the mean neural activations were modulated by different conditions and how neural signals in ASD participants differed from those in healthy control subjects, focusing on the rTPJ. At the individual level, we incorporated the onsets of the decision phase of all conditions (i.e., Private_{Good}, Private_{Bad}, Public_{Good}, Public_{Bad}) in valid trials as regressors of interest. Similarly, the onsets of button press together with invalid trials as well as head motion parameters were also modeled as separate regressors of no interest. After the parameter estimation, we constructed the following contrasts concerning the main effect of Audience (i.e., Public-Private) and Moral Context (i.e., Good-Bad). These contrast images were

Table 3. Results of mixed-effect logistic regressions predicting log-transformed decision time (in ms)

	All ^b (SE)	Good ^b (SE)	Good: private ^b (SE)	Good: public ^b (SE)	Bad ^b (SE)
Intercept	7.22*** (0.07)	7.19*** (0.06)	7.19*** (0.06)	7.18*** (0.06)	7.19*** (0.10)
Group	0.07 (0.11)	0.04 (0.09)	0.04 (0.10)	0.09 (0.09)	−0.04 (0.14)
Audience	−0.01 (0.02)	−0.02 (0.01)			0.01 (0.01)
Moral context	−0.08*** (0.02)				
Group × Audience	0.04 (0.02)	0.04* (0.02)			−0.04† (0.02)
Group × Moral context	−0.13*** (0.02)				
Audience × Moral context	0.02 (0.02)				
Group × Audience × Moral context	−0.08* (0.03)				
Decision	−0.02† (0.01)	0.04* (0.01)	0.05* (0.02)	0.03 (0.02)	−0.10*** (0.02)
Payoff for oneself ^{a,b}		0.03*** (0.01)	0.02* (0.01)	0.04*** (0.01)	0.02*** (0.01)
Payoff for association ^{a,b}		−0.03*** (0.01)	−0.03*** (0.01)	−0.03*** (0.01)	−0.01* (0.006)
Age ^a	0.03 (0.05)	−0.001 (0.05)	−0.01 (0.05)	0.005 (0.05)	0.05 (0.07)
AIC	15,114.8	6095.5	3074.3	3058.6	7203.8
BIC	15,203.4	6162.4	3122.2	3106.5	7270.6
<i>N</i> (observation)	11,823	5912	2952	2960	5911
<i>N</i> (participant)	47	47	47	47	47

Values are the mean (SE). Reference levels were set as follows: Group, NC; Audience, private; Association, good cause (charity). Table also shows goodness-of-fit statistics: BIC, Bayesian information criterion.

† $p < 0.06$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

^aWe standardized these variables for the analyses.

^bThese variables were added as covariates only when the regressor Association (and its interaction) was not in the regression model, as the regressor “payoff for oneself” qualitatively covaried with Association, which might cause the collinear issue.

ASD participants do not appear to consider social reputation and rigorously conform to a rule in curbing their immoral behaviors

Mixed-effect logistic regressions revealed that participants were more likely to behave morally in the Bad Context than in the Good Context (i.e., rejecting more frequently the offer in the Bad Context than accepting it in the Good Context; a main effect of Moral Context: $\chi^2(1) = 632.68$, $p < 0.001$). More importantly, significant interaction effects were identified between Group and Audience ($\chi^2(1) = 4.50$, $p = 0.034$) as well as between Group and Moral Context ($\chi^2(1) = 59.33$, $p < 0.001$) on choosing the moral option (i.e., accepting the offer to benefit a charity or rejecting the offer to benefit a morally bad cause; Fig. 2). No other main effect (p values > 0.09) or interaction effect was detected (p values > 0.57).

To understand the first interaction effect, we performed *post hoc* analyses on the dataset of the ASD and the HC groups, respectively. For each analysis, we ran a similar logistic regression, including the main effect of audience and context as the fixed-effects predictors. The Audience × Moral Context interaction was dropped from these analyses as neither this effect ($\chi^2(1) = 0.31$, $p = 0.580$) nor the three-way interaction effect ($\chi^2(1) = 0.30$, $p = 0.586$) was significant in the main analysis. The results showed that while healthy control subjects were more likely to make the moral choice when they were observed in the Public condition (vs Private; OR = 1.16, $b = 0.15$, SE = 0.06, $p = 0.012$), ASD participants did not change their behaviors significantly depending on the presence or absence of a witness (OR = 0.93, $b = -0.08$, SE = 0.08, $p = 0.371$).

To understand the second interaction effect, we performed similar regression analyses using trials in the Good and Bad Contexts separately. For each *post hoc* regression analysis, we incorporated Group and Audience, along with their interaction as the fixed-effects predictors, while controlling for the effect of the payoff for participants and associations in these analyses (same below for analyses on decision time). We observed only a strong main effect of Group ($\chi^2(1) = 5.05$, $p = 0.025$) and a Group × Audience interaction effect in the Bad Context ($\chi^2(1) = 4.04$, $p = 0.044$), which was mainly driven by a drastically enhanced probability of behaving morally in the ASD group (vs

HC group) when deciding privately (OR = 64.25, $b = 4.16$, SE = 1.53, $p = 0.006$). Neither of these effects was significant in the Good Context (p values > 0.12 ; Table 2, details of regression outputs).

ASD participants evaluate the immoral gains more severely for both themselves and the bad cause

We developed eight models with different utility functions characterizing participants' choices in the ASD and HC groups separately. Model estimation and comparison was performed with an HBA approach (Gelman et al., 2014) via the “hBayesDM” package (see Materials and Methods for details). R-hat values of all estimated parameters of all models are close to 1.0 (i.e., < 1.06 in the worst case), which showed sufficient convergence of the MCMC chains (Gelman and Rubin, 1992). Hierarchical Bayesian model comparison showed that model 4 (see below for the utility function) has the lowest LOOIC scores, indicating that it fits to the current dataset the best compared with other competitive models (Fig. 3A), as follows:

$$SV(M_S, M_O) = \begin{cases} -(\alpha_{\text{Good}} - q * \theta) * M_S + (\beta_{\text{Good}} + q * \theta) * M_O & \text{if Good} \\ (\alpha_{\text{Bad}} - q * \theta) * M_S + (\beta_{\text{Bad}} - q * \theta) * M_O & \text{if Bad} \end{cases}$$

Here, SV denotes the subjective value of the given trial depending on the specific choice made by the participant. M_S and M_O represent the payoff (gain or loss) for oneself and each association respectively. Established on the basis of a donation task (Lopez-Persem et al., 2017), the winning model assumed that people weighed their own payoff (measured by α : α_{good} , α_{bad}) and the benefits for associations (measured by β : β_{good} , β_{bad}) separately given the moral contexts involved in the decisions. θ measured the audience effect, which was modulated by an indicator function, q (0 for private, 1 for public; see Materials and Methods for details). A posterior predictive check further confirmed that the simulated choice behaviors in light of the parameter estimates of the winning model can nicely capture the actual behaviors by showing a high correlation between each

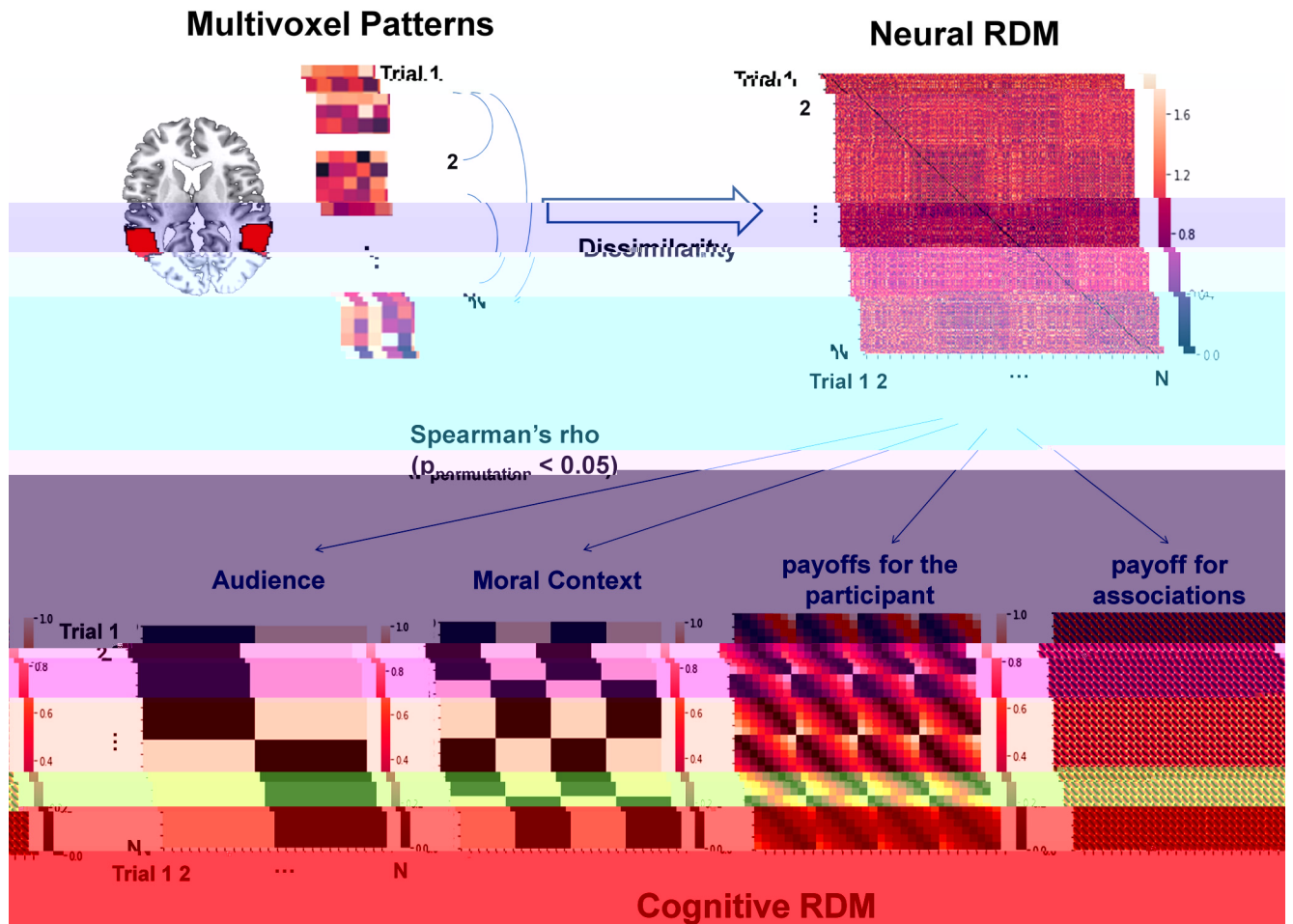


Figure 6. Illustration of within-subject RSAs. For each individual, we first constructed a nen830.9(e-2172f.641/F15)

cea 490(chdedDocu)0(t)s216.acalg0.9(e-2)23.3(jeext)TJ3118(ng)-0(y)-346.66(

other (i.e., for both the HC group and ASD group: Pearson's r values > 0.99 , p values < 0.001 ; Fig. 3B).

Next, we examined how parameters derived from the winning model vary in terms of groups and experimental conditions. To this end, we extracted the individual-level posterior mean of key parameters (i.e., α , β , and θ) and performed linear regression, including Group as the predictor on each of them, respectively. To test the Group \times Association interaction on α and β , we regressed groups on the difference score between two contexts for each of the parameters. For all these regression analyses, we also added age as a covariate to control for its confounding effect.

We first showed a significant Group \times Association interaction on both α ($b = 7.93$, $SE = 3.91$, $t_{(44)} = 2.03$, $p = 0.049$; $b_z = 0.30$) and β ($b = -10.88$, $SE = 3.46$, $t_{(44)} = -3.14$, $p = 0.003$; $b_z = -0.43$). Simple-effect analyses showed a significant decrease of decision weights on payoffs, in ASD participants, for both themselves ($\alpha_{\text{HC}} \text{ vs } \alpha_{\text{ASD}}$: 0.90 ± 11.74 vs -9.27 ± 10.13 , $t_{(44)} = -3.14$, $p = 0.003$; $b_z = -0.45$) and the morally bad cause ($\beta_{\text{HC}} \text{ vs } \beta_{\text{ASD}}$: -4.87 ± 5.02 vs -11.52 ± 8.67 , $t_{(44)} = -2.96$, $p = 0.005$; $b_z = -0.41$). No between-group difference was observed in either parameter when participants weighed the tradeoff between personal financial losses ($\alpha_{\text{HC}} \text{ vs } \alpha_{\text{ASD}}$: 11.18 ± 6.41 vs 8.89 ± 9.09 , $t_{(44)} = -1.26$, $p = 0.216$; $b_z = -0.19$) and the donation to a charity

($\beta_{\text{HC}} \text{ vs } \beta_{\text{ASD}}$: 4.38 ± 4.04 vs 6.78 ± 7.83 , $t_{(44)} = 1.56$, $p = 0.126$; $b_z = 0.24$; Fig. 4A). Notably, the correlation between α and β was not significant across moral contexts in either group (ASD group: Good Context: $r = -0.177$, $p = 0.469$; Bad Context: $r = -0.242$, $p = 0.319$; HC group: Good Context: $r = -0.018$, $p = 0.928$; Bad Context: $r = -0.003$, $p = 0.989$; Fig. 4B). This indicates that participants value payoffs for oneself and the causes (associations) independently. Consistent with the behavioral finding, we also observed a trend to significance for the between-group difference in θ ; namely, that ASD participants exhibited a reduced audience effect compared with HC participants during moral decision-making ($\theta_{\text{HC}} \text{ vs } \theta_{\text{ASD}}$: 0.39 ± 0.67 vs 0.17 ± 0.12 , $t_{(44)} = -1.80$, $p = 0.080$, $b_z = -0.27$).

ASD participants do not differ from HCs in decision time in either moral context

Mixed-effect linear regression on log-transformed decision time showed a significant three-way interaction among Group, Audience, and Moral Context ($F_{(1,11,769)} = 6.02$, $p = 0.014$), along with a Group \times Moral Context interaction effect ($F_{(1,11,769)} = 100.20$, $p < 0.001$) and a main effect of Moral Context ($F_{(1,11,772)} = 299.76$, $p < 0.001$) after controlling for the effect of specific choices ($F_{(1,11,804)} = 3.76$, $p = 0.052$; Fig. 5). Splitting the

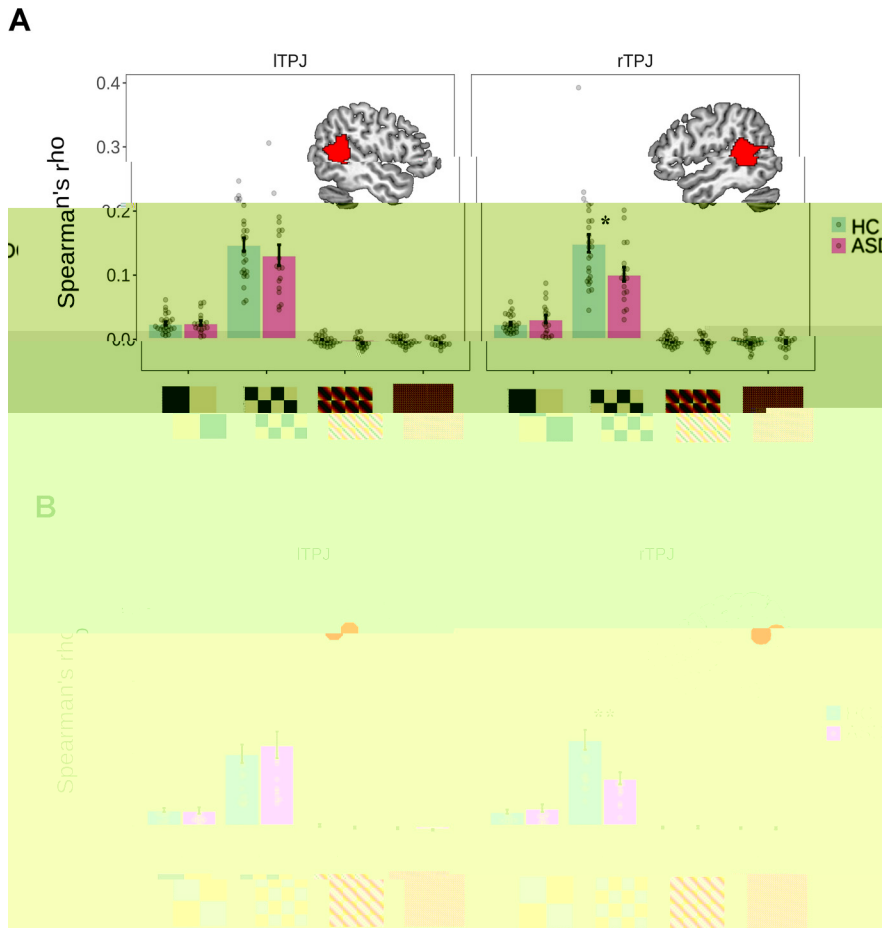


Figure 7. *A, B*, Within-subject RSA results using the parcellation-based ROI (*A*) and the coordinate-based ROI (*B*) of TPJ. For each participant, we only adopted valid trials (see Materials and Methods for details) in these analyses. Each dot represents the data of a single participant. Error bars represent the SEM; significance: * $p_{\text{permutation}} < 0.05$, ** $p_{\text{permutation}} < 0.01$, after controlling for the age difference.

dataset according to Moral Context, *post hoc* analyses revealed a significant Group \times Audience interaction effect when participants decided whether to serve a good cause at a personal cost ($F_{(1,5860)} = 4.28$, $p = 0.039$) and a trend-to-significant interaction effect in the Bad Context ($F_{(1,5859)} = 3.76$, $p = 0.053$). However, neither the main effect of Group in the Good Context (ASD: 1676.5 ± 527.7 ms; HC: 1490.0 ± 399.5 ms; $F_{(1,44)} = 0.51$, $p = 0.479$) nor in the Bad Context (ASD: 1525.7 ± 828.1 ms; HC: 1445.5 ± 500.9 ms; $F_{(1,44)} = 0.17$, $p = 0.682$) was significant. The interaction effect in the Good Context was driven by a slightly larger difference in decision time between groups when they made decisions in public (ASD: 1709.5 ± 558.8 ms; HC: 1467.9 ± 379.3 ms) compared with those made in private (ASD: 1645.5 ± 526.0 ms; HC: 1514.1 ± 448.9 ms). However, neither of these between-group differences was statistically significant (public: $b = 0.09$, $SE = 0.09$, $t_{(44)} = 0.98$, $p = 0.334$, $b_z = 0.29$; private: $b = 0.04$, $SE = 0.10$, $t_{(44)} = 0.42$, $p = 0.678$, $b_z = 0.13$; Table 3, details of regression output).

Imaging results

Decreased neural representation of moral contexts in the rTPJ of ASD participants

To examine how the decision-related neural patterns differ in representing information contributing to the value computation and final decisions between ASD participants and HC

participants, we performed a within-subject RSA (Fig. 6, illustration of RSA procedure). Given our hypotheses, we focused our analysis on the rTPJ. To avoid bias on results caused by ROI selection to the maximum degree, we defined the rTPJ in two different ways, either via a whole-brain parcellation based on meta-analytic functional coactivation of the Neurosynth database (i.e., the parcellation-based ROI) or via a coordinate-based manner given a recent meta-analysis on neural correlates of ToM (Schurz et al., 2014; i.e., the coordinate-based ROI; for details, see Materials and Methods).

Regardless of the ROI approach, we consistently found that, compared with the HC group, ASD participants only showed a reduced representation of the information of the identity of associations in the rTPJ (ASD vs HC: the parcellation-based ROI: Spearman's $\rho = 0.101 \pm 0.047$ vs 0.150 ± 0.071 ; $p_{\text{permutation}} = 0.013$; the coordinate-based ROI: 0.066 ± 0.036 vs 0.119 ± 0.070 ; $p_{\text{permutation}} = 0.006$). These significant differences held after ruling out the confounding effect of age. Importantly, such a between-group difference of similarity was not observed between the neural RDM in the rTPJ and other cognitive RDMs (the parcellation-based ROI: $p_{\text{permutation}}$ values > 0.20 ; the coordinate-based ROI: $p_{\text{permutation}}$ values > 0.38) or between the neural RDM in the ITPJ and all the cognitive RDMs (the

parcellation-based ROI: $p_{\text{permutation}}$ values > 0.17 ; the coordinate-based ROI: $p_{\text{permutation}}$ values > 0.30 ; Fig. 7, Table 4, details). *Post hoc* 2 (group) \times 4 (cognitive RDM) mixed ANOVA on the Fisher r -to- z transformed Spearman's ρ revealed a strong interaction between group and cognitive RDM only in rTPJ (the parcellation-based ROI: $F_{(3,126)} = 6.09$, $p < 0.001$; the coordinate-based ROI: $F_{(3,126)} = 8.37$, $p < 0.001$) but not in ITPJ (the parcellation-based ROI: $F_{(3,126)} = 0.65$, $p = 0.585$; the coordinate-based ROI: $F_{(3,126)} = 0.42$, $p = 0.743$) after controlling for the age difference, which further confirmed that the reduced ability to represent the information of moral context in ASD participants was uniquely reflected in rTPJ. Finally, to further examine the robustness of the above findings, we also applied the above analyses using all 256 trials, which did not affect the results (Fig. 8, Table 5, details).

Univariate results in rTPJ

We first investigated whether the neural audience effect in rTPJ (i.e., Public $>$ Private) in healthy control subjects reported in the study by Qu et al. (2019) could be replicated in the present study. The results showed that the rTPJ activity was not significantly higher in the Public (vs Private) condition (no voxel survived under a threshold of $p < 0.005$ uncorrected at the voxel level

Table 4. Within-subject RSA results in TPJ using valid trials

			Spearman's ρ (mean \pm SD)		$p_{\text{permutation}}$	$p_{\text{permutation}}^b$
			ASD	HC		
Neurosynth	ITPJ	Audience	0.026 \pm 0.016 ^{***}	0.025 \pm 0.014 ^{***}	0.848	0.493
		Moral context	0.131 \pm 0.069 ^{***}	0.148 \pm 0.054 ^{***}	0.403	0.493
		Payoff for oneself	−0.005 \pm 0.008	−0.002 \pm 0.007	0.174	0.271
	rTPJ	Payoff for association	−0.004 \pm 0.006	−0.003 \pm 0.006	0.413	0.447
		Audience	0.032 \pm 0.025 ^{***}	0.024 \pm 0.013 ^{***}	0.201	0.163
		Moral context	0.101 \pm 0.047 ^{***}	0.150 \pm 0.071 ^{***}	0.013	0.018
Meta-analysis ^a	ITPJ	Payoff for oneself	−0.004 \pm 0.009	−0.003 \pm 0.007	0.723	0.311
		Payoff for association	−0.004 \pm 0.010	−0.006 \pm 0.009	0.578	0.995
		Audience	0.021 \pm 0.019 ^{***}	0.022 \pm 0.014 ^{***}	0.912	0.931
	rTPJ	Moral context	0.112 \pm 0.079 ^{***}	0.100 \pm 0.065 ^{***}	0.566	0.551
		Payoff for oneself	−0.002 \pm 0.007	0.0005 \pm 0.009	0.304	0.472
		Payoff for association	−0.005 \pm 0.005	−0.003 \pm 0.007	0.308	0.262
	ITPJ	Audience	0.025 \pm 0.022 ^{***}	0.020 \pm 0.014 ^{***}	0.383	0.230
		Moral context	0.066 \pm 0.036 ^{***}	0.119 \pm 0.070 ^{***}	0.006	0.002
		Payoff for oneself	−0.002 \pm 0.008	−0.002 \pm 0.007	0.900	0.685
	rTPJ	Payoff for association	−0.003 \pm 0.007	−0.003 \pm 0.006	0.924	0.400

We excluded trials that did not reach the behavioral criterion (i.e., those with a decision time of <200 ms or longer than the mean \pm 3 SDs of that individual) or fMRI criterion (all trials in a run with an excessive head motion: ASD, >5 mm; HC, >3 mm). l, Left; r, right.

^a These masks were spheres with a radius of 10 mm centering on the MNI coordinates based on a recent meta-analysis involving the mentalizing process (peak MNI coordinates: left TPJ/pSTS: $-53/-59/20$; right TPJ/pSTS: $56/-56/18$).

^b We added the standardized age as the covariates to the regression, using the ImPerm package.

*** These effects are significantly higher than 0 (i.e., one-sample t test with 5000 permutations; $p_{\text{permutation}} < 0.001$).

with $k=10$, in either rTPJ mask; Fig. 9A). One possibility could be that the neural audience effect of rTPJ was modulated by large individual differences in the behavioral audience effect across individuals, which blurred the main effect. To test this possibility, we extracted the mean activity (contrast value) of the rTPJ from each condition, and then computed a neural index of audience effect for each individual (i.e., $0.5 * [(Public_{\text{Good}} + Public_{\text{Bad}}) - (Private_{\text{Good}} + Private_{\text{Bad}})]$). We also defined a behavioral index of audience effect on the proportion of moral choice, which was calculated with the same equation. Results showed that the Pearson correlation between these two indices was not significant (the parcellation-based ROI: $r(24) = 0.02$, $p = 0.914$; the coordinate-based ROI: $r(24) = -0.06$, $p = 0.761$; Fig. 9B). Furthermore, the between-group comparison did not reveal a significant result in the audience effect in rTPJ (i.e., no voxel survived under the threshold mentioned above; Fig. 10). Besides, no significant difference in the neural activity was observed in the rTPJ between the Good and Bad Contexts in the HC group or between two groups (i.e., no voxel survived under the threshold mentioned above). For the completeness of the analyses, we also applied the same analyses to ITPJ, yielding similar results (Figs. 9, 10, Table 6, whole-brain results under a liberal threshold).

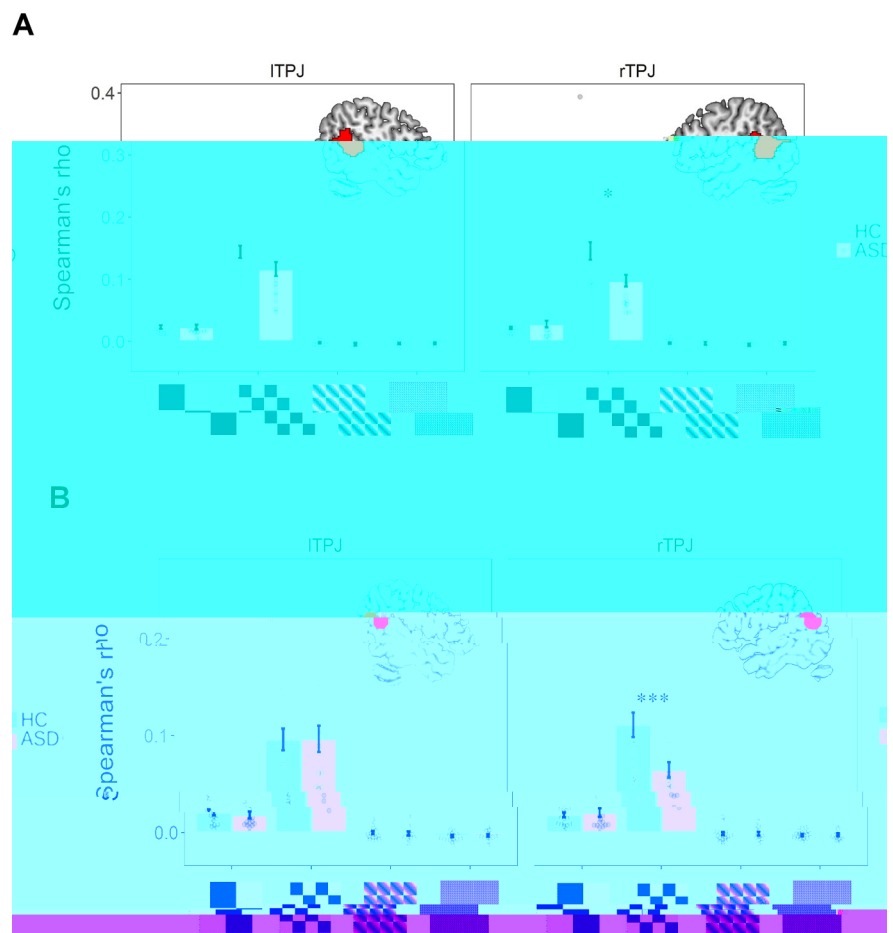


Figure 8. *A, B*, Robustness check of within-subject RSA results using the parcellation-based ROI (*A*) and the coordinate-based ROI (*B*) of TPJ. For each participant, we adopted all 256 trials in these analyses. Each dot represents the data of a single participant. Error bars represent the SEM; significance: $*p_{\text{permutation}} < 0.05$, $***p_{\text{permutation}} < 0.001$, after controlling for the age difference.

Table 5. Within-subject RSA results in TPJ using all 256 trials

			Spearman's ρ (mean \pm SD)		$p_{\text{permutation}}$	$p_{\text{permutation}}^b$
			ASD	HC		
Neurosynth	ITPJ	Audience	0.023 \pm 0.016 ^{***}	0.023 \pm 0.014 ^{***}	0.967	0.538
		Moral context	0.117 \pm 0.047 ^{***}	0.144 \pm 0.050 ^{***}	0.063	0.094
		Payoff for oneself	−0.004 \pm 0.009	−0.002 \pm 0.006	0.299	0.358
		Payoff for association	−0.003 \pm 0.006	−0.003 \pm 0.006	0.932	0.919
	rTPJ	Audience	0.028 \pm 0.022 ^{***}	0.022 \pm 0.010 ^{***}	0.203	0.169
		Moral context	0.098 \pm 0.040 ^{***}	0.146 \pm 0.070 ^{***}	0.009	0.027
		Payoff for oneself	−0.003 \pm 0.009	−0.003 \pm 0.007	0.854	0.387
		Payoff for association	−0.003 \pm 0.009	−0.005 \pm 0.008	0.386	0.667
Meta-analysis ^a	ITPJ	Audience	0.018 \pm 0.015 ^{***}	0.021 \pm 0.016 ^{***}	0.579	0.850
		Moral context	0.097 \pm 0.058 ^{***}	0.096 \pm 0.058 ^{***}	0.972	0.914
		Payoff for oneself	−0.001 \pm 0.008	−0.0004 \pm 0.008	0.721	0.745
		Payoff for association	−0.003 \pm 0.005	−0.004 \pm 0.007	0.775	0.871
	rTPJ	Audience	0.021 \pm 0.019 ^{***}	0.018 \pm 0.013 ^{***}	0.613	0.451
		Moral context	0.067 \pm 0.034 ^{***}	0.111 \pm 0.064 ^{***}	0.006	< 0.001
		Payoff for oneself	−0.001 \pm 0.008	−0.001 \pm 0.008	0.990	0.528
		Payoff for association	−0.002 \pm 0.006	−0.003 \pm 0.006	0.796	0.689

l, left; r, right. *Post hoc* 2 (group) \times 4 (cognitive RDM) mixed ANOVA on the Fisher *r*-to-*z* transformed Spearman's ρ revealed a strong interaction between group and cognitive RDM only in rTPJ regardless of the way we defined the ROI (the parcellation-based ROI: $F_{(3,126)} = 6.59$, $p < 0.001$; the coordinate-based ROI: $F_{(3,126)} = 7.37$, $p < 0.001$), which was not true in ITPJ (the parcellation-based ROI: $F_{(3,126)} = 3.00$, $p = 0.033$; the coordinate-based ROI: $F_{(3,126)} = 0.03$, $p = 0.994$) after controlling for the age difference, which further confirmed that the specific between-group effect in representing information of Moral context was unique in rTPJ.

^a These masks were spheres with a radius of 10 mm centering on the MNI coordinates based on a recent meta-analysis involving the mentalizing process (peak MNI coordinates: left TPJ/pSTS: $-53/-59/20$; right TPJ/pSTS: $56/-56/18$).

^b We added the standardized age as the covariates to the regression, using the ImPerm package.

^{***} These effects are significantly higher than 0 (i.e., one-sample *t* test with 5000 permutations; $p_{\text{permutation}} < 0.001$).

Discussion

When facing moral dilemmas such as earning ill gotten money by supporting a bad cause or donating to a charity at a personal cost, how do autistic individuals choose? Do they vary their immoral/moral behaviors with respect to the presence or absence of someone else or contingent on moral concerns elicited by specific contexts (i.e., serving a good or a bad cause)? What neuro-computational mechanisms underlie such behavioral changes? In the present model-based fMRI study, we attempted to answer these questions by adopting a novel task in which individuals decided among tradeoffs between personal benefits/losses and context-sensitive moral concerns while also, perhaps, considering their social reputation. Our behavioral results reveal that the moral behavior of ASD individuals differs from healthy control subjects in two aspects.

First, ASD individuals, unlike healthy control subjects, blurred the distinction between private and public conditions while making moral decisions. This finding not only coheres with the ToM deficit hypothesis of ASD individuals (Baron-Cohen et al., 1985; Baron-Cohen, 2001), but also agrees with previous findings using a tradeoff between suffering personal losses and donating to a good cause (Izuma et al., 2011). Moreover, it extends the lack of attention to social reputation in autism to include an immoral context where individuals are confronted with a moral conflict between personal profits and a cost brought by benefiting an immoral cause. This first finding confirms that ASD individuals do not appear to take into account their social reputation while making immoral/moral choices consistently across contexts (Izuma et al., 2011).

Second, a robust behavioral difference between ASD individuals and healthy control subjects was found specifically in one moral context. ASD individuals generally refused more offers in the Bad Context that could have earned extra money for themselves but resulted in an immoral consequence. No similar between-group difference was observed in the Good Context. Note that decision difficulty cannot explain these behavioral effects because no decision time difference was observed between the two groups. Furthermore, this effect cannot be attributed to

their greater dislike/like for the morally bad cause because there was no significant between-group difference on subjective ratings.

Our computational modeling approach provides crucial insights to understand further this difference in ASD individuals, which is specific to moral behaviors serving a bad cause. In parallel to the choice findings, ASD individuals drastically lowered their decision weights on payoffs that would be earned both for themselves and the morally bad cause, whereas they valued the personal losses and the benefits of the charity similarly to healthy control subjects. These findings strongly indicate an atypical valuation of morally tainted personal profits and moral costs brought by benefiting a bad cause in autistic individuals. This probably led to their extremely high rejection rate for immoral offers. Our results fit the literature on moral judgment, which has shown that ASD individuals exhibit an excessive valuation of negative consequences when judging the moral appropriateness or permissibility of actions. For example, Moran et al. (2011) reported that ASD participants considered accidental negative outcomes less permissible than healthy control subjects, whereas both groups rated other types of events as having similar moral appropriateness. In a more recent study, a similar effect was observed; namely, ASD individuals judged a protagonist's immoral but understandable action (e.g., a husband stealing medicine sold at an unaffordable price to save his fatally sick wife) as less morally acceptable than did healthy control subjects (Schaller et al., 2019). In agreement with these findings, our results suggest that autistic individuals may apply a rule of refusing to serve an immoral cause because they evaluate the negative consequences of their actions more severely. This might result in insensitivity in ASD individuals who have difficulty in adjusting their behaviors regarding their personal interests that might be associated with immoral consequences.

Another possible explanatory factor of ASD participants' tendency to make more moral decisions in the Bad Context is behavioral rigidity, a core symptom for clinical diagnosis of ASD (American Psychiatric Association, 2013). Previous studies have revealed that, compared with healthy control subjects, individuals with ASD were more likely to show repetitive behaviors in a

variety of cognitive tasks (D’Cruz et al., 2013; Watanabe et al., 2019). Hence, it is possible that behavioral rigidity, at least to some extent, is a more general mechanism that contributes to the inflexibly moral behaviors in the Bad Context (i.e., rejecting >85% of the trials). Nonetheless, this explanation should be treated with caution because it seems not to account well for the behaviors of ASD participants in the Good Context, where they behaved in a comparatively more flexible fashion (i.e., accepting ~60% of the trials).

At the brain level, we performed within-subject RSA to examine how different types of information (social reputation, moral contexts, payoffs for each party) that contribute to the final decision were represented in the rTPJ, and how distinct rTPJ representations distinguish ASD participants from healthy control subjects. Compared with the traditional univariate approach, RSA takes advantage of neural patterns from multiple voxels and proves to be more sensitive to subtle experimental effects that might be masked by the averaged local neural responses (Norman et al., 2006; Hebart and Baker, 2018). RSA is also considered to be more informative, because it takes into account the variability within multivoxel patterns (Kriegeskorte et al., 2008; Popal et al., 2019). We observed a reduced association (representation similarity) in ASD participants (vs healthy control subjects) between the trial-by-trial multivariate rTPJ patterns and the information structure unique to the moral contexts, despite that, such a representation in rTPJ is present in both groups. The representations of other types of information (i.e., social reputation and payoffs for each party) did not differ between groups. Together with a much higher rejection rate, as well as atypical weights on payoffs in the bad context, this RSA finding provides a neural account for previous findings that autistic individuals are inclined to judge moral culpability more severely than HCs on the basis of its consequences. This distinguishes ASD individuals from HCs, who prioritize intentions to guide their moral judgments (Fadda et al., 2016; Salvano-Pardieu et al., 2016; Bellesi et al., 2018). Notably, our results showed that the group difference in representational similarity was only detected in rTPJ but not in ITPJ, further indicating a unique role of rTPJ in specifically representing information concerning moral contexts.

Regarding the function of the rTPJ, our RSA finding is consistent with a recent TMS study in healthy volunteers that revealed a context-sensitive moral role of rTPJ in signaling moral conflicts between personal benefits and moral values (Obeso et al., 2018). That study evidenced an asymmetrical TMS effect of rTPJ on moral behaviors depending on the moral context.

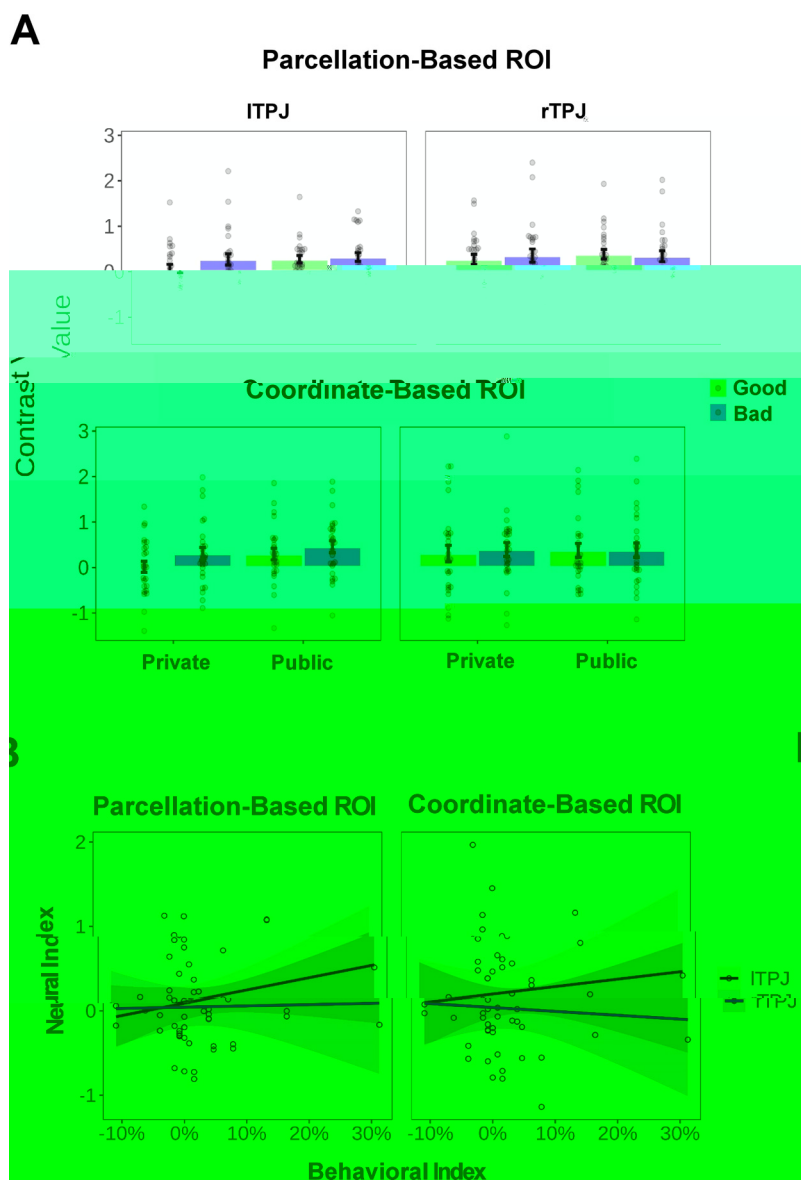


Figure 9. Univariate results of TPJ in healthy control subjects. **A**, Bar plot of TPJ signals. For visualization, we extracted the mean activity (contrast value) of ITPJ and rTPJ from the parcellation-based or coordinate-based mask as a function of reputation (Private or Public) and context (Good or Bad). Each dot represents the data of a single participant. Error bars represent the SEM. **B**, Relationship between neural audience effect in TPJ and behavioral audience effect across individuals. Each dot repre-

Specifically, healthy participants under rTPJ stimulation were more altruistic such that they accepted more offers of donating to a charity at a personal cost regardless of donation amounts, whereas rTPJ disruption inhibited participants from accepting offers to earn morally tainted money only when benefits to the bad cause were large. Building on this finding, the present study provides further evidence using a different approach to reveal that rTPJ is critically involved in representing the moral contexts that flexibly modulate the tradeoff between personal benefits and other’s welfare during decision-making, which extends our understanding of the rTPJ function.

Notably, our univariate fMRI results did not reveal a neural audience effect in rTPJ in the healthy control subjects as was initially expected. Although previous studies provided evidence (Izuma, 2012; Qu et al., 2019) suggesting that TPJ is involved in social reputation, negative evidence also exists. For instance, a

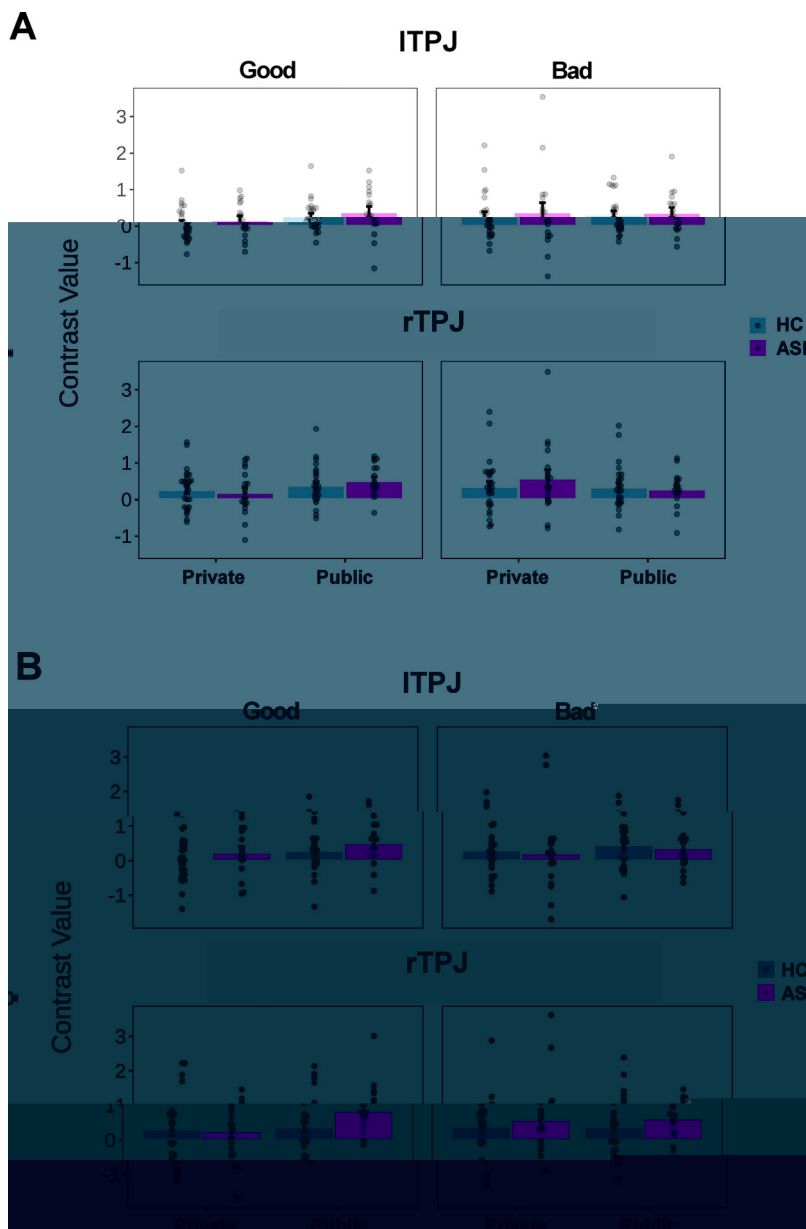


Figure 10. *A, B*, Univariate results of TPJ in the HC and ASD groups using the parcellation-based mask (*A*) and the coordinate-based mask (*B*). For visualization, we extracted the mean activity (contrast value) of ITPJ and rTPJ from the corresponding masks as a function of group (ASD or HC), reputation (Private or Public), and context (Good or Bad). Each dot represents the data of a single participant. Error bars represent the SEM.

recent transcranial magnetic stimulation (TMS) study using a similar experimental paradigm has shown that disrupting rTPJ (vs sham) does not influence the audience effect on moral decisions in healthy individuals (Obeso et al., 2018). In addition, two earlier fMRI studies failed to find an increased activation of rTPJ in response to the presence (vs absence) of observers while healthy participants made charitable decisions (Izuma et al., 2010b) or social evaluation (Izuma et al., 2010a). However, it is also worth noting that nonsignificant results do not necessarily reflect a true null effect (Makin and Xivry, 2019). Also, our RSA result suggests that multivoxel patterns of rTPJ represent the information of social reputation in healthy control subjects. Further studies are needed to clarify whether and how rTPJ plays a role in reputation-based decision-making.

Intriguingly, we did not observe a between-group difference of rTPJ in representing information about social reputation,

although, as expected, a small but significant effect of social reputation on moral behaviors was observed only in healthy control subjects rather than in ASD participants. At first glance, this finding may seem at odds with the well established role of the rTPJ in mentalizing (and relevant social abilities) in both healthy participants (Hampton et al., 2008; Young et al., 2010; Carter et al., 2012; Morishima et al., 2012; Schurz et al., 2014; Hutcherson et al., 2015; Strombach et al., 2015; Hill et al., 2017; Hu et al., 2018; Qu et al., 2019) and ASD populations (Kana et al., 2009; Lombardo et al., 2011; Koster-Hale et al., 2013). These previous findings indicate that the alteration of ToM ability, reflected by the functional changes of rTPJ, determines the anomaly in moral behaviors in autistic cohorts. However, it should be noted that evidence also exists, revealing that ASD individuals may preserve some degree of ToM ability to guide their intent-based moral judgments. For instance, one study showed that autistic adults not only exhibit performance comparable to that of healthy control subjects in a false belief task but also report similar moral permissibility when judging intended harms with neutral outcomes (Moran et al., 2011). Another study even reported an increased sensitivity to intention during moral judgment in Asperger's syndrome compared with healthy control subjects (Channon et al., 2011). Consistent with these studies, our RSA results also suggest that the ability to represent the information on social reputation in rTPJ is partially intact in ASD participants. These findings indicate that the ability to infer and base moral judgments on intentionality may still be present in ASD individuals, and potentially explains why we did not observe a between-group difference of rTPJ in representing social reputation in our

task. It has also been proposed that the method of inferring intentionality differs between autistic and neurotypical participants (Dempsey et al., 2020). Here, a reduced rTPJ representation similarity in ASD, unique to the moral context, explains that ASD individuals prioritize the negative consequences of an immoral action. This may block further recruitment of the intent-based system and thus lead to a lack of consideration for social reputation when making choices. Future studies may consider adopting tasks that involve both moral judgment and decision-making and implement noninvasive brain stimulation methods to target the rTPJ of ASD individuals to provide causal evidence for this possibility.

Despite the strengths of this study, there are two potential limitations. First, the sample size is relatively small for the ASD group, which could have lowered the statistical power

Table 6. Supplementary univariate GLM results

Brain region	Hemisphere	Cluster size	MNI			BA	t value	p(d-FWE)
			x	y	z			
ASD								
Public > private								
Cingulate gyrus/corpus callosum	L	96	−16	8	30		4.99	0.162
HC > ASD								
Private > public								
Cingulate gyrus/corpus callosum	L	71	−10	2	30		4.33	0.371
Good > bad								
Prec/SOG	L	95	−18	−58	32	7	4.84	0.229
IPL/PoCG	L	56	−38	−32	44	40	3.59	0.525

We excluded trials that did not reach the behavioral criterion (i.e., those with a decision time <200 ms or longer than mean \pm 3 SDs of that individual) or fMRI criterion (all trials in a run with an excessive head motion). Regions shown here met an uncorrected voxel-level threshold of $p < 0.001$ with $k = 50$. Coordinates shown here were based on the MNI coordinate system. L, Left; B, bilateral; BA, Brodmann Area; d-FWE, cluster-level FWE (corrected); CC, corpus callosum; CG, cingulate gyrus; IPL, inferior parietal lobule; PoCG, post-central gyrus; Prec, precuneus; SOG, superior occipital gyrus.

for the fMRI data analyses. Second, our sample has a relatively wide age range that covers the transition period from adolescence to early adulthood, during which time changes in sociocognitive processes and moral cognition continue to occur (Eisenberg and Morris, 2004; Blakemore and Mills, 2014; Kilford et al., 2016). Evidence indicates that mentalizing ability is still undergoing development in late adolescence (Dumontheil et al., 2010). More relevantly, previous studies have shown a distinct pattern in adolescents (vs adults) for prosocial behaviors (Padilla-Walker et al., 2018) or the susceptibility to the audience effect (Wolf et al., 2015). Importantly, these changes are considered to be crucially associated with the development of the social brain network in adolescence (Blakemore, 2008; Kilford et al., 2016). Taking TPJ as an example, evidence from brain imaging studies showed that both structural and functional features of this region vary during this transition period (Blakemore et al., 2007; Mills et al., 2014). Hence, the age-related heterogeneity of our sample may have had some impact on our results, although we controlled for age-related differences in our between-group analyses. Future studies with a larger sample or less age heterogeneity would allow more definite conclusions.

To conclude, the present study, combining computational modeling with multivariate fMRI analyses, uncovers the neuro-computational changes of the rTPJ during moral behaviors in autistic individuals. They are characterized not only by a lack of consideration for social reputation but also, more predominantly, by an increased sensitivity to the negative consequences caused by immoral actions. This difference in moral cognition and behaviors in ASD individuals is specifically associated with rTPJ and consists of a reduced capability to represent information concerning moral contexts. Our findings provide novel insights for a better understanding of the neurobiological basis underlying atypical moral behaviors in ASD individuals.

References

- Ahn W-Y, Krawitz A, Kim W, Busmeyer JR, Brown JW (2011) A model-based fMRI analysis with hierarchical Bayesian parameter estimation. *J Neurosci Psychol Econ* 4:95–110.
- Ahn W-Y, Haines N, Zhang L (2017) Revealing neuro-computational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *Comput Psychiatr* 1:24–57.
- American Psychiatric Association (2013) Diagnostic and statistical manual of mental disorders: DSM-5, Ed 5. Arlington, VA: American Psychiatric Association.
- Baron-Cohen S (2001) Theory of mind and autism: a review. *Int Rev Res Ment Retard* 23:169–184.
- Baron-Cohen S, Leslie AM, Frith U (1985) Does the autistic child have a “theory of mind”? *Cognition* 21:37–46.
- Bates D, Maechler M, Bolker B (2013) lme4: linear mixed-effects models using Eigen and syntax. Vienna, Austria: R Foundation.
- Bault N, Pelloux B, Fahrenfort JJ, Ridderinkhof KR, van Winden F (2015) Neural dynamics of social tie formation in economic decision-making. *Soc Cogn Affect Neurosci* 10:877–884.
- Bellesi G, Vyas K, Jameel L, Channon S (2018) Moral reasoning about everyday situations in adults with autism spectrum disorder. *Research in Autism Spectrum Disorders* 52:1–11.
- Blakemore S-J (2008) The social brain in adolescence. *Nat Rev Neurosci* 9:267–277.
- Blakemore S-J, Mills KL (2014) Is adolescence a sensitive period for sociocultural processing? *Annu Rev Psychol* 65:187–207.
- Blakemore S-J, Ouden D H, Choudhury S, Frith C (2007) Adolescent development of the neural circuitry for thinking about intentions. *Soc Cogn Affect Neurosci* 2:130–139.
- Buon M, Dupoux E, Jacob P, Chaste P, Leboyer M, Zalla T (2013) The role of causal and intentional judgments in moral reasoning in individuals with high functioning autism. *J Autism Dev Disord* 43:458–470.
- Burnham KP, Anderson DR (2004) Multimodel inference: understanding AIC and BIC in model selection. *Sociol Methods Res* 33:261–304.
- Carter RM, Bowling DL, Reeck C, Huettel SA (2012) A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science* 337:109–111.
- Channon S, Lagnado D, Fitzpatrick S, Drury H, Taylor I (2011) Judgments of cause and blame: sensitivity to intentionality in Asperger’s syndrome. *J Autism Dev Disord* 41:1534–1542.
- Crockett MJ (2016) How formal models can illuminate mechanisms of moral judgment and decision making. *Curr Dir Psychol Sci* 25:85–90.
- Crockett MJ, Kurth-Nelson Z, Siegel JZ, Dayan P, Dolan RJ (2014) Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences* 111:17320–17325.
- Crockett MJ, Siegel JZ, Kurth-Nelson Z, Dayan P, Dolan RJ (2017) Moral transgressions corrupt neural representations of value. *Nature Neuroscience* 20:879–885.
- D’Cruz A-M, Ragozzino ME, Mosconi MW, Shrestha S, Cook EH, Sweeney JA (2013) Reduced behavioral flexibility in autism spectrum disorders. *Neuropsychology* 27:152–160.
- Dempsey E, Moore C, Johnson S, Stewart S, Smith I (2020) Morality in autism spectrum disorder: a systematic review. *Dev Psychopathol* 32:1069–1085.
- Dumontheil I, Apperly IA, Blakemore S-J (2010) Online usage of theory of mind continues to develop in late adolescence. *Dev Sci* 13:331–338.
- Eisenberg N, Morris AS (2004) Moral cognitions and prosocial responding in adolescence. In: *Handbook of adolescent psychology*, Ed 2 (Lerner RM, Steinberg L, eds), pp 155–188. New York: Wiley.
- Eklund A, Nichols TE, Knutsson H (2016) Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci U S A* 113:7900–7905.
- Fadda R, Parisi M, Ferretti L, Saba G, Foscoliano M, Salvago A, Doneddu G (2016) Exploring the role of theory of mind in moral judgment: the case of children with autism spectrum disorder. *Front Psychol* 7:523.

- Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. *Q J Econ* 114:817–868.
- Fox J, Weisberg S, Adler D, Bates D, Baud-Bovy G, Ellison S, Firth D, Friendly M, Gorjanc G, Graves S (2016) Package “car”. Vienna, Austria: R Foundation.
- Frith U, Frith C (2011) Reputation management: in autism, generosity is its own reward. *Curr Biol* 21:R994–R995.
- Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Stat Sci* 7:457–472.
- Gelman A, Carlin JB, Stern HS, Rubin DB (2014) Bayesian data analysis. Boca Raton, FL: Chapman & Hall/CRC.
- Gelman A, Lee D, Guo J (2015) Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics* 40:530–543.
- Haley KJ, Fessler DMT (2005) Nobody’s watching?: Subtle cues affect generosity in an anonymous economic game. *Evol Hum Behav* 26:245–256.
- Hampton AN, Bossaerts P, O’Doherty JP (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proc Natl Acad Sci U S A* 105:6741–6746.
- Hebart MN, Baker CI (2018) Deconstructing multivariate decoding for the study of brain function. *Neuroimage* 180:4–18.
- Hill CA, Suzuki S, Polania R, Moisa M, O’Doherty JP, Ruff CC (2017) A causal account of the brain network computations underlying strategic social behavior. *Nat Neurosci* 20:1142–1149.
- Hu Y, He L, Zhang L, Wölk T, Dreher JC, Weber B (2018) Spreading inequality: neural computations underlying paying-it-forward reciprocity. *Soc Cogn Affect Neurosci* 13:578–589.
- Hutcherson C, Bushong B, Rangel A (2015) A neurocomputational model of altruistic choice and its implications. *Neuron* 87:451–462.
- Izuma K (2012) The social neuroscience of reputation. *Neurosci Res* 72:283–288.
- Izuma K, Saito DN, Sadato N (2008) Processing of social and monetary rewards in the human striatum. *Neuron* 58:284–294.
- Izuma K, Saito DN, Sadato N (2010a) The roles of the medial prefrontal cortex and striatum in reputation processing. *Soc Neurosci* 5:133–147.
- Izuma K, Saito DN, Sadato N (2010b) Processing of the incentive for social approval in the ventral striatum during charitable donation. *J Cogn Neurosci* 22:621–631.
- Izuma K, Matsumoto K, Camerer CF, Adolphs R (2011) Insensitivity to social reputation in autism. *Proc Natl Acad Sci U S A* 108:17302–17307.
- Kana RK, Keller TA, Cherkassky VL, Minshew NJ, Just MA (2009) Atypical frontal-posterior synchronization of theory of mind regions in autism during mental state attribution. *Soc Neurosci* 4:135–152.
- Kilford EJ, Garrett E, Blakemore SJ (2016) The development of social cognition in adolescence: an integrated perspective. *Neurosci Biobehav Rev* 70:106–120.
- Kononov A, Hu J, Ruff CC (2018) Neurocomputational approaches to social behavior. *Curr Opin Psychol* 24:41–47.
- Koster-Hale J, Saxe R, Dungan J, Young LL (2013) Decoding moral judgments from neural representations of intentions. *Proc Natl Acad Sci U S A* 110:5648–5653.
- Kriegeskorte N, Mur M, Bandettini PA (2008) Representational similarity analysis-connecting the branches of systems neuroscience. *Front Syst Neurosci* 2:4.
- Lombardo MV, Chakrabarti B, Bullmore ET, Baron-Cohen S, Consortium MA (2011) Specialization of right temporo-parietal junction for mentalizing and its relation to social impairments in autism. *Neuroimage* 56:1832–1838.