



**HAL**  
open science

## Right temporoparietal junction underlies avoidance of moral transgression in Autism Spectrum Disorder

Y Hu, Am Pereira, X Gao, Bm Campos, Edmund A. Derrington, Brice Corgnet, X Zhou, F Cendes, Jean-Claude Dreher

### ► To cite this version:

Y Hu, Am Pereira, X Gao, Bm Campos, Edmund A. Derrington, et al.. Right temporoparietal junction underlies avoidance of moral transgression in Autism Spectrum Disorder. *Journal of Neuroscience*, 2021, 41 (8), pp.1699-1715. 10.1523/jneurosci.1237-20.2020 . hal-02990821

**HAL Id: hal-02990821**

**<https://hal.science/hal-02990821>**

Submitted on 9 Nov 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

*Research Articles: Behavioral/Cognitive*

## Right temporoparietal junction underlies avoidance of moral transgression in Autism Spectrum Disorder

<https://doi.org/10.1523/JNEUROSCI.1237-20.2020>

**Cite as:** J. Neurosci 2020; 10.1523/JNEUROSCI.1237-20.2020

Received: 25 May 2020

Revised: 27 October 2020

Accepted: 28 October 2020

---

*This Early Release article has been peer-reviewed and accepted, but has not been through the composition and copyediting processes. The final version may differ slightly in style or formatting and will contain links to any extended data.*

**Alerts:** Sign up at [www.jneurosci.org/alerts](http://www.jneurosci.org/alerts) to receive customized email alerts when the fully formatted version of this article is published.

---

1 **Title:** Right temporoparietal junction underlies avoidance of moral transgression in Autism  
2 Spectrum Disorder

3 **Abbreviated title:** Moral decisions and autism

4 **Authors information:**

5 Yang Hu<sup>1,2,5†</sup>, Alessandra M. Pereira<sup>3†</sup>, Xiaoxue Gao<sup>5</sup>,

6 Brunno M. Campos<sup>3</sup>, Edmund Derrington<sup>2,4</sup>, Brice Corgnet<sup>7</sup>,

7 Xiaolin Zhou<sup>1,5,6</sup>, Fernando Cendes<sup>3</sup>, Jean-Claude Dreher<sup>2,4\*</sup>

8 <sup>1</sup>Key Laboratory of Applied Brain and Cognitive Sciences, School of Business and  
9 Management, Shanghai International Studies University, 201620 Shanghai, China

10 <sup>2</sup>Laboratory of Neuroeconomics, Institut des Sciences Cognitives Marc Jeannerod, CNRS,  
11 69675 Bron, France

12 <sup>3</sup>Neuroimaging Laboratory, School of Medical Sciences, The Brazilian Institute of Neuroscience  
13 and Neurotechnology, University of Campinas (UNICAMP), 13083-970 Campinas, Brazil

14 <sup>4</sup>Université Claude Bernard Lyon 1, 69100 Villeurbanne, France

15 <sup>5</sup>School of Psychological and Cognitive Sciences, Peking University, 100871 Beijing, China

16 <sup>6</sup>PKU-IDG/McGovern Institute for Brain Research, Peking University, 100871 Beijing, China

17 <sup>7</sup>EmLyon, 69130 Ecully, France

18

19 \*Correspondence to: [dreher@isc.cnrs.fr](mailto:dreher@isc.cnrs.fr)

20 †These authors equally contributed to this study.

21 **Number of pages: 61**

22 **Number of figures: 10**

23 **Number of tables: 6**

24 **Number of words for abstract: 232**

25 **Number of words for introduction: 646**

26 **Number of words for discussion: 2,077**

---

27 **Competing interests:** The authors declare no competing interests.

28 **Acknowledgments:** J-C.D. was funded by the IDEX-LYON from Université de Lyon (project  
29 INDEPTH) within the Programme Investissements d'Avenir (ANR-16-IDEX-0005) and of the  
30 LABEX CORTEX (ANR-11-LABX-0042) of Université de Lyon, within the program  
31 Investissements d'Avenir (ANR-11-IDEX-007) operated by the French National Research  
32 Agency, and grants from the Agence Nationale pour la Recherche and NSF in the CRCNS  
33 program (ANR n°16-NEUC-0003-01) and from Fondation de France (n°89590). F.C. was funded  
34 by São Paulo Research Foundation (FAPESP n° 2013/07559-3) of Brazil. Y.H. was funded by  
35 China Postdoctoral Science Foundation (2019M660007). X.G. and X.Z. were supported by  
36 National Basic Research Program of China (973 Program: 2015CB856400) and National  
37 Natural Science Foundation of China (91232708, 31170972, 31630034, 71942001). X.G. was  
38 supported by China Postdoctoral Science Foundation (2019M650008) and National Natural  
39 Science Foundation of China (31900798). We thank the staff of the Imaging Center of the  
40 University of Campinas for helpful assistance with data collection for the fMRI study.

---

41 **Abstract**

42 Autism spectrum disorder (ASD) is characterized by a core deficit in theory-of-mind (ToM) ability,  
43 which extends to perturbations in moral judgment and decision-making. Although the function of  
44 the right temporoparietal junction (rTPJ), a key neural marker of ToM and morality, is known to  
45 be altered in autistic individuals, the neurocomputational mechanisms underlying its specific  
46 impairment in moral decision-making remain unclear. Here, we addressed this question by  
47 employing a novel fMRI task together with computational modeling and representational  
48 similarity analysis (RSA). ASD patients and healthy controls (HC) decided in public or private  
49 whether to incur a personal cost for funding a morally-good cause (Good Context) or receive a  
50 personal gain for benefiting a morally-bad cause (Bad Context). Compared with HC, individuals  
51 with ASD were much more likely to reject the opportunity to earn ill-gotten money by supporting  
52 a bad cause than HC. Computational modeling revealed that this resulted from unduly weighing  
53 benefits for themselves and the bad cause, suggesting that ASD patients apply a rule of  
54 refusing to serve a bad cause because they over-evaluate the negative consequences of their  
55 actions. Moreover, RSA revealed a reduced rTPJ representation of the information specific to  
56 moral contexts in ASD patients. Together, these findings indicate the contribution of rTPJ in  
57 representing information concerning moral rules and provide new insights for the  
58 neurobiological basis underpinning moral behaviors illustrated by a specific dysfunction of rTPJ  
59 in ASD patients.

60

61 **Significance**

62 Previous investigations have found an altered pattern of moral behaviors in individuals with  
63 autism spectrum disorder (ASD), closely associated with a dysfunction in the right  
64 temporoparietal junction (rTPJ). However, the specific neurocomputational mechanisms at play

---

65 that drive the dysfunction of the rTPJ in moral decision-making remain unclear. Here, we show  
66 that ASD individuals are more inflexible when following a moral rule even though an immoral  
67 action can benefit themselves, and suffer an undue concern about their ill-gotten gains and the  
68 moral cost. Moreover, a selectively reduced rTPJ representation of information concerning  
69 moral rules was observed in ASD patients. These findings deepen our understanding of the  
70 neurobiological roots that underlie atypical moral behaviors in ASD patients.

---

71 **Introduction**

72           Autism spectrum disorder (ASD) is a complex neurodevelopment disorder with evident  
73 impairments in social interaction, communication, and interpersonal relationships (APA, 2013),  
74 which are critically dependent on theory-of-mind (ToM) ability (Young et al., 2007; Margoni and  
75 Surian, 2016). These social deficits are also associated with atypical moral cognition. Indeed,  
76 individuals with ASD have difficulties in evaluating the moral appropriateness of actions in terms  
77 of the intentions of the protagonists in hypothetical scenarios (Moran et al., 2011; Buon et al.,  
78 2013; Fadda et al., 2016). ASD patients also conduct aberrant moral behaviors that lead to real  
79 consequences. For example, they are less sensitive to observation by others while making  
80 charitable decisions (Izuma et al., 2011).

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

---

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

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

113           Our second hypothesis was inspired by studies of moral judgments that reveal autistic  
114 individuals tend to judge moral culpability more often in terms of consequences (Moran et al.,  
115 2011; Fadda et al., 2016; Salvano-Pardieu et al., 2016), and often over-evaluate the negative  
116 moral consequences (Moran et al., 2011; Bellesi et al., 2018). Hence, it was possible that  
117 compared to HC, ASD patients would display increased aversion to the consequences of an  
118 immoral action and therefore reject more offers that earn themselves morally-tainted profits. We



---

119 further explored whether such behavioral differences could be explained by a reduced rTPJ  
120 representation of information concerning moral contexts in ASD patients.

---

121 **Materials and Methods**

122 **Participants**

123 A total of 48 participants were recruited for the present fMRI experiment. Specifically, 20  
124 individuals with autism spectrum disorder (ASD; 4 females;  $17.0 \pm 3.0$  years, ranging from 14 to  
125 24 years; 3 left handedness) were recruited via those who attended psychiatric and pediatric  
126 neurology clinics as outpatients and fulfilled the inclusion criteria. 28 healthy controls (HC; 10  
127 females;  $18.9 \pm 3.0$  years, ranging from 14 to 25 years; 1 left handed) were recruited from the  
128 local community via fliers. Diagnoses of ASD were performed by a clinical pediatric neurologist  
129 according to the Autism Diagnostic Interview-revised (ADI-R; see **Table 1** for all clinical tests  
130 describing the two samples). There were no significant between-group differences in gender  
131 ( $\chi(1)^2 = 1.395$ ,  $p = 0.238$ ) and IQ (total:  $t(43) = -1.795$ ,  $p = 0.080$ ; verbal:  $t(43) = -1.379$ ,  $p =$   
132  $0.175$ ; execution:  $t(43) = -1.421$ ,  $p = 0.162$ ), except that ASD participants were slightly younger  
133 than HC participants ( $t(46) = -2.121$ ,  $p = 0.039$ ).

134 The study was performed at the Imaging Center of the University of Campinas and  
135 approved by the local ethics committee (plataformabrasil.saude.gov.br; reference number:  
136 CAAE 02388012.5.0000.5404; approved ethical statement: No. 1904090). All experimental  
137 protocols and procedures were conducted in accordance with the IRB guidelines for  
138 experimental testing and complied with the latest revision of the Declaration of Helsinki (BMJ  
139 1991; 302: 1194).

140

141 **Experimental Design and Task**

142 We adopted a  $2 \times 2$  within-subject design with a novel paradigm (also see Obeso et al.,  
143 2018; Qu et al., 2019). Specifically, participants decided whether to accept or reject a series of  
144 offers consisting of a personal profit or loss and a donation to a certain association, either in  
145 absence or presence of an audience (i.e., Audience: Private vs. Public; see Procedure for

---

146 details). In half of the trials, participants were confronted with offers involving a monetary loss  
147 for themselves but a financial gain to a local charity, “*The Child Hope Campaign*”  
148 ([www.redeglobo.globo.com/criancaesperanca](http://www.redeglobo.globo.com/criancaesperanca)), which supports the education of children and  
149 adolescents in Brazil. In the other trials, participants considered offers that comprised a  
150 monetary gain for themselves but also a financial gain benefiting a morally-bad cause, “*No Dogs  
151 and Cats*” ([www.naoaesegatos.net](http://www.naoaesegatos.net)), which aims to clean the street by exterminating street  
152 animals. In other words, we manipulated moral contexts (i.e., Good vs. Bad) according to the  
153 cause involved in offers. In total, the present design yielded four experimental conditions,  
154  $\text{Public}_{\text{Good}}$ ,  $\text{Public}_{\text{Bad}}$ ,  $\text{Private}_{\text{Good}}$ , and  $\text{Private}_{\text{Bad}}$ . Crucially, participants were informed that their  
155 decisions could have real consequences. Thus, if participants accepted the offer in the Good  
156 context, they would lose a certain amount of money and the charity would be paid. If they do so  
157 in the Bad context, they would earn the money and the bad cause would also be paid. However,  
158 if participants rejected the offer, neither they nor the involved association would gain or lose any  
159 money (see **Figure 1**). Participants were also informed that all trials (decisions) were  
160 independent from each other so that the incentive consequences would not accumulate across  
161 the experiment. Only one trial would be randomly selected and paid at the end of the  
162 experiment

163       One key aspect of the present design was that we varied the monetary stakes for the  
164 participants and the associations independently across trials within each condition. Personal  
165 payoffs (i.e., profits or losses) ranged from 1 to 8 in steps of 1 (unit: Brazillian Real; 1 Brazilian  
166 Real  $\approx$  0.2 USD). Donations to both associations ranged from 4 to 32 in the steps of 4. The  
167 personal payoff and the donation were orthogonal, which led to 64 different offers. Each offer  
168 appeared only once in each condition and thus summed up to 256 trials in total.

---

169           The functional scanning comprised four runs of 64 trials. Each moral context was  
170 assigned to either the first or the second of two runs. Each run consisted of two blocks, which  
171 included 32 trials presenting unique offers in either the Private or the Public condition. The order  
172 of runs involving Good/Bad moral contexts and Public/Private blocks were counterbalanced  
173 across participants. The trial order was randomized within each block. For each trial,  
174 participants were presented with the decision screen consisting of the payoff information for the  
175 participant (monetary gain or loss), and the association indicated by the corresponding symbol.  
176 The cue that signaled whether it was a Public (a picture of eyes) or a Private (i.e., a picture of a  
177 padlock) trial was also shown on the same screen. Here a cue of being watched was used as  
178 previous studies have consistently shown that it influences individuals' behaviors (Haley and  
179 Fessler, 2005; Izuma et al., 2008). Participants decided whether to accept or reject the offer by  
180 pressing the corresponding button on the button box with the right index or middle finger at their  
181 own pace. In the Private condition, once a response was made, the screen was unchanged for  
182 0.5s to keep the chosen option private. In the Public condition, the chosen option was  
183 highlighted with a larger font, and the non-chosen option disappeared, which lasted slightly  
184 longer (1.5s) to further emphasize the presence of a witness (Qu et al., 2019). This was  
185 followed by a uniformly jittered fixation (2.5 – 6.5s), which ended the trial.

186           All visual stimuli were presented using Presentation v14 (Neurobehavioral Systems Inc.,  
187 *Albany, CA, USA*) back-projected on a screen outside the scanner, using a mirror system  
188 attached to the head coil.

### 189 **Procedure**

190           On the day of scanning, participants (and their legal guardians when necessary) first  
191 signed the written informed consent and then were given the instructions. After that, they  
192 completed a series of comprehension questions to ensure that they fully understood the task.  
193 Importantly, they met with an independent audience and were informed that this person would

---

194 sit in the control room to witness their choices in some trials (i.e., in the Public condition) during  
195 the experiment. In the scanner, participants completed a practice session to get familiar with the  
196 paradigm and the response button. The scanning part consisted of four functional runs lasting  
197 around 35 min, which was followed by a 6-min structural scan. After that, participants indicated  
198 their liking for each association on an 11-point Likert scale (0 indicated “dislike very much”, 10  
199 indicated “like very much”). Finally, participants were debriefed, paid, and thanked.

200

### 201 **Data Acquisition**

202 The imaging data were acquired on a 3-Tesla Philips Achieva MRI system with a 32-  
203 channel head coil (Best, *The Netherlands*) at the Imaging Center of University of Campinas.  
204 Functional data were acquired using T2\*-weighted echo-planar imaging (EPI) sequences  
205 employing a BOLD contrast (TR = 2000 ms, TE = 30 ms; flip angle = 90°; slice thickness = 3  
206 mm without gap, matrix = 80 × 80, FoV = 240 × 240 mm<sup>2</sup>) in 40 axial slices. Slices were axially  
207 oriented along the AC-PC plane and acquired in an ascending order. A high-resolution structural  
208 T1-weighted image was also collected for every participant using a 3D MRI sequence (TR = 7  
209 ms, TE = 3.2 ms; flip angle = 8°; slice thickness = 1 mm, matrix = 240 × 240, FoV = 240 × 240  
210 mm<sup>2</sup>).

211

### 212 **Statistical Analyses**

213 One ASD participant was excluded from behavioral analyses due to the invariant  
214 response pattern (i.e., rejecting all trials in the task). After checking the preprocessed fMRI data,  
215 we excluded two more HC participants (one because half of the scanning data was lost due to a  
216 technical reason, the other for excessive head motion [i.e., > 3 mm] in two out of four runs) and  
217 one more ASD participant (due to excessive head motion [i.e., > 5 mm] in three out of four

---

218 functional runs). Thus, 26 HC participants and 18 ASD participants were included for the fMRI  
219 analyses.

220

221

222

### 223 *Behavioral Analyses*

224 All behavioral analyses were conducted using R (<http://www.r-project.org/>) (R Core  
225 Team, 2014). All reported p values are two-tailed and  $p < 0.05$  was considered statistically  
226 significant. Data visualization was performed via the “ggplot2” package (Wickham, 2016). We  
227 excluded trials with either extremely fast responses (i.e.,  $< 200$  ms) or extremely slow response  
228 (i.e., exceeding 3 SD of the individual mean decision time) from both the behavioral analyses  
229 and the model-based analyses. The percentage of trials excluded due to the criteria of decision  
230 time was 1.63% for the HC group and 1.89% for the ASD group.

231 For ease of interpretation, we defined the moral choices as those in which the participant  
232 accepted offers in the Good context or rejected offers in the Bad context. We performed the  
233 repeated mixed-effect logistic regression predicting the moral choice by the glmer function in  
234 “lme4” package (Bates et al., 2013), with Group (dummy variable; reference level: HC; same  
235 below), Audience (dummy variable; reference level: private; same below), Moral Context  
236 (dummy variable; reference level: good; same below), and their interactions (i.e., 3 two-way  
237 interactions and 1 three-way interaction) as the fixed-effect predictors. We also incorporated  
238 age as a covariate in the analyses to rule out its possible confounding effect. We included  
239 random-effect predictors that allowed varying intercepts across participants. For the statistical  
240 inference on each predictor, we performed the Type II Wald chi-square test on the model fits by  
241 using the Anova function in “car” package (Fox et al., 2016). Once the interactions were  
242 detected, we ran post-hoc regressions on the subset of data given the different groups and then

---

243 conditions. We reported the odds ratio (OR) as an index of effect size of each predictor on moral  
244 choices.

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

### 258 *Computational Modeling*

259 To examine how participants evaluated payoffs of each party and integrated them into a  
260 subjective value (SV), we compared the following 8 models with different utility functions  
261 characterizing participants’ choices.

262 Model 1 was adapted from a recent study on moral decision making by Crockett *et al*  
263 (2014, 2015, 2017), which could be formally represented as follows:

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

264 where SV denotes the SV of the given trial if the participant chooses to accept. For rejection  
 265 trials, SV is always 0 given the rule of the task (i.e., neither beneficiaries would gain the money;  
 266 same for all models).  $M_S$  and  $M_O$  represent the payoff (gain or loss) for oneself and payoffs  
 267 donated to the corresponding association.  $\alpha$  ( $0 < \alpha < 1$ ) is the unknown parameter of social  
 268 preference which arbitrates the relative weight on the payoff for the participant in the decision.  $\theta$   
 269 ( $0 < \theta < 1$ ) is the unknown parameter characterizing the audience effect, which is modulated by  
 270 an indicator function  $q$  (0 for private, 1 for public; same below). This model assumes that the  
 271 subjective value was computed as a weighted summation of personal payoffs and payoffs  
 272 donated to the association, and that people cared less about their own payoffs but increased the  
 273 weights on the benefits donated to the association in public (vs. private). Model 2 was similar to  
 274 Model 1 except that it adopted two separate  $\alpha$  depending on the moral context in that trial.

275 Model 3 has a logic similar to Model 1, and was built upon studies adopting a donation  
 276 task (Lopez-Persem et al., 2017; Qu et al., 2020):

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

277 where  $\alpha$  and  $\beta$  are unknown parameters which capture the weight of the payoff for either the  
 278 participant or the association involved in the trial ( $-20 < \alpha, \beta < 20$ ). Again,  $\theta$  ( $0 < \theta < 10$ )  
 279 describes the audience effect which is represented by the indicator function  $q$ . Model 4 was  
 280 similar to Model 3 except that it adopted two separate pairs of  $\alpha$  and  $\beta$  according to the  
 281 association involved in that trial (i.e., good cause or the bad cause).

282 Models 5-8 were established on the basis of the Fehr-Schmidt model (Fehr and Schmidt,  
 283 1999):

$$SV(M_S, M_O) = \begin{cases} -M_S - \alpha * \max(M_O + M_S, 0) - \beta * \max(-M_S - M_O, 0) & \text{if Good} \\ M_S - \alpha * \max(M_O - M_S, 0) - \beta * \max(M_S - M_O, 0) & \text{if Bad} \end{cases}$$



---

284 where  $\alpha$  and  $\beta$  measure the degree of aversion to payoff inequality in disadvantageous  
285 and advantageous situations respectively (i.e., how participants dislike that they themselves  
286 gained less/more than the association;  $0 < \alpha, \beta < 5$ ). Among them, Model 5 adopted a fixed pair  
287 of  $\alpha$  and  $\beta$  in all four conditions. Model 6 and Model 7 took different pairs of  $\alpha$  and  $\beta$  either in  
288 terms of the audience or the moral context. Model 8 assumed that people showed distinct  
289 advantageous and disadvantageous inequality aversion that changed in each of the four  
290 conditions.

291 Given the softmax rule, we could estimate the probability of making a moral choice (i.e.,  
292 accept in the Good context or reject in the Bad context) as below:

293

$$p(SV_{moral}) = \frac{e^{\tau SV_{moral}}}{e^{\tau SV_{moral}} + e^{\tau SV_{immoral}}}$$

294 where  $\tau$  refers to the inverse softmax temperature ( $0 < \tau < 10$ ) which denotes the sensitivity of  
295 individual's behavior to the difference in SV between moral and immoral choices.

296 We leveraged a Hierarchical Bayesian Analysis (HBA) approach (Gelman et al., 2014) to  
297 fit all the above candidate models via the "hBayesDM" package (Ahn et al., 2017). In general,  
298 HBA has several advantages over the traditional maximal likelihood estimation (MLE) approach  
299 such that it could provide more stable and accurate estimates, and estimate the posterior  
300 distribution of both the group-level and individual-level parameters simultaneously (Ahn et al.,  
301 2011). The "hBayesDM" package performs a full Bayesian inference and provides actual  
302 posterior distribution using a Markov Chain Monte Carlo (MCMC) sampling manner through the  
303 Stan language (Stan Development Team, 2016). Conforming to the default setting in this  
304 package, we assumed the individual-level parameters were drawn from a group-level normal  
305 distribution: *individual-level parameters*  $\sim$  *Normal* ( $\mu, \sigma$ ). We fit each candidate model with four

---

306 independent MCMC chains using 1,000 iterations after 2,000 iterations for the initial algorithm  
307 warmup per chain that results in 4,000 valid posterior samples. The convergence of the MCMC  
308 chains was assessed through Gelman-Rubin R-hat Statistics (Gelman and Rubin, 1992).

309 For model comparisons, we computed the leave-one-out information criterion (LOOIC)  
310 score for each candidate model (Bault et al., 2015). LOOIC score provides the estimate of out-  
311 of-sample predictive accuracy in a fully Bayesian way, which makes it more reliable than the  
312 point-estimate information criterion (e.g., AIC). By convention, the lower LOOIC score indicates  
313 better out-of-sample prediction accuracy of the candidate model. A difference score of 10 on the  
314 information criterion scale is considered decisive (Burnham and Anderson, 2004). We selected  
315 the model with the lowest LOOIC as the winning model for subsequent analysis of key  
316 parameters. A posterior predictive check was additionally implemented to examine the absolute  
317 performance of the winning model. In other words, we tested whether the prediction of the  
318 winning model could capture the actual behaviors. In terms of the actual trial-wise stimuli  
319 sequences, we employed each individual's joint posterior MCMC samples (i.e., 4,000 times) to  
320 generate new choice datasets correspondingly (i.e., 4,000 choices per trial per participant).  
321 Then we calculated the mean proportion of moral choices of each experimental condition in  
322 these new datasets for each subject, respectively. We performed *Pearson* correlation to  
323 examine to what degree the predicted proportion of moral choice correlated with the actual  
324 proportion across individuals in each condition, respectively.

325

#### 326 *fMRI Data Preprocessing*

327 Functional imaging data were analyzed using SPM12 (Wellcome Trust Centre for  
328 Neuroimaging, University College London, *London, UK*). The preprocessing procedure followed  
329 the pipeline recommended by SPM12. In particular, functional images (EPI) were first realigned

---

330 to the first volume to correct motion artifacts, unwarped, and corrected for slice timing. Next, the  
331 structural  $T_1$  image was segmented into white-matter, grey-matter and cerebrospinal fluid with  
332 the skull removed, and co-registered to the mean functional images. Then all functional images  
333 were normalized to the MNI space, resampled with a  $2 \times 2 \times 2 \text{ mm}^3$  resolution, in terms of  
334 parameters generated in the previous step. Last, the normalized functional images were  
335 smoothed using an 8-mm isotropic full width half maximum (FWHM) based on a Gaussian  
336 kernel.

337

### 338 *Within-Subject Representational Similarity Analyses (RSA)*

339 To clarify what information rTPJ exactly represents during the decision period that  
340 distinguished ASD patients from HC participants, we carried out a within-subject RSA in Python  
341 3.6.8 using the *nltools* package (version 0.3.14; <https://github.com/cosanlab/nltools>). Some  
342 preparation was performed before implementing RSA. In particular, we established a trial-wise  
343 general linear model (GLM) for each participant, which included the onsets of the decision  
344 screen with the duration of decision time of each valid trial. Here, valid trials were those that  
345 conformed to neither the exclusion criterion for the behavioral data (trials with extremely fast or  
346 slow responses; see above for details) nor the fMRI data (trials in runs with excessive head  
347 motion). The onsets of button press and invalid trials were also modeled as separate regressors  
348 of no interest. In addition, six movement parameters were added to this GLM as covariates to  
349 account for artifacts of head motion. The canonical hemodynamic response function (HRF) was  
350 used and a high-pass temporal filtering was performed with a default cut-off value of 128s to  
351 remove low-frequency drifts. After the parameter estimation, we built up the trial-wise contrasts  
352 which were used for subsequent RSA.

---

353 Our analyses concentrated on rTPJ given our hypotheses. Notably, we took two different  
354 ways to define the cluster of rTPJ to circumvent the potential effect of ROI selection on results.  
355 These included defining it via a whole-brain parcellation based on meta-analytic functional  
356 coactivation of the Neurosynth database (i.e., the parcellation-based ROI;  
357 <https://neurovault.org/collections/2099/>; including a total of 1,750 voxels, with a volume of  
358  $2 \times 2 \times 2 \text{ mm}^3$  per voxel, same below) or via a coordinate-based manner given a recent meta-  
359 analysis on neural correlates of ToM (Schurz et al., 2014) (i.e., the coordinate-based ROI; a  
360 sphere with a radius of 10mm centering on the MNI coordinates of 56/-56/18; 515 voxels in  
361 total).

362 We first extracted the parameter estimates (i.e., contrast value in arbitrary units) of rTPJ  
363 from these 1<sup>st</sup>-level contrast images of valid trials for each participant respectively. Next, we  
364 constructed the individual-level neural representation distance matrix (RDM) by computing the  
365 pairwise correlation dissimilarity of activation patterns within this mask between each pair of  
366 valid trials. We also built up the same neural RDM for ITPJ as a control region (i.e., the  
367 parcellation-based ROI: 1,626 voxels in total; the coordinate-based ROI (Schurz et al., 2014), a  
368 sphere with a radius of 10mm centering on the MNI coordinates of -53/-59/20; 515 voxels in  
369 total). In line with our research goal, we constructed two main cognitive RDMs in light of the  
370 trial-wise information of reputation (i.e., arbitrary code: 0 = Private, 1 = Public), and Moral  
371 Context (i.e., 0 = Bad, 1 = Good) by calculating the Euclidean distance between each pair of  
372 trials. We also built up two additional cognitive RDMs using the trial-wise information of payoffs  
373 for the participant (i.e., from 1 to 8 in step of 1), and payoffs for associations (i.e., from 4 to 32 in  
374 step of 4) as controls. These cognitive RDMs measured the dissimilarity between trials given  
375 corresponding information. Notably, we sorted all trials according to the order of Audience,  
376 Moral Context, payoff for the participant, and payoff for associations (the charity or the bad  
377 cause) to guarantee the information contained by both the neural and cognitive RDMs was

---

378 matched with each other. To make these cognitive RDMs comparable, we rescaled them within  
379 the range from 0 (i.e., the most similar) to 1 (i.e., the most dissimilar). Then we performed  
380 Spearman's rank-order correlation between the neural RDM and the cognitive RDM for each  
381 participant.

382 For the group-level statistical tests, we first implemented the Fisher r-to-z transformation  
383 on the Spearman's rho, and then performed the permutation-based two-sample t-test (i.e., the  
384 number of permutations: 5,000) on these statistics between the two groups for each cognitive  
385 RDM separately. To further examine the robustness of these findings, we applied the above  
386 analyses using all 256 trials. To this end, a new GLM was established that modeled the onset of  
387 the decision screen of all trials to further construct the neural RDM. The remaining details and  
388 procedures were the same as mentioned above.

#### 389 *Supplementary Univariate Analyses*

390 We also performed a traditional univariate GLM analysis to examine whether the mean  
391 neural activations were modulated by different conditions and how neural signals in ASD  
392 patients differed from healthy controls, focusing on the rTPJ. At the individual level, we  
393 incorporated the onsets of the decision phase of all conditions (i.e., Private<sub>Good</sub>, Private<sub>Bad</sub>,  
394 Public<sub>Good</sub>, Public<sub>Bad</sub>) in valid trials as regressors of interest. Similarly, the onsets of button press  
395 together with invalid trials as well as head motion parameters were also modelled as separate  
396 regressors of no interest. After the parameter estimation, we constructed the following contrasts  
397 concerning the main effect of Audience (i.e., Public - Private) and Moral Context (i.e., Good -  
398 Bad). These contrast images were fed to the group-level one-sample T-test for within-group  
399 analyses or independent two-sample T-tests for between-group analyses. Given the goal of this  
400 analysis, we performed a small volume correction (SVC) within the rTPJ mask. To match the  
401 multivariate analyses, we adopted two independent rTPJ masks from different sources (i.e., the  
402 parcellation-based ROI and the coordinate-based ROI; see above for details). For the

---

403 completeness of the analyses, we also performed the same analyses using the ITPJ mask.  
404 Otherwise, we adopted a whole-brain threshold of  $p < 0.001$  uncorrected at the voxel-level  
405 together with  $p < 0.05$  FWE corrected at the cluster-level (Eklund et al., 2016).

---

406 **Results**

407 **Subjective evaluation on associations**

408 Post-task rating on a 0-10 Likert scale (0 indicates “do not like the association at all”, 10  
409 indicates “like the association very much”) revealed that both ASD patients and healthy controls  
410 favored the charity (ASD vs. healthy controls:  $9.3 \pm 1.4$  vs.  $8.8 \pm 1.2$ ) and disliked the bad cause  
411 ( $0.0 \pm 0.0$  vs.  $0.3 \pm 0.6$ ). No between-group difference was observed in the subjective rating for  
412 the charity ( $b = 0.43$ ,  $SE = 0.39$ ,  $t(44) = 1.11$ ,  $p = 0.274$ ,  $b_z = 0.17$ ) and bad cause ( $b = -0.22$ ,  $SE$   
413  $= 0.14$ ,  $t(44) = -1.56$ ,  $p = 0.125$ ,  $b_z = -0.23$ ).

414

415 **ASD patients not only fail to consider social reputation but also rigorously conform to a**  
416 **rule in curbing their immoral behaviors**

417 Mixed-effect logistic regressions revealed that participants were more likely to behave  
418 morally in the Bad than the Good context (i.e., rejecting more frequently the offer in the Bad  
419 context than accepting it in the Good context; a main effect of Moral Context:  $\chi^2(1) = 632.68$ ,  $p <$   
420  $0.001$ ). More importantly, significant interaction effects were identified between Group and  
421 Audience ( $\chi^2(1) = 4.50$ ,  $p = 0.034$ ) as well as between Group and Moral Context ( $\chi^2(1) = 59.33$ ,  
422  $p < 0.001$ ) on choosing the moral option (i.e., accepting the offer to benefit a charity or rejecting  
423 the offer to benefit a morally-bad cause; see **Figure 2**). No other main effect ( $ps > 0.09$ ) or  
424 interaction effect was detected ( $ps > 0.57$ ).

425 To understand the first interaction effect, we performed *post-hoc* analyses on the dataset  
426 of the ASD and the HC groups, respectively. For each analysis, we ran a similar logistic  
427 regression, including the main effect of audience and context as the fixed-effect predictors. The  
428 Audience  $\times$  Moral Context interaction was dropped from these analyses as neither this effect  
429 ( $\chi^2(1) = 0.31$ ,  $p = 0.580$ ) nor the three-way interaction effect ( $\chi^2(1) = 0.30$ ,  $p = 0.586$ ) was

---

430 significant in the main analysis. The results showed that while healthy controls were more likely  
431 to make the moral choice when they were observed in the Public condition (vs. Private; OR =  
432 1.16,  $b = 0.15$ ,  $SE = 0.06$ ,  $p = 0.012$ ), ASD patients did not change their behaviors significantly  
433 depending on the presence or absence of a witness (OR = 0.93,  $b = -0.08$ ,  $SE = 0.08$ ,  $p =$   
434 0.371).

435 To understand the second interaction effect, we performed similar regression analyses  
436 using trials in the Good and Bad context separately. For each *post-hoc* regression analysis, we  
437 incorporated Group and Audience, along with their interaction as the fixed-effect predictors  
438 while controlling for the effect of the payoff for participants and associations in these analyses  
439 (same below for analyses on decision time). We only observed a strong main effect of Group  
440 ( $\chi^2(1) = 5.05$ ,  $p = 0.025$ ) and a Group  $\times$  Audience interaction effect in the Bad context ( $\chi^2(1) =$   
441 4.04,  $p = 0.044$ ), which was mainly driven by a drastically enhanced probability of behaving  
442 morally in the ASD group (vs. HC) when deciding privately (OR = 64.25,  $b = 4.16$ ,  $SE = 1.53$ ,  $p$   
443 = 0.006). Neither of these effects was significant in the Good context ( $ps > 0.12$ ; see **Table 2** for  
444 details of regression outputs).

445

#### 446 **ASD patients over-evaluate the immoral gains for both themselves and the bad cause**

447 We developed 8 models with different utility functions characterizing participants'  
448 choices in ASD and HC groups separately. Model estimation and comparison was performed  
449 with a Hierarchical Bayesian Analysis (HBA) approach (Gelman et al., 2014) via the "hBayesDM"  
450 package (see Methods for details). R-hat values of all estimated parameters of all models are  
451 close to 1.0 (i.e., smaller than 1.06 in the worst case), which showed sufficient convergence of  
452 the MCMC chains (Gelman and Rubin, 1992). Hierarchical Bayesian model comparison showed  
453 that model 4 (see below for the utility function) has the lowest leave-one-out information criterion



454 (LOOIC) scores, indicating that it fits to the current dataset the best compared with other  
 455 competitive models (see **Figure 3A**).

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

456 Here, SV denotes the subjective value of the given trial depending on the specific choice  
 457 made by the participant. Ms and Mo represent the payoff (gain or loss) for oneself and each  
 458 association respectively. Established on the basis of a donation task (Lopez-Persem et al.,  
 459 2017), the winning model assumed that people weighed their own payoff (measured by  $\alpha$ :  $\alpha_{good}$ ,  
 460  $\alpha_{bad}$ ) and the benefits for associations (measured by  $\beta$ :  $\beta_{good}$ ,  $\beta_{bad}$ ) separately given the moral  
 461 contexts involved in the decisions.  $\theta$  measured the audience effect, which was modulated by an  
 462 indicator function  $q$  (0 for private, 1 for public; see Methods for details). Posterior predictive  
 463 check further confirmed that the simulated choice behaviors in light of the parameter estimates  
 464 of the winning model can nicely capture the actual behaviors by showing a high correlation  
 465 between each other (i.e., for both HC and ASD group: Pearson's  $r_s > 0.99$ ,  $p_s < 0.001$ ; see  
 466 **Figure 3B**).

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

474 We first showed a significant group  $\times$  association interaction on both  $\alpha$  ( $b = 7.93$ ,  $SE =$   
 475  $3.91$ ,  $t(44) = 2.03$ ,  $p = 0.049$ ;  $b_z = 0.30$ ) and  $\beta$  ( $b = -10.88$ ,  $SE = 3.46$ ,  $t(44) = -3.14$ ,  $p = 0.003$ ;  $b_z$   
 476  $= -0.43$ ). Simple-effect analyses showed a significant decrease of decision weights on payoffs,

---

477 in ASD patients, for both themselves ( $\alpha_{\text{HC}}$  vs.  $\alpha_{\text{ASD}}$ :  $0.90 \pm 11.74$  vs.  $-9.27 \pm 10.13$ ,  $t(44) = -3.14$ ,  
478  $p = 0.003$ ;  $b_z = -0.45$ ) and the morally-bad cause ( $\beta_{\text{HC}}$  vs.  $\beta_{\text{ASD}}$ :  $-4.87 \pm 5.02$  vs.  $-11.52 \pm 8.67$ ,  
479  $t(44) = -2.96$ ,  $p = 0.005$ ;  $b_z = -0.41$ ). No between-group difference was observed in either  
480 parameter when participants weighed the trade-off between personal financial losses ( $\alpha_{\text{HC}}$  vs.  
481  $\alpha_{\text{ASD}}$ :  $11.18 \pm 6.41$  vs.  $8.89 \pm 9.09$ ,  $t(44) = -1.26$ ,  $p = 0.216$ ;  $b_z = -0.19$ ) and the donation to a  
482 charity ( $\beta_{\text{HC}}$  vs.  $\beta_{\text{ASD}}$ :  $4.38 \pm 4.04$  vs.  $6.78 \pm 7.83$ ,  $t(44) = 1.56$ ,  $p = 0.126$ ;  $b_z = 0.24$ ; see **Figure**  
483 **4A**). Notably, the correlation between  $\alpha$  and  $\beta$  was not significant across moral contexts in either  
484 group (ASD group: Good context:  $r = -0.177$ ,  $p = 0.469$ ; Bad context:  $r = -0.242$ ,  $p = 0.319$ ; HC  
485 group: Good context:  $r = -0.018$ ,  $p = 0.928$ ; Bad context:  $r = -0.003$ ,  $p = 0.989$ ; see **Figure 4B**).  
486 This indicates that participants value payoffs for oneself and the causes (associations)  
487 independently. Consistent with the behavioral finding, we also observed a trend-to-significance  
488 for the between-group difference in  $\theta$ , namely, that ASD patients exhibited a reduced audience  
489 effect compared to HC participants during moral decision-making ( $\theta_{\text{HC}}$  vs.  $\theta_{\text{ASD}}$ :  $0.39 \pm 0.67$  vs.  
490  $0.17 \pm 0.12$ ,  $t(44) = -1.80$ ,  $p = 0.080$ ,  $b_z = -0.27$ ).

491

#### 492 **ASD patients do not differ from HC in decision time in either moral context**

493 Mixed-effect linear regression on log-transformed decision time showed a significant  
494 three-way interaction between Group, Audience, and Moral Context ( $F(1,11769) = 6.02$ ,  $p =$   
495  $0.014$ ), along with a Group  $\times$  Moral Context interaction effect ( $F(1,11769) = 100.20$ ,  $p < 0.001$ )  
496 and a main effect of Moral Context ( $F(1,11772) = 299.76$ ,  $p < 0.001$ ) after controlling for the  
497 effect of specific choices ( $F(1,11804) = 3.76$ ,  $p = 0.052$ ; see **Figure 5**). Splitting the dataset  
498 according to Moral Context, *post-hoc* analyses revealed a significant Group  $\times$  Audience  
499 interaction effect when participants decided whether to serve a good cause at a personal cost  
500 ( $F(1,5860) = 4.28$ ,  $p = 0.039$ ) and a trend-to-significant interaction effect in the Bad context  
501 ( $F(1,5859) = 3.76$ ,  $p = 0.053$ ). However, neither the main effect of Group in the Good (ASD:

---

502 1676.5  $\pm$  527.7 ms; HC: 1490.0  $\pm$  399.5 ms;  $F(1,44) = 0.51$ ,  $p = 0.479$ ) nor the Bad context  
503 (ASD: 1525.7  $\pm$  828.1 ms; HC: 1445.5  $\pm$  500.9 ms;  $F(1,44) = 0.17$ ,  $p = 0.682$ ) was significant.  
504 The interaction effect in the Good context was driven by a slightly larger difference in decision  
505 time between groups when they made decisions in public (ASD: 1709.5  $\pm$  558.8 ms; HC: 1467.9  
506  $\pm$  379.3 ms) as compared to private (ASD: 1645.5  $\pm$  526.0 ms; HC: 1514.1  $\pm$  448.9 ms).  
507 However, neither of these between-group differences was statistically significant (public:  $b =$   
508 0.09,  $SE = 0.09$ ,  $t(44) = 0.98$ ,  $p = 0.334$ ,  $b_z = 0.29$ ; private:  $b = 0.04$ ,  $SE = 0.10$ ,  $t(44) = 0.42$ ,  $p =$   
509 0.678,  $b_z = 0.13$ ; see **Table 3** for details of regression output).

510

## 511 **Imaging Results**

### 512 *Decreased neural representation of moral contexts in the rTPJ of ASD patients*

513 To examine how the decision-related neural patterns differ in representing information  
514 contributing to the value computation and final decisions between ASD patients and HC  
515 participants, we performed a within-subject RSA (see **Figure 6** for the illustration of RSA  
516 procedure). Given our hypotheses, we focused our analysis on the rTPJ. To avoid bias on  
517 results caused by ROI selection to the maximum degree, we defined the rTPJ in two different  
518 ways, either via a whole-brain parcellation based on meta-analytic functional coactivation of the  
519 Neurosynth database (i.e., the Parcellation-Based ROI) or via a coordinate-based manner given  
520 a recent meta-analysis on neural correlates of ToM (Schurz et al., 2014) (i.e., the Coordinate-  
521 Based ROI; see Methods for details).

522 Regardless of the ROI approach, we consistently found that compared with HC group,  
523 ASD patients only showed a reduced representation of the information of the identity of  
524 associations in the rTPJ (ASD vs. HC: the Parcellation-Based ROI: Spearman's  $\rho = 0.101 \pm$   
525 0.047 vs.  $0.150 \pm 0.071$ ;  $p_{\text{permutation}} = 0.013$ ; the Coordinate-Based ROI:  $0.066 \pm 0.036$  vs.  $0.119$

---

526  $\pm 0.070$ ;  $p_{\text{permutation}} = 0.006$ ). These significant differences held after ruling out the confounding  
527 effect of age. Importantly, such a between-group difference of similarity was not observed  
528 between the neural RDM in the rTPJ and other cognitive RDMs (the Parcellation-Based ROI:  
529  $p_{\text{Spermutation}} > 0.20$ ; the Coordinate-Based ROI:  $p_{\text{Spermutation}} > 0.38$ ) or between the neural RDM in  
530 the left TPJ and all the cognitive RDMs (the Parcellation-Based ROI:  $p_{\text{Spermutation}} > 0.17$ ; the  
531 Coordinate-Based ROI:  $p_{\text{Spermutation}} > 0.30$ ; see **Figure 7**; also see **Table 4** for details). *Post-hoc*  
532  $2$  (group)  $\times$   $4$  (cognitive RDM) mixed ANOVA on the Fisher *r*-to-*z* transformed Spearman's rho  
533 revealed a strong interaction between group and cognitive RDM only in rTPJ (the Parcellation-  
534 Based ROI:  $F(3,126) = 6.09$ ,  $p < 0.001$ ; the Coordinate-Based ROI:  $F(3,126) = 8.37$ ,  $p < 0.001$ )  
535 but not in lTPJ (the Parcellation-Based ROI:  $F(3,126) = 0.65$ ,  $p = 0.585$ ; the Coordinate-Based  
536 ROI:  $F(3,126) = 0.42$ ,  $p = 0.743$ ) after controlling for the age difference, which further confirmed  
537 that the reduced ability to represent the information of moral context in ASD patients was  
538 uniquely reflected in rTPJ. Finally, to further examine the robustness of the above findings, we  
539 also applied the above analyses using all 256 trials, which did not affect the results (see **Figure**  
540 **8**; also see **Table 5** for details).

#### 541 *Univariate Results in rTPJ*

542 We first investigated whether the neural audience effect in rTPJ (i.e., Public > Private) in  
543 healthy controls reported in Qu et al. (2019) could be replicated in the present study. The  
544 results showed that the rTPJ activity was not significantly higher in the Public (vs. Private)  
545 condition (no voxel survived under a threshold of  $p < 0.005$  uncorrected at the voxel-level with  $k$   
546  $= 10$ , in either rTPJ mask; see **Figure 9A**). One possibility could be that the neural audience  
547 effect of rTPJ was modulated by large individual differences in the behavioral audience effect  
548 across individuals, which blurred the main effect. To test this possibility, we extracted the mean  
549 activity (contrast value) of the rTPJ from each condition, and then computed a neural index of

---

550 audience effect for each individual (i.e.,  $0.5 * [(Public_{Good} + Public_{Bad}) - (Private_{Good} + Private_{Bad})]$ ).

551 We also defined a behavioral index of audience effect on the proportion of moral choice, which

552 was calculated with the same equation. Results showed that the *Pearson* correlation between

553 these two indices was not significant (the parcellation-based ROI:  $r(24) = 0.02$ ,  $p = 0.914$ ; the

554 coordinate-based ROI:  $r(24) = -0.06$ ,  $p = 0.761$ ; see **Figure 9B**). Furthermore, the between-

555 group comparison did not reveal a significant result in the audience effect in rTPJ (i.e., no voxel

556 survived under the threshold mentioned above; see **Figure 10**). Besides, no significant

557 difference in the neural activity was observed in the rTPJ between the Good and Bad context in

558 the HC group or between two groups (i.e., no voxel survived under the threshold mentioned

559 above). For the completeness of the analyses, we also applied the same analyses to ITPJ,

560 yielding similar results (see **Figure 9 and 10**; also see **Table 6** for the whole-brain results under

561 a liberal threshold).

---

562 **Discussion**

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

573           First, ASD patients, unlike healthy controls, blurred the distinction between private and  
574 public conditions while making moral decisions. This finding not only coheres with the ToM  
575 deficit hypothesis of ASD patients (Baron-Cohen et al., 1985; Baron-Cohen, 2001) but also  
576 agrees with previous findings using a trade-off between suffering personal losses and donating  
577 to a good cause (Izuma et al., 2011). Moreover, it extends the unawareness of social reputation  
578 in autism to include an immoral context where individuals are confronted with a moral conflict  
579 between personal profits and a cost brought by benefiting an immoral cause. This first finding  
580 confirms that ASD patients seem unable to take into account their social reputation while  
581 making (im)moral choices consistently across contexts (Izuma et al., 2011).

582           Second, a robust behavioral difference between ASD patients and healthy controls was  
583 found in specifically one moral context. ASD patients generally refused more offers in the Bad  
584 context that could have earned extra money for themselves but which resulted in an immoral  
585 consequence. No similar between-group difference was observed in the Good context. Note

---

586 that decision difficulty cannot explain these behavioral effects because no decision time  
587 difference was observed between the two groups. Furthermore, this effect cannot be attributed  
588 to their greater (dis)like for the morally-bad cause because there was no significant between-  
589 group difference on subjective ratings.

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

---

610 Another possible explanatory factor of ASD participants' tendency to make overly-moral  
611 decisions in the Bad context is behavioral rigidity, a core symptom for clinical diagnosis of ASD  
612 (APA, 2013). Previous studies have revealed that compared with healthy controls, individuals  
613 with ASD were more likely to show repetitive behaviors in a variety of cognitive tasks (D'Cruz et  
614 al., 2013; Watanabe et al., 2019). Hence, it is possible that behavioral rigidity, at least to some  
615 extent, is a more general mechanism that contributes to the overly-moral behaviors in the Bad  
616 context (i.e., rejecting over 85% of the trials). Nonetheless, this explanation should be treated  
617 with caution because it seems not to account well for the behaviors of ASD patients in the Good  
618 context, where they behaved in a comparatively more flexible fashion (i.e., accepting around 60%  
619 of the trials).

620 At the brain level, we performed within-subject RSA to examine how different types of  
621 information (social reputation, moral contexts, payoffs for each party), that contribute to the final  
622 decision, were represented in the rTPJ, and how distinct rTPJ representations distinguish ASD  
623 patients from healthy controls. Compared with the traditional univariate approach, RSA takes  
624 advantage of neural patterns from multiple voxels, and proves to be more sensitive to subtle  
625 experimental effects that might be masked by the averaged local neural responses (Norman et  
626 al., 2006; Hebart and Baker, 2018). RSA is also considered to be more informative, because it  
627 takes into account the variability within multi-voxel patterns (Kriegeskorte et al., 2008; Popal et  
628 al., 2020). We observed a reduced association (representation similarity) in ASD patients (vs.  
629 healthy controls) between the trial-by-trial multivariate rTPJ patterns and the information  
630 structure unique to the moral contexts, despite that, such a representation in rTPJ is present in  
631 both groups. The representations of other types of information (i.e., social reputation and  
632 payoffs for each party) did not differ between groups. Together with a much higher rejection rate,  
633 as well as atypical weights on payoffs in the bad context, this RSA finding provides a neural  
634 account for previous findings that autistic individuals are inclined to judge moral culpability more



---

635 severely than HC on the basis of its consequences. This distinguishes ASD patients from HC,  
636 who prioritize intentions to guide their moral judgments (Fadda et al., 2016; Salvano-Pardieu et  
637 al., 2016; Bellesi et al., 2018). Notably, our results showed the group-difference in  
638 representational similarity was only detected in rTPJ but not in ITPJ, further indicating a unique  
639 role of rTPJ in specifically representing information concerning moral contexts.

640       Regarding the rTPJ's function, our RSA finding is consistent with a recent TMS study in  
641 healthy volunteers that revealed a context-sensitive moral role of rTPJ in signaling moral  
642 conflicts between personal benefits and moral values (Obeso et al., 2018). That study  
643 evidenced an asymmetrical TMS effect of rTPJ on moral behaviors depending on the moral  
644 context. Specifically, healthy participants under rTPJ stimulation were more altruistic such that  
645 they accepted more offers of donating to a charity at a personal cost regardless of donation  
646 amounts, whereas rTPJ disruption inhibited participants from accepting offers to earn morally-  
647 tainted money only when benefits to the bad cause were large. Building upon this finding, the  
648 present study provides further evidence using a different approach to reveal that rTPJ is  
649 critically involved in representing the moral contexts that flexibly modulate the trade-off between  
650 personal benefits and other's welfare during decision-making, which extends our understanding  
651 of the rTPJ function.

652       Notably, our univariate fMRI results did not reveal a neural audience effect in rTPJ in the  
653 healthy controls as was initially expected. Although previous studies provided evidence (Izuma,  
654 2012; Qu et al., 2019) suggesting that TPJ is involved in social reputation, negative evidence  
655 also exists. For instance, a recent transcranial magnetic stimulation (TMS) study using a similar  
656 experimental paradigm has shown that disrupting rTPJ (vs. sham) does not influence the  
657 audience effect on moral decisions in healthy individuals (Obeso et al., 2018). In addition, two  
658 earlier fMRI studies failed to find an increased activation of rTPJ in response to the presence (vs.  
659 absence) of observers while healthy participants made charitable decisions (Izuma et al., 2010b)

---

660 or social evaluation (Izuma et al., 2010a). However, it is also worth noting that non-significant  
661 results do not necessarily reflect a true null effect (Makin and Xivry, 2019). Also, our RSA result  
662 suggests that multi-voxel patterns of rTPJ represent the information of social reputation in  
663 healthy controls. Further studies are needed to clarify whether and how rTPJ plays a role in  
664 reputation-based decision-making.

665 Intriguingly, we did not observe a between-group difference of rTPJ in representing  
666 information about social reputation, although, as expected, a small-but-significant effect of social  
667 reputation on moral behaviors was observed only in healthy controls rather than ASD patients.  
668 At first glance, this finding may seem at odds with the well-established role of the rTPJ in  
669 mentalizing (and relevant social abilities) in both healthy participants (Hampton et al., 2008;  
670 Young et al., 2010; Carter et al., 2012; Morishima et al., 2012; Schurz et al., 2014; Hutcherson  
671 et al., 2015; Strombach et al., 2015; Hill et al., 2017; Hu et al., 2018; Qu et al., 2019) and ASD  
672 populations (Kana et al., 2009; Lombardo et al., 2011; Koster-Hale et al., 2013). These previous  
673 findings indicate that the deficiency of ToM ability, reflected by the dysfunction of rTPJ,  
674 determines the anomaly in moral behaviors in autistic cohorts. However, it should be noted that  
675 evidence also exists, revealing that ASD patients may preserve some degree of ToM ability to  
676 guide their intent-based moral judgments. For instance, one study showed that autistic adults  
677 not only exhibit performance comparable to healthy controls in a false belief task but also report  
678 similar moral permissibility when judging intended harms with neutral outcomes (Moran et al.,  
679 2011). Another study even reported an increased sensitivity to intention during moral judgment  
680 in Asperger syndrome compared with healthy controls (Channon et al., 2011). Consistent with  
681 these studies, our RSA results also suggest that the ability to represent the information of social  
682 reputation in rTPJ is partially intact in ASD patients. These findings indicate that the ability to  
683 infer and base moral judgments on intentionality may be still present in ASD individuals, and  
684 potentially explains why we did not observe a between-group difference of rTPJ in representing

---

685 social reputation in our task. It has also been proposed that the method of inferring intentionality  
686 differs between autistic and neurotypical participants (Dempsey et al., 2019). Here, a reduced  
687 rTPJ representation similarity in ASD, unique to the moral context, explains that patients  
688 excessively consider the negative consequences of an immoral action. This may block further  
689 recruitment of the intent-based system and thus lead to a failure to consider social reputation  
690 when making choices. Future studies may consider adopting tasks that involve both moral  
691 judgment and decision-making and implement non-invasive brain stimulation methods to target  
692 the rTPJ of ASD patients to provide causal evidence for this possibility.

693         Despite the strengths of this study, there are two potential limitations. First, the sample  
694 size is relatively small for the ASD group, which could have lowered the statistical power for the  
695 fMRI data analyses. Second, our sample has a relatively wide age range that covers the  
696 transition period from adolescence to early adulthood, during which time changes in  
697 sociocognitive processes and moral cognition continue to occur (Eisenberg and Morris, 2004;  
698 Blakemore and Mills, 2014; Kilford et al., 2016). Evidence indicates that mentalizing ability is still  
699 undergoing development in late adolescence (Dumontheil et al., 2010). More relevantly,  
700 previous studies have shown a distinct pattern in adolescents (vs. adults) for prosocial  
701 behaviors (Padilla-Walker et al., 2018) or the susceptibility to the audience effect (Wolf et al.,  
702 2015). Importantly, these changes are considered to be crucially associated with the  
703 development of the social brain network in adolescence (Blakemore, 2008; Kilford et al., 2016).  
704 Taking TPJ as an example, evidence from brain imaging studies showed that both structural  
705 and functional features of this region vary during this transition period (Blakemore et al., 2007;  
706 Mills et al., 2014). Hence, the age-related heterogeneity of our sample may have had some  
707 impact on our results, although we controlled for age-related differences in our between-group  
708 analyses. Future studies with a larger sample or less age heterogeneity would allow more  
709 definite conclusions.

---

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

718 **References**

- 719 Ahn W-Y, Haines N, Zhang L (2017) Revealing neuro-computational mechanisms of  
720 reinforcement learning and decision-making with the hBayesDM package.  
721 *Computational Psychiatry* 1:24-57.
- 722 Ahn W-Y, Krawitz A, Kim W, Busemeyer JR, Brown JW (2011) A model-based fMRI analysis  
723 with hierarchical Bayesian parameter estimation. *Journal of neuroscience, psychology,  
724 and economics* 4:95.
- 725 APA (2013) *Diagnostic and statistical manual of mental disorders: DSM-5 (5th ed.)*.
- 726 Baron-Cohen S (2001) Theory of mind and autism: A review. *International review of research in  
727 mental retardation* 23:169-184.
- 728 Baron-Cohen S, Leslie AM, Frith U (1985) Does the autistic child have a “theory of mind”?  
729 *Cognition* 21:37-46.
- 730 Bates D, Maechler M, Bolker B (2013) lme4: Linear mixed-effects models using Eigen and  
731 Eigen. R package version 0.999999-0. 2012. URL: <http://CRAN.R-project.org/package=lme4>.
- 732 Bault N, Pelloux B, Fahrenfort JJ, Ridderinkhof KR, van Winda F (2015) Neural dynamics of  
733 social tie formation in economic decision-making. *Social cognitive and affective  
734 neuroscience* 10:877-884.
- 735 Bellesi G, Vyas K, Jameel L, Channon S (2018) Moral reasoning about everyday situations in  
736 adults with autism spectrum disorder. *Research in Autism Spectrum Disorders* 52:1-11.
- 737 Blakemore, Sarah-Jayne, Ouden D, Hanneke, Choudhury, Suparna, Frith, Chris (2007)  
738 Adolescent development of the neural circuitry for thinking about intentions. *Social  
739 Cognitive & Affective Neuroscience* 2:130-139.
- 740 Blakemore S-J (2008) The social brain in adolescence. *Nature Reviews Neuroscience* 9:267-277.
- 741 Blakemore S-J, Mills KL (2014) Is adolescence a sensitive period for sociocultural processing?  
742 *Annual Review of Psychology* 65:187-207.
- 743 Buon M, Dupoux E, Jacob P, Chaste P, Leboyer M, Zalla T (2013) The role of causal and  
744 intentional judgments in moral reasoning in individuals with high functioning autism.  
745 *Journal of autism and developmental disorders* 43:458-470.
- 746 Burnham KP, Anderson DR (2004) Multimodel inference: understanding AIC and BIC in model  
747 selection. *Sociological methods & research* 33:261-304.
- 748 Carter RM, Bowling DL, Reeck C, Huettel SA (2012) A distinct role of the temporal-parietal  
749 junction in predicting socially guided decisions. *Science* 337:109-111.
- 750 Channon S, Lagnado D, Fitzpatrick S, Drury H, Taylor I (2011) Judgments of cause and blame:  
751 Sensitivity to intentionality in Asperger’s syndrome. *Journal of autism and  
752 developmental disorders* 41:1534-1542.
- 753 Crockett MJ (2016) How Formal Models Can Illuminate Mechanisms of Moral Judgment and  
754 Decision Making. *Current Directions in Psychological Science* 25:85-90.
- 755 D’Cruz A-M, Ragozzino ME, Mosconi MW, Shrestha S, Cook EH, Sweeney JA (2013) Reduced  
756 behavioral flexibility in autism spectrum disorders. *Neuropsychology* 27:152-160.
- 757 Dempsey E, Moore C, Johnson S, Stewart S, Smith I (2019) Morality in autism spectrum  
758 disorder: A systematic review. *Development and psychopathology*:1-17.
- 759 Dumontheil I, Apperly IA, Blakemore S-J (2010) Online usage of theory of mind continues to  
760 develop in late adolescence. *Developmental Science* 13:331-338.
- 761 Eisenberg N, Morris AS (2004) Moral cognitions and prosocial responding in adolescence. In:  
762 *Handbook of Adolescent Psychology, 2nd Edition (Lerner RM, Steinberg L, eds), pp  
763 155-188. New York: Wiley.*

- 
- 764 Eklund A, Nichols TE, Knutsson H (2016) Cluster failure: Why fMRI inferences for spatial  
765 extent have inflated false-positive rates. *Proceedings of the National Academy of*  
766 *Sciences* 113:7900-7905.
- 767 Fadda R, Parisi M, Ferretti L, Saba G, Foscoliano M, Salvago A, Doneddu G (2016) Exploring  
768 the role of Theory of Mind in moral judgment: the case of children with autism spectrum  
769 disorder. *Frontiers in psychology* 7:523.
- 770 Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. *Quarterly*  
771 *Journal of Economics*:817-868.
- 772 Fox J, Weisberg S, Adler D, Bates D, Baud-Bovy G, Ellison S, Firth D, Friendly M, Gorjanc G,  
773 Graves S (2016) Package ‘car’.
- 774 Frith U, Frith C (2011) Reputation Management: In Autism, Generosity Is Its Own Reward.  
775 *Current Biology* 21:R994-995.
- 776 Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences.  
777 *Statistical Science*:457-472.
- 778 Gelman A, Carlin JB, Stern HS, Rubin DB (2014) *Bayesian data analysis*: Chapman &  
779 Hall/CRC Boca Raton, FL, USA.
- 780 Haley KJ, Fessler DMT (2005) Nobody's watching? Subtle cues affect generosity in an  
781 anonymous economic game. *Evolution and Human Behavior* 26:245-256.
- 782 Hampton AN, Bossaerts P, O'Doherty JP (2008) Neural correlates of mentalizing-related  
783 computations during strategic interactions in humans. *Proceedings of the National*  
784 *Academy of Sciences* 105:6741-6746.
- 785 Hebart MN, Baker CI (2018) Deconstructing multivariate decoding for the study of brain  
786 function. *Neuroimage* 180:4-18.
- 787 Hill CA, Suzuki S, Polania R, Moisa M, O'Doherty JP, Ruff CC (2017) A causal account of the  
788 brain network computations underlying strategic social behavior. *Nature Neuroscience*  
789 20:1142.
- 790 Hu Y, He L, Zhang L, Wolk T, Dreher JC, Weber B (2018) Spreading inequality: neural  
791 computations underlying paying-it-forward reciprocity. *Social Cognitive and Affective*  
792 *Neuroscience* 13:578-589.
- 793 Hutcherson C, Bushong B, Rangel A (2015) A Neurocomputational Model of Altruistic Choice  
794 and Its Implications. *Neuron* 87:451-462.
- 795 Izuma K (2012) The social neuroscience of reputation. *Neuroscience research* 72:283-288.
- 796 Izuma K, Saito DN, Sadato N (2008) Processing of social and monetary rewards in the human  
797 striatum. *Neuron* 58:284-294.
- 798 Izuma K, Saito DN, Sadato N (2010a) The roles of the medial prefrontal cortex and striatum in  
799 reputation processing. *Social Neuroscience* 5:133-147.
- 800 Izuma K, Saito DN, Sadato N (2010b) Processing of the incentive for social approval in the  
801 ventral striatum during charitable donation. *Journal of Cognitive Neuroscience* 22:621-  
802 631.
- 803 Izuma K, Matsumoto K, Camerer CF, Adolphs R (2011) Insensitivity to social reputation in  
804 autism. *Proceedings of the National Academy of Sciences* 108:17302-17307.
- 805 Kana RK, Keller TA, Cherkassky VL, Minshew NJ, Just MA (2009) Atypical frontal-posterior  
806 synchronization of Theory of Mind regions in autism during mental state attribution.  
807 *Social Neuroscience* 4:135-152.
- 808 Kilford EJ, Garrett E, Blakemore SJ (2016) The Development of Social Cognition in  
809 Adolescence: An Integrated Perspective. *Neuroence & Biobehavioral Reviews*:106-120.



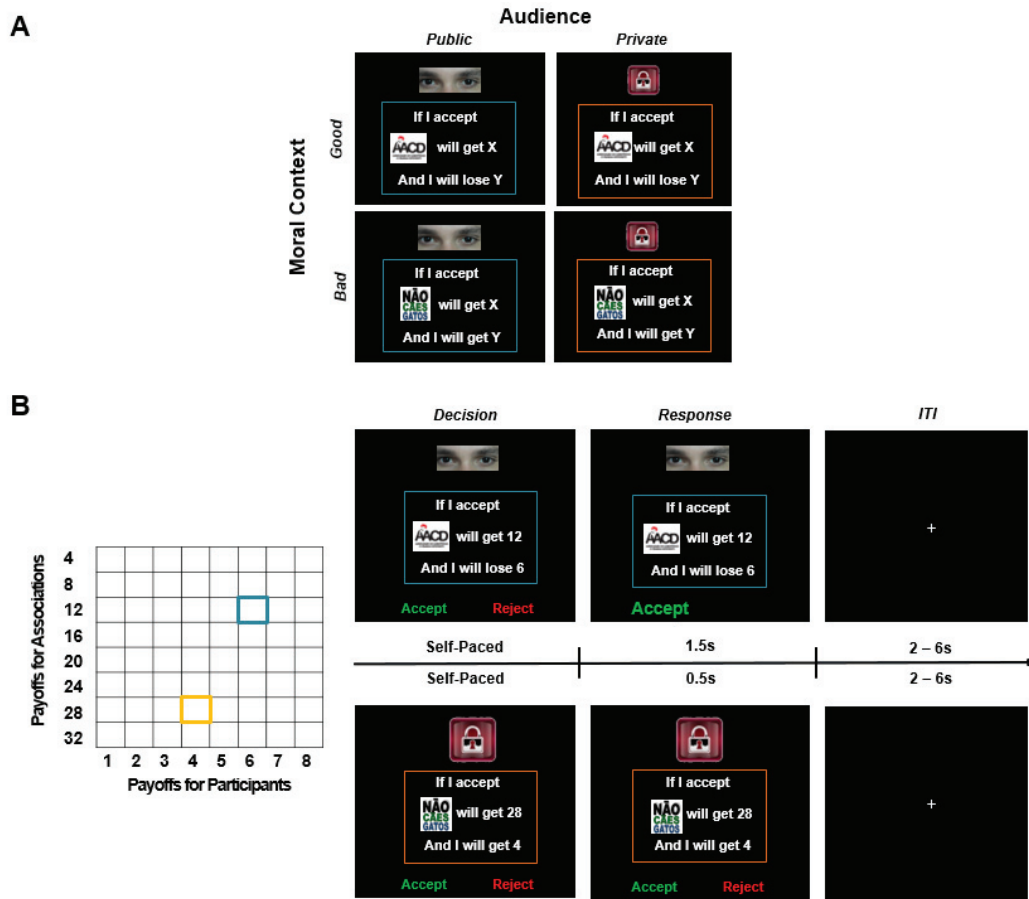
- 
- 810 Konovalov A, Hu J, Ruff CC (2018) Neurocomputational approaches to social behavior. *Current*  
811 *Opinion in Psychology*.
- 812 Koster-Hale J, Saxe R, Dungan J, Young LL (2013) Decoding moral judgments from neural  
813 representations of intentions. *Proceedings of the National Academy of Sciences*  
814 110:5648-5653.
- 815 Kriegeskorte N, Mur M, Bandettini PA (2008) Representational similarity analysis-connecting  
816 the branches of systems neuroscience. *Frontiers in Systems Neuroscience* 2:4.
- 817 Lombardo MV, Chakrabarti B, Bullmore ET, Baron-Cohen S, Consortium MA (2011)  
818 Specialization of right temporo-parietal junction for mentalizing and its relation to social  
819 impairments in autism. *Neuroimage* 56:1832-1838.
- 820 Lopez-Persem A, Rigoux L, Bourgeois-Gironde S, Daunizeau J, Pessiglione M (2017) Choose,  
821 rate or squeeze: Comparison of economic value functions elicited by different behavioral  
822 tasks. *PLoS computational biology* 13:e1005848.
- 823 Luke SG (2017) Evaluating significance in linear mixed-effects models in R. *Behavior Research*  
824 *Methods* 49:1494-1502.
- 825 Makin TR, Xivry JJOD (2019) Ten common statistical mistakes to watch out for when writing or  
826 reviewing a manuscript. *Elife Sciences* 8:e48175.
- 827 Margoni F, Surian L (2016) Mental state understanding and moral judgment in children with  
828 autistic spectrum disorder. *Frontiers in psychology* 7:1478.
- 829 Mills KL, Francois L, Clasen LS, Giedd JN, Sarah-Jayne B (2014) Developmental changes in the  
830 structure of the social brain in late childhood and adolescence. *Social Cognitive and*  
831 *Affective Neuroscience* 9:123-131.
- 832 Moran JM, Young LL, Saxe R, Lee SM, O'Young D, Mavros PL, Gabrieli JD (2011) Impaired  
833 theory of mind for moral judgment in high-functioning autism. *Proceedings of the*  
834 *National Academy of Sciences* 108:2688-2692.
- 835 Morishima Y, Schunk D, Bruhin A, Ruff CC, Fehr E (2012) Linking brain structure and  
836 activation in temporoparietal junction to explain the neurobiology of human altruism.  
837 *Neuron* 75:73-79.
- 838 Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: multi-voxel pattern  
839 analysis of fMRI data. *Trends in cognitive sciences* 10:424-430.
- 840 Obeso I, Moisa M, Ruff CC, Dreher J-C (2018) A causal role for right temporo-parietal junction  
841 in signaling moral conflict. *eLife* 7:e40671.
- 842 Padilla-Walker LM, Carlo G, Memmott-Elison MK (2018) Longitudinal Change in Adolescents'  
843 Prosocial Behavior Toward Strangers, Friends, and Family. *Journal of Research on*  
844 *Adolescence*.
- 845 Pantelis PC, Byrge L, Tyszka JM, Adolphs R, Kennedy DP (2015) A specific hypoactivation of  
846 right temporo-parietal junction/posterior superior temporal sulcus in response to socially  
847 awkward situations in autism. *Social cognitive and affective neuroscience* 10:1348-1356.
- 848 Popal HS, Olson IR, Wang Y (2020) A Guide to Representational Similarity Analysis for Social  
849 Neuroscience. *Social Cognitive and Affective Neuroscience*.
- 850 Qu C, Météreau E, Butera L, Villeval MC, Dreher J-C (2019) Neurocomputational mechanisms  
851 at play when weighing concerns for extrinsic rewards, moral values, and social image.  
852 *PLOS Biology* 17:e3000283.
- 853 Qu C, Hu Y, Tang Z, Derrington E, Dreher JC (2020) Neurocomputational mechanisms  
854 underlying immoral decisions benefiting self or others. *Social Cognitive and Affective*  
855 *Neuroscience* nsaa029.

- 
- 856 R Core Team (2014) R: A language and environment for statistical computing. In.  
857 Salvano-Pardieu V, Blanc R, Combalbert N, Pierratte A, Manktelow K, Maintier C, Lepeltier S,  
858 Gimenes G, Barthelemy C, Fontaine R (2016) Judgment of blame in teenagers with  
859 Asperger's syndrome. *Thinking & Reasoning* 22:251-273.
- 860 Schaafsma SM, Pfaff DW, Spunt RP, Adolphs R (2014) Deconstructing and reconstructing  
861 theory of mind. *Trends in Cognitive Sciences* 19:65-72.
- 862 Schaller UM, Biscaldi M, Fangmeier T, van Elst LT, Rauh R (2019) Intuitive Moral Reasoning  
863 in High-Functioning Autism Spectrum Disorder: A Matter of Social Schemas? *Journal of*  
864 *autism and developmental disorders* 49:1807-1824.
- 865 Schurz M, Radua J, Aichhorn M, Richlan F, Perner J (2014) Fractionating theory of mind: A  
866 meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral*  
867 *Reviews* 42:9-34.
- 868 Stan Development Team (2016) Stan: A C++ library for probability and sampling.
- 869 Strombach T, Weber B, Hangebrauk Z, Kenning P, Karipidis II, Tobler PN, Kalenscher T (2015)  
870 Social discounting involves modulation of neural value signals by temporoparietal  
871 junction. *Proceedings of the National Academy of Sciences* 112:1619-1624.
- 872 Tusche A, Böckler A, Kanske P, Trautwein F-M, Singer T (2016) Decoding the Charitable Brain:  
873 Empathy, Perspective Taking, and Attention Shifts Differentially Predict Altruistic  
874 Giving. *The Journal of Neuroscience* 36:4719-4732.
- 875 Watanabe T, Lawson RP, Walldén YSE, Rees G (2019) A neuroanatomical substrate linking  
876 perceptual stability to cognitive rigidity in autism. *The Journal of Neuroence*:2831-2818.
- 877 Wickham H (2016) *ggplot2: elegant graphics for data analysis*.
- 878 Wolf LK, Bazargani N, Kilford EJ, Dumontheil I, Blakemore SJ (2015) The audience effect in  
879 adolescence depends on who's looking over your shoulder. *Journal of Adolescence* 43:5-  
880 14.
- 881 Young L, Cushman F, Hauser M, Saxe R (2007) The neural basis of the interaction between  
882 theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*  
883 104:8235-8240.
- 884 Young L, Camprodon JA, Hauser M, Pascual-Leone A, Saxe R (2010) Disruption of the right  
885 temporoparietal junction with transcranial magnetic stimulation reduces the role of  
886 beliefs in moral judgments. *Proceedings of the National Academy of Sciences* 107:6753-  
887 6758.
- 888



889 **Figures**

890

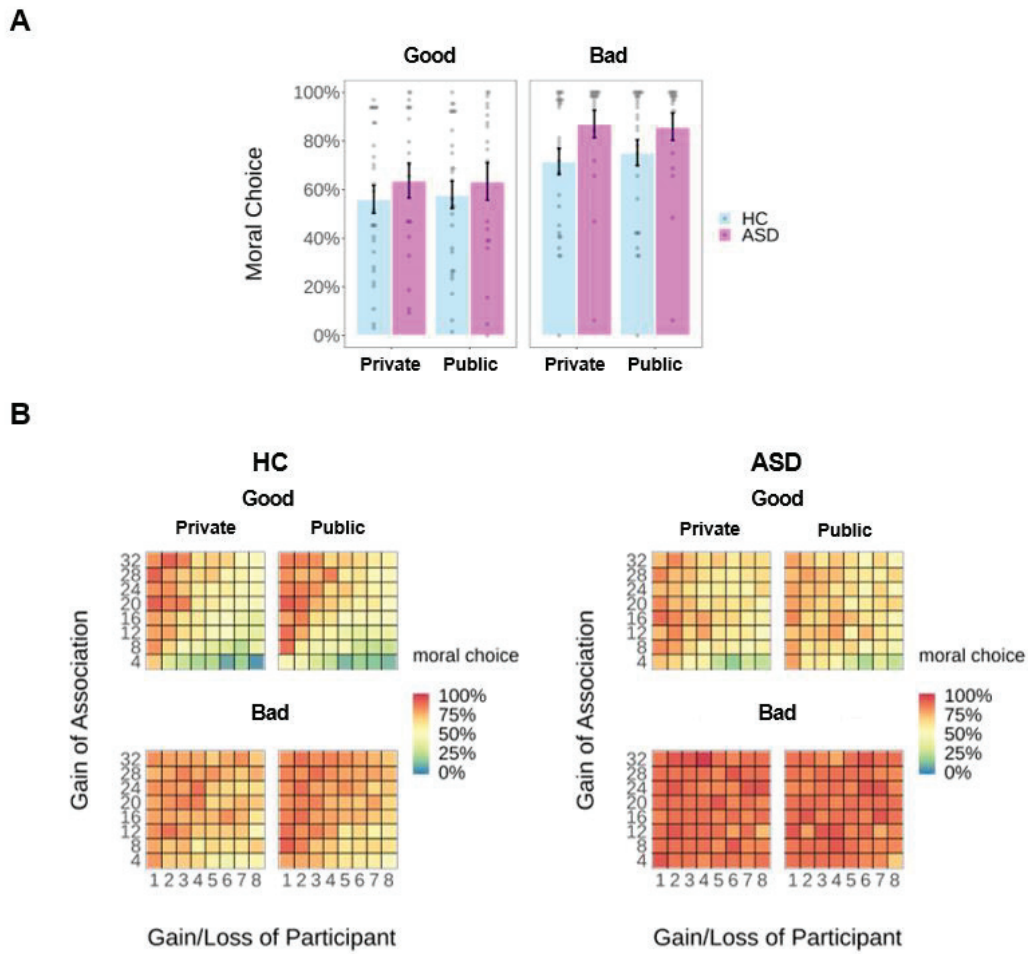


891

892 **Figure 1. Illustration of Experimental Design and Trial Procedure.** (A) We employed a 2 × 2  
 893 within-subject design by independently manipulating Audience (Private or Public) and Moral  
 894 Context (Good or Bad), which yielded four experimental conditions (i.e., Public<sub>Good</sub>, Public<sub>Bad</sub>,  
 895 Private<sub>Good</sub>, and Private<sub>Bad</sub>). The Public condition was indicated by the picture of “eyes”, and the  
 896 Private condition was indicated by the picture of a “lock”. The Good context involved a trade-off  
 897 between personal losses and benefits for a charity, whereas in the Bad context participants  
 898 traded personal benefits against benefits for a morally-bad cause. (B) Monetary payoffs (in

---

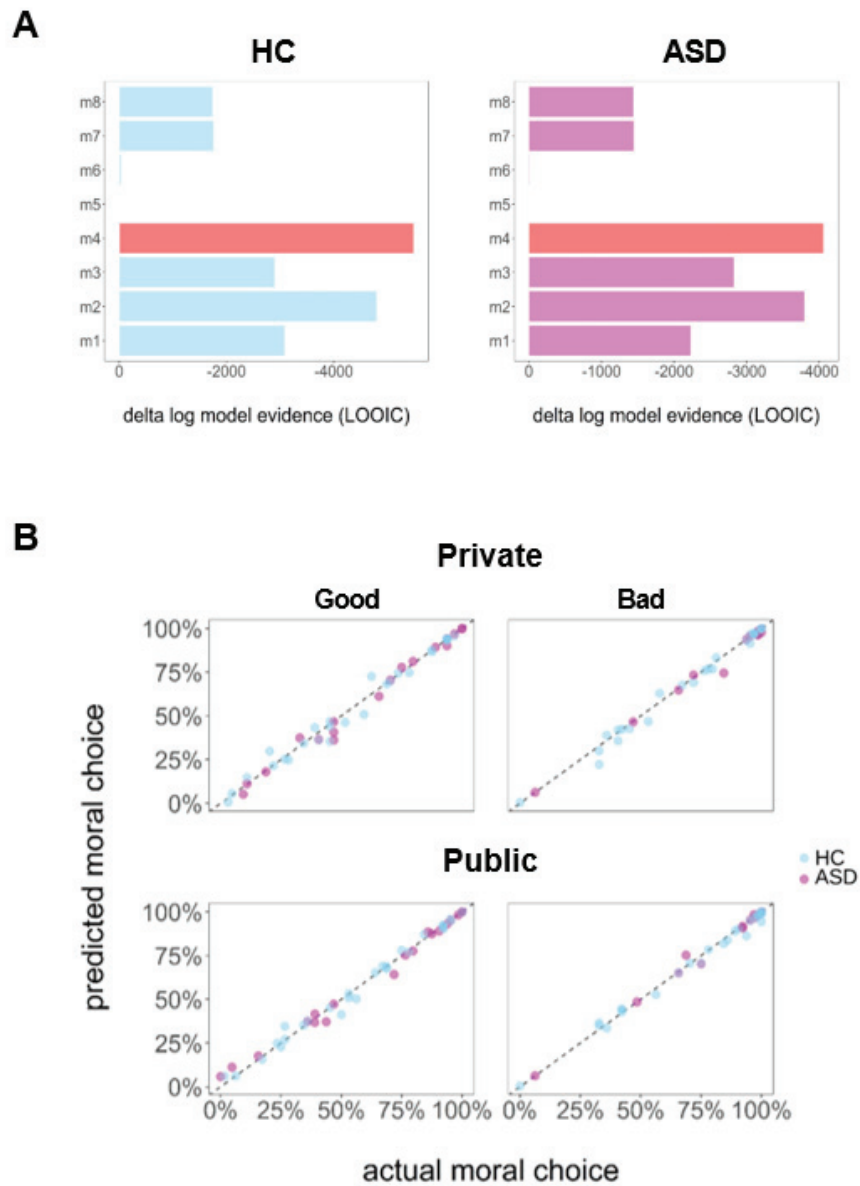
899 Brazilian Real) for participants (8 levels: from 1 to 8, in steps of 1) and the association (8 levels:  
900 from 4 to 32, in steps of 4) were orthogonally varied, yielding 64 unique offers for each condition.  
901 In the example trial (one for the Public<sub>Good</sub> and the other for the Private<sub>Bad</sub> condition), participants  
902 were presented with an offer and decided whether to accept or reject the offer with no time limit.  
903 If they accepted the offer, both parties involved (i.e., the participant and the association) might  
904 undergo the financial consequences as proposed. If they rejected the offer, neither party would  
905 profit. In the Private condition, once a response was made, the screen was unchanged for 0.5s  
906 to keep the chosen option private. In the Public condition, the chosen option was highlighted  
907 with a larger font and the non-chosen option disappeared, this lasted slightly longer (1.5s) to  
908 further emphasize the presence of a witness. Each trial was ended with an inter-trial interval (ITI)  
909 showing a jittered fixation (2.5 ~ 6.5s).



910

911 **Figure 2. Results of choice behavior.** (A) Rate of choosing the moral option as a function of  
 912 group (ASD or HC), reputation (Private or Public), and context (Good or Bad). (B) Heat map of  
 913 the mean proportion (%) of moral choices as a function of payoffs (monetary units, MU) for  
 914 participants and for associations in each experimental condition for each group. Each dot  
 915 represents the data of a single participant. Error bars represent the SEM; Abbreviation: HC:  
 916 healthy control, ASD: autism spectrum disorder.

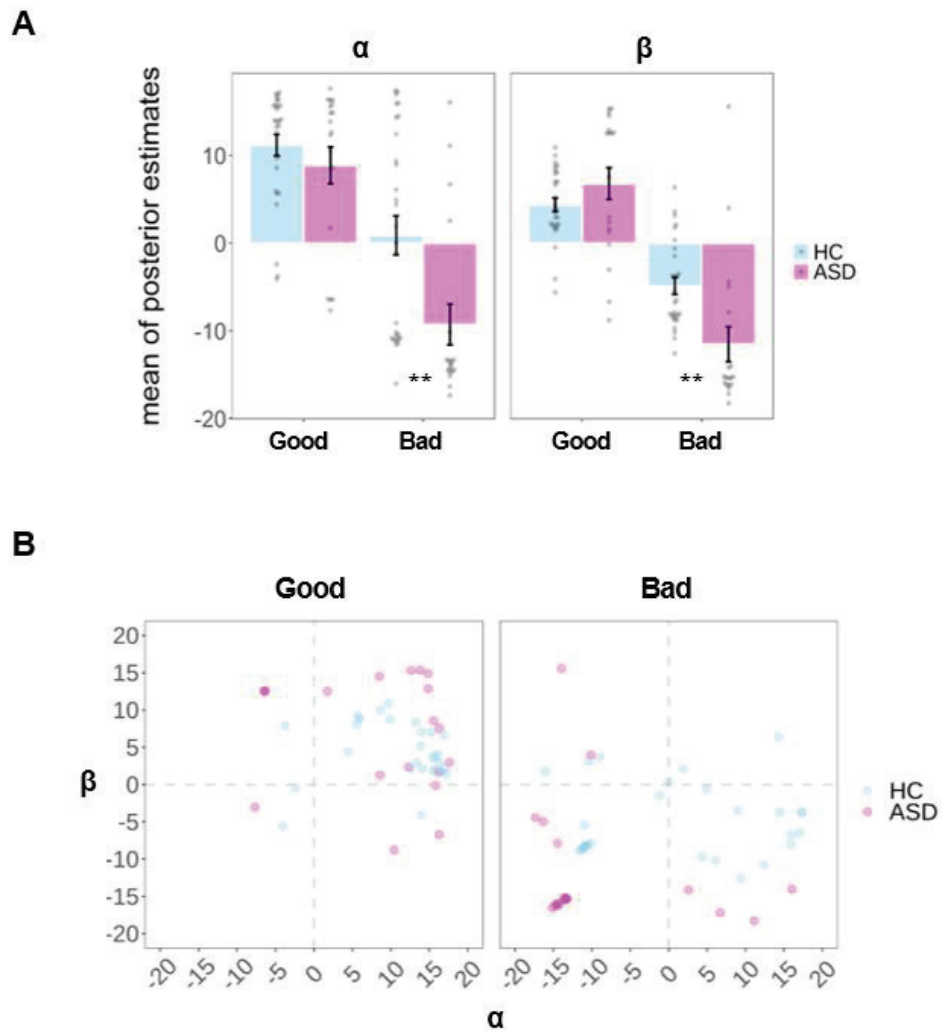
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935



936 **Figure 3. Model comparison and validation.** (A) Bayesian model evidence. Model evidence  
937 (relative to the model with the worst accuracy of out-of-sample prediction, i.e., model 5), clearly  
938 favors model 4 (m4). Lower (i.e., more negative) leave-one-out information criterion (LOOIC)  
939 scores indicate a better model. (B) Posterior predictive check of the winning model. Each dot

---

940 represents the data of a single participant. For each participant, we calculated the mean of the  
941 predicted proportion of moral choice (%; y axis) by averaging moral choices generated using the  
942 whole posterior distribution of estimated parameters specific to that participant based on the  
943 winning model. Regardless of experimental conditions, these dots almost fell on the diagonal,  
944 indicating that the winning model captured the actual behaviors of all participants in this task.  
945 Abbreviation: HC: healthy control, ASD: autism spectrum disorder.



946

947 **Figure 4. Results of parameter estimates.** (A) Group-level mean of individual-level posterior

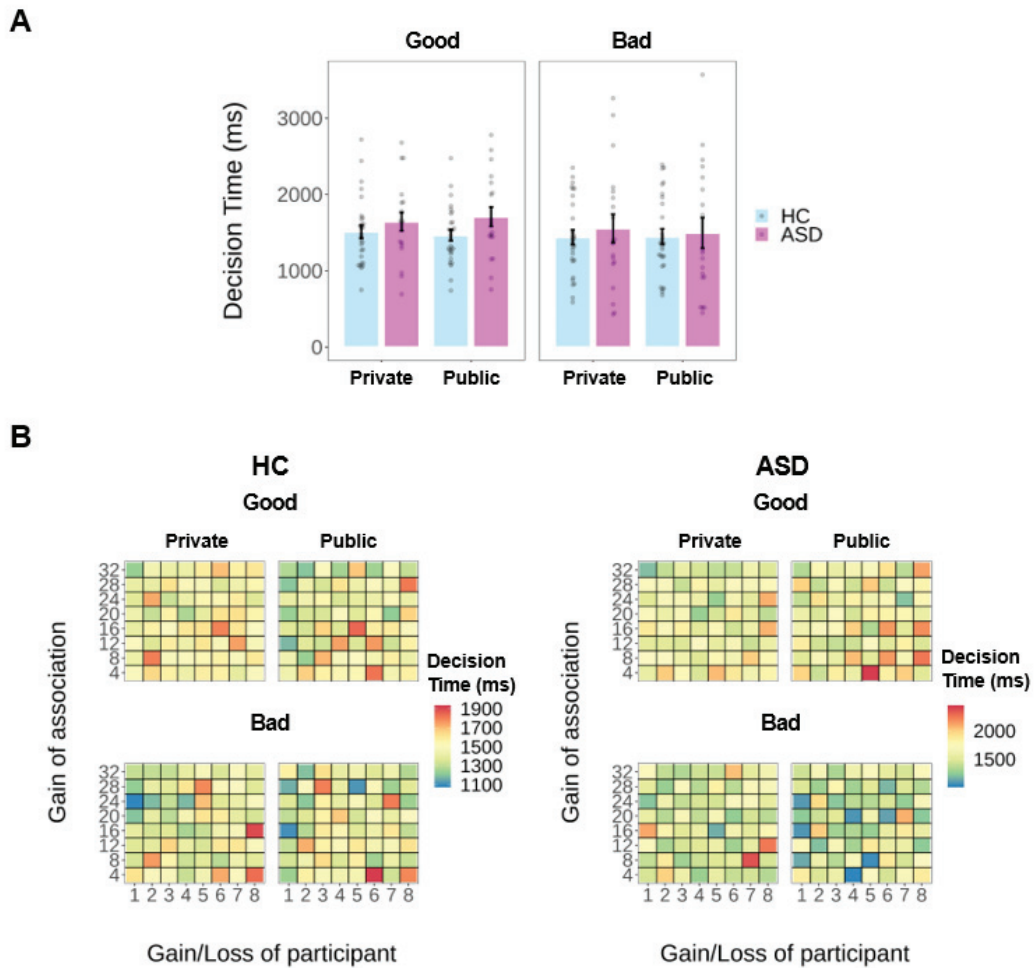
948 mean of  $\alpha$  and  $\beta$  across moral contexts (good or bad) derived from the winning model. (B)

949 Scatter plot of individual-level posterior mean of  $\alpha$  and  $\beta$  across moral contexts (Good or Bad) in

950 each group. Each dot represents the data of a single participant. Error bars represent the SEM;

951 significance: \*\* $p < 0.01$ , after controlling for the age difference between groups. Abbreviation:

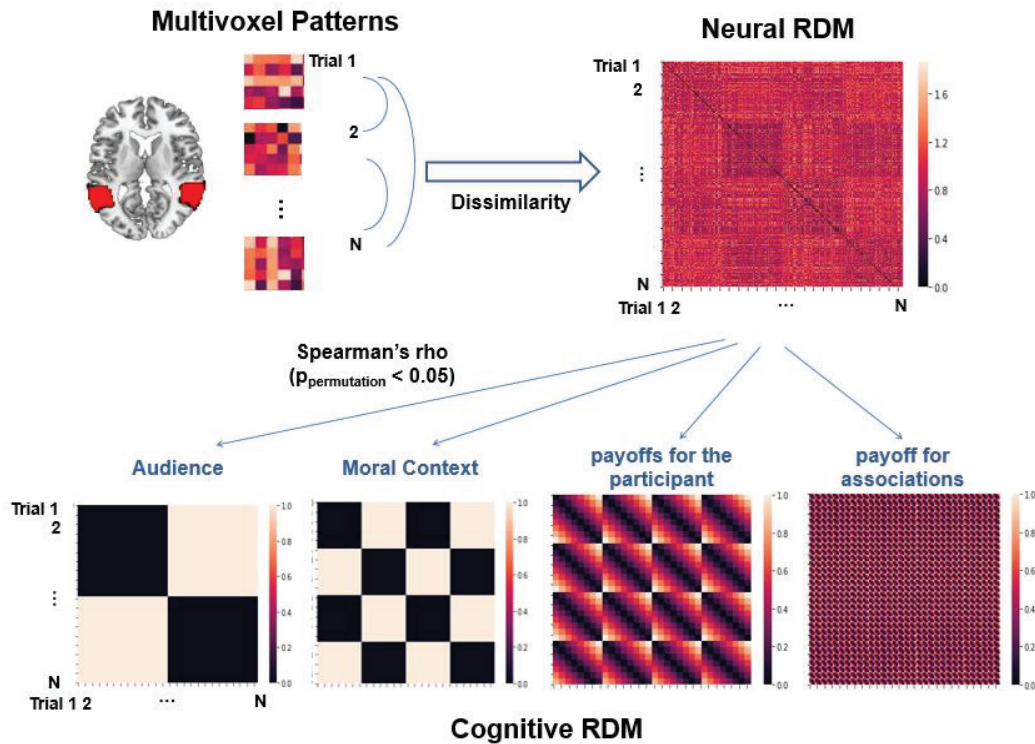
952 HC: healthy control, ASD: autism spectrum disorder.



953

954 **Figure 5. Results of decision time (in ms).** (A) Bar plot of the mean decision time as a  
 955 function of group (ASD or HC), reputation (Private or Public), and context (Good or Bad). (B)  
 956 Heat map of the mean decision time regardless of specific choices as a function of payoffs  
 957 (monetary units, MU) for participants and for associations in each experimental condition of  
 958 each group. Abbreviation: HC: healthy control, ASD: autism spectrum disorder.





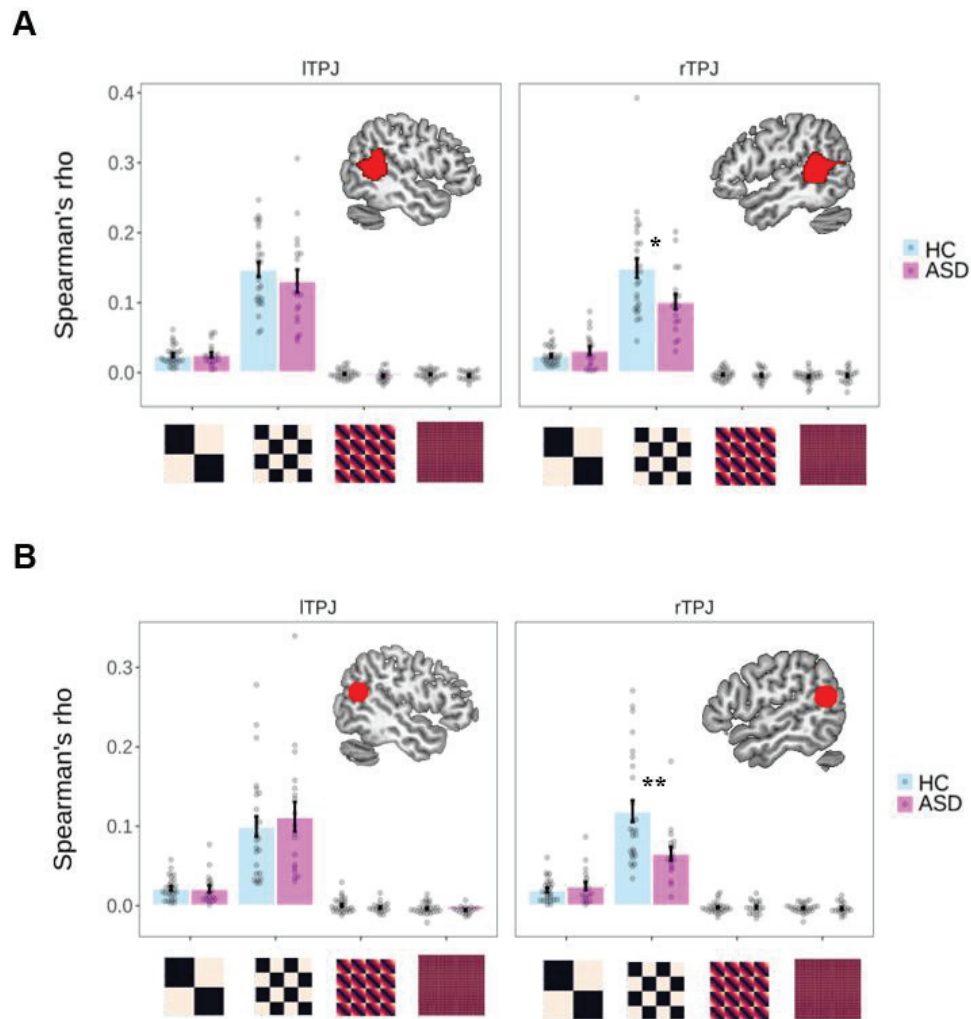
960

961 **Figure 6. Illustration of within-subject representational similarity analyses (RSA).** For  
 962 each individual, we first constructed a neural RDM measuring the correlational distances of  
 963 multi-voxel patterns of the decision-relevant neural activities within either left or right TPJ  
 964 between each pair of valid trials respectively. Next, we constructed four cognitive RDMs by  
 965 calculating the Euclidean distances between each pair of valid trials with respect to the following  
 966 information: 1) Audience (i.e., social reputation; Private or Public), 2) Moral Context (i.e., Good  
 967 or Bad), 3) payoffs for the participant, and 4) payoffs for associations. Notably, we sorted all  
 968 trials according to the order of Audience, Moral Context, payoff for the participant, and payoff for  
 969 associations to guarantee the information contained by both the neural and cognitive RDMs was  
 970 matched with each other. Then we performed the Spearman rank-ordered correlation between



---

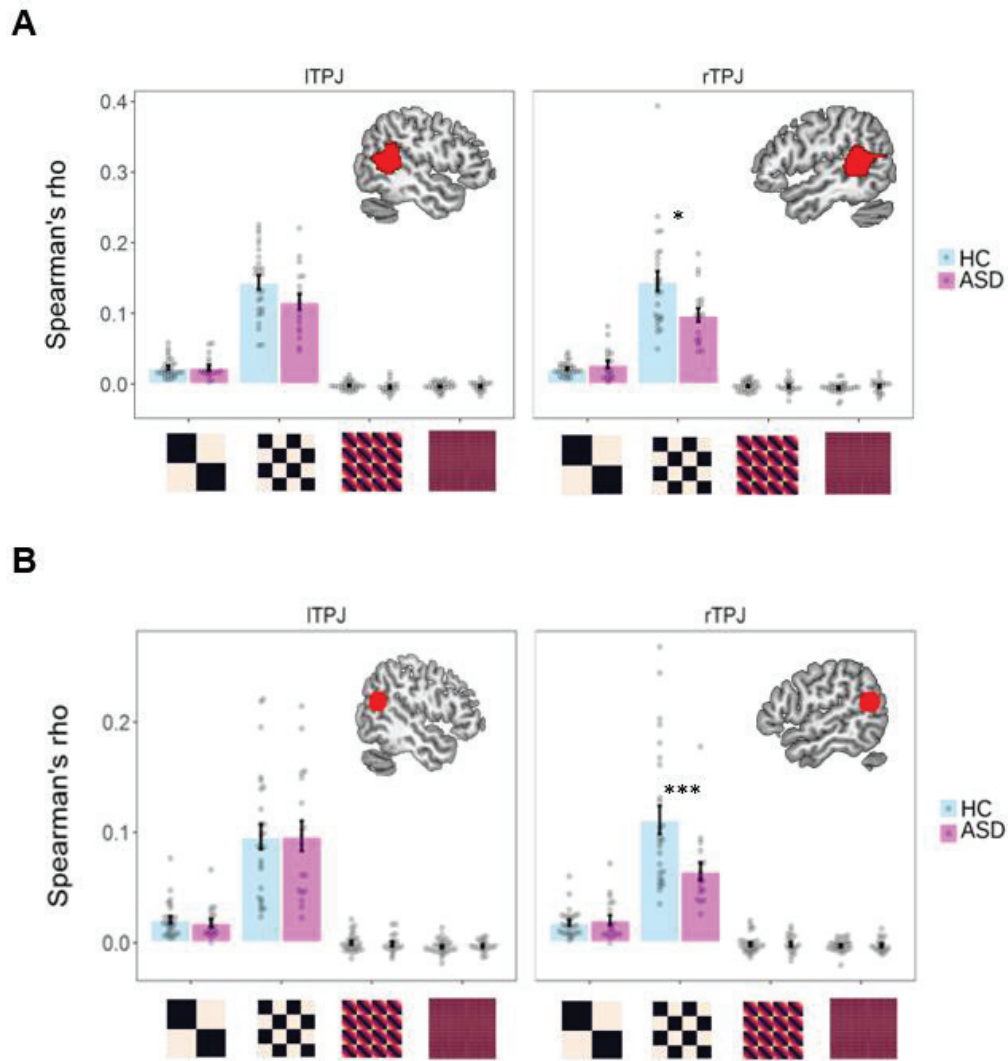
971 the neural and the cognitive RDMs. Finally, an independent two-sample permutation-based T-  
972 test was conducted to compare the between-group difference on the z-transformed Spearman's  
973 rho. Abbreviations: RDM: representational dissimilarity matrix.



974

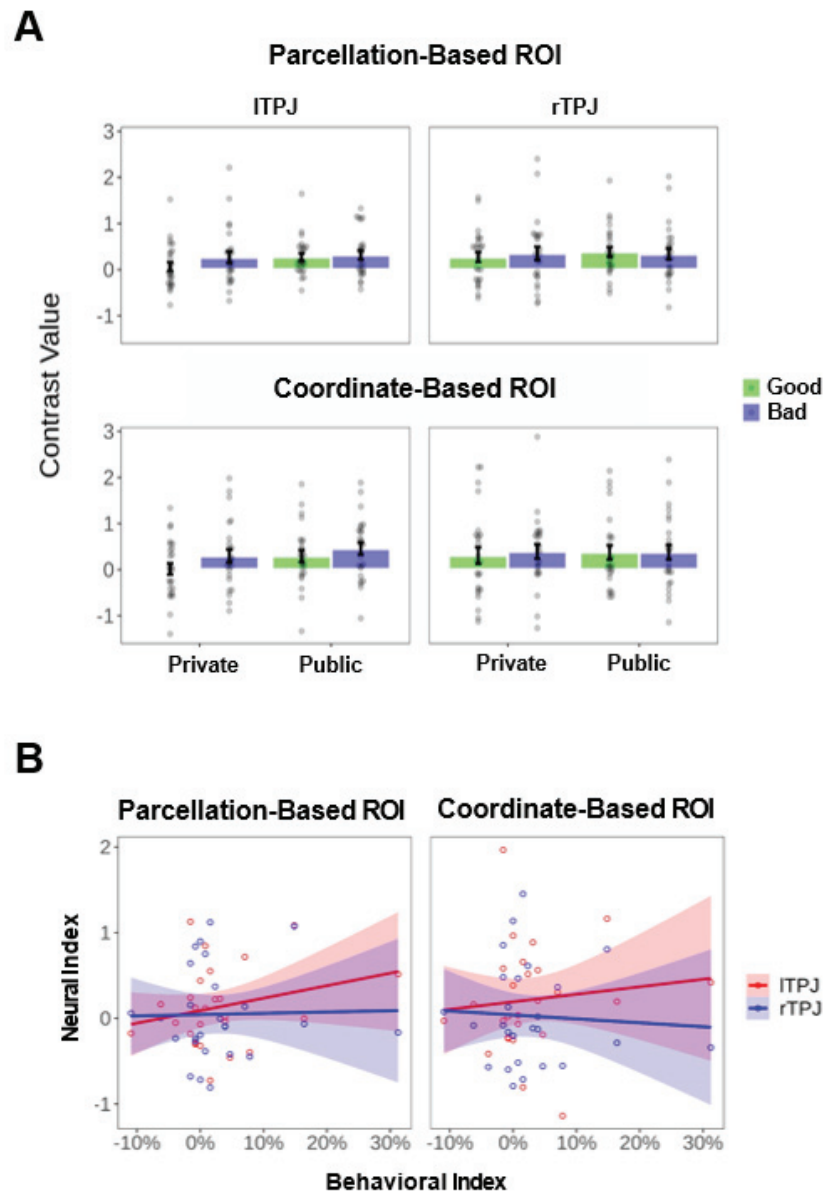
975

976 **Figure 7. Within-subject RSA results using (A) the Parcellation-Based ROI and (B) the**  
 977 **Coordinate-Based ROI of TPJ.** For each participant, we only adopted valid trials (see Methods  
 978 for details) in these analyses. Each dot represents the data of a single participant. Error bars  
 979 represent the SEM; significance: \* $p_{\text{permutation}} < 0.05$ , \*\* $p_{\text{permutation}} < 0.01$ , after controlling for the  
 980 age difference. Abbreviation: RSA: representational similarity analysis; TPJ: temporoparietal  
 981 junction; HC: healthy control, ASD: autism spectrum disorder.



982

983 **Figure 8. Robustness check of within-subject RSA results using (A) the Parcellation-**  
 984 **Based ROI and (B) the Coordinate-Based ROI of TPJ.** For each participant, we adopted all  
 985 256 trials in these analyses. Each dot represents the data of a single participant. Error bars  
 986 represent the SEM; significance: \* $p_{\text{permutation}} < 0.05$ , \*\*\* $p_{\text{permutation}} < 0.001$ , after controlling for the  
 987 age difference. Abbreviation: RSA: representational similarity analysis; TPJ: temporoparietal  
 988 junction; HC: healthy control, ASD: autism spectrum disorder.



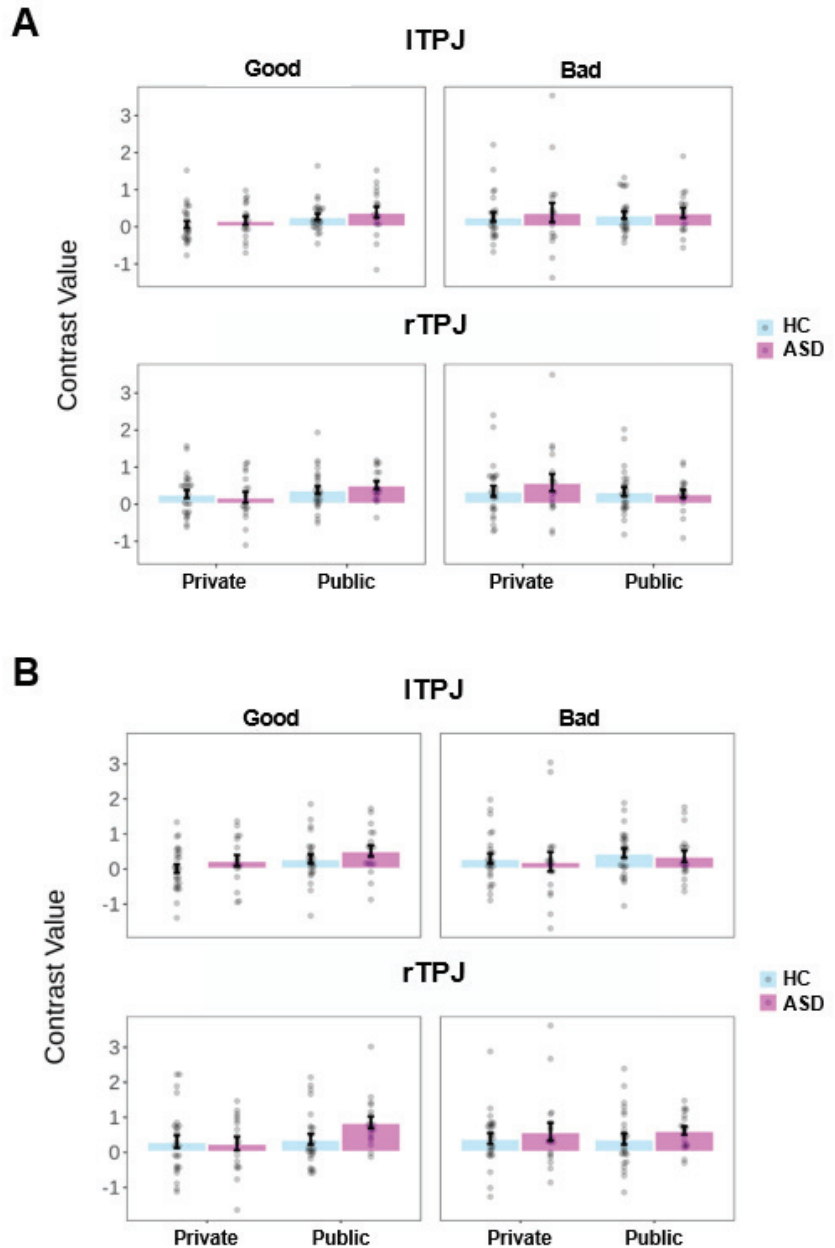
989

990 **Figure 9** Univariate results of TPJ in healthy controls. **(A)** Bar plot of TPJ signals. For  
 991 visualization, we extracted the mean activity (contrast value) of ITPJ and rTPJ from the  
 992 parcellation-based or coordinate-based mask as a function of reputation (Private or Public), and  
 993 context (Good or Bad). Each dot represents the data of a single participant. Error bars represent

---

994 the SEM. **(B) Relationship between neural audience effect in TPJ and behavioral audience**  
995 **effect across individuals.** Each dot represents the data of a single participant. Each line  
996 represents the linear fit. Shaded areas represent the 95% confidence interval. Abbreviation: ROI:  
997 region of interest; TPJ: temporoparietal junction.

998



999

1000 **Figure 10** Univariate results of TPJ in HC and ASD group using (A) the parcellation-based  
 1001 **mask** and (B) the coordinate-based mask. For visualization, we extracted the mean activity

---

1002 (contrast value) of ITPJ and rTPJ from the corresponding masks as a function of group (ASD or  
1003 HC), reputation (Private or Public), and context (Good or Bad). Each dot represents the data of  
1004 a single participant. Error bars represent the SEM.

---

1005 **Tables**

1006 **Table 1 Summary of clinical measures in two groups**

	ASD	HC
IQ: total <sup>a</sup>	100.0 ± 10.0	105.0 ± 8.9
IQ: verbal <sup>a</sup>	103.2 ± 9.9	103.2 ± 9.2
IQ: execution <sup>a</sup>	106.7 ± 12.4	106.7 ± 11.5
ADI-R: social	21.0 ± 5.2	
ADI-R: communication	14.0 ± 4.5	
ADI-R: repetitive	6.7 ± 1.7	

1007

1008 Note: <sup>a</sup> IQ was measured by Wechsler Intelligence Scale for Children (WISC); Data of 3 HC were

1009 missing. Abbreviations: IQ: intelligence quotient, ADI: autism diagnostic interview; HC: healthy

1010 control, ASD: autism spectrum disorder.



1011 **Table 2 Results of mixed-effect logistic regressions predicting moral choices**

	All	Good	Bad	Bad: Private	Bad: Public
	<i>b</i> (SE)	<i>b</i> (SE)	<i>b</i> (SE)	<i>b</i> (SE)	<i>b</i> (SE)
Intercept	0.63 (0.39)	0.54 (0.53)	2.64** (0.81)	2.57** (0.84)	3.26*** (0.88)
Group	0.86 (0.64)	1.31 (0.86)	3.41* (1.42)	4.16** (1.53)	2.32 (1.44)
Audience	0.11 (0.08)	0.15 (0.10)	0.27** (0.10)		
Moral context	0.95*** (0.08)				
Group × Audience	-0.17 (0.13)	-0.23 (0.16)	-0.44* (0.22)		
Group × Moral context	0.89*** (0.15)				
Audience × Moral context	0.09 (0.12)				
Group × Audience × Moral context	-0.11 (0.21)				
Payoff for oneself <sup>a,b</sup>		-0.99*** (0.04)	-0.46*** (0.05)	-0.39*** (0.07)	-0.56*** (0.07)
Payoff for association <sup>a,b</sup>		0.83*** (0.04)	0.33*** (0.05)	0.34*** (0.06)	0.35*** (0.07)
Age <sup>a</sup>	0.19 (0.32)	0.50 (0.43)	0.26 (0.70)	0.23 (0.70)	0.12 (0.73)
AIC	10501.0	4340.7	3148.7	1649.7	1551.1
BIC	10574.8	4394.1	3202.2	1685.6	1587.1
N (Observation)	11823	5912	5911	2948	2963
N (Participant)	47	47	47	47	47

1012

1013 Note: <sup>a</sup> We standardized these variables for the analyses.

1014 <sup>b</sup> These variables were added as covariates only when the regressor “Association” (and its interaction)

1015 was not in the regression model, as the regressor “payoff for oneself” qualitatively co-varied with

1016 “Association”, which might cause the collinear issue.

1017 Reference levels were set as follows: Group = healthy controls (HC), Audience = Private, Moral Context =

1018 Good. Table also shows goodness-of-fit statistics: AIC = Akaike Information Criterion, BIC = Bayesian

1019 Information Criterion. Significance: \*p < 0.05, \*\*p < 0.01, \*\*\*p < 0.001.

1020 **Table 3 Results of mixed-effect logistic regressions predicting log-transformed decision**

	All	Good	Good: Private	Good: Public	Bad
	<i>b</i> (SE)	<i>b</i> (SE)	<i>b</i> (SE)	<i>b</i> (SE)	<i>b</i> (SE)
Intercept	7.22 <sup>***</sup> (0.07)	7.19 <sup>***</sup> (0.06)	7.19 <sup>***</sup> (0.06)	7.18 <sup>***</sup> (0.06)	7.19 <sup>***</sup> (0.10)
Group	0.07	0.04	0.04	0.09	-0.04

1021 **time (in ms)**

	(0.11)	(0.09)	(0.10)	(0.09)	(0.14)
Audience	-0.01	-0.02			0.01
	(0.02)	(0.01)			(0.01)
Moral context	-0.08***				
	(0.02)				
Group × Audience	0.04	0.04 <sup>†</sup>			-0.04 <sup>†</sup>
	(0.02)	(0.02)			(0.02)
Group × Moral context	-0.13***				
	(0.02)				
Audience × Moral context	0.02				
	(0.02)				
Group × Audience × Moral context	-0.08 <sup>†</sup>				
	(0.03)				
Decision	-0.02 <sup>†</sup>	0.04 <sup>†</sup>	0.05 <sup>†</sup>	0.03	-0.10***
	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Payoff for oneself <sup>a,b</sup>		0.03***	0.02 <sup>†</sup>	0.04***	0.02**
		(0.01)	(0.01)	(0.01)	(0.01)
Payoff for association <sup>a,b</sup>		-0.03***	-0.03***	-0.03***	-0.01 <sup>†</sup>
		(0.01)	(0.01)	(0.01)	(0.006)
Age <sup>a</sup>	0.03	-0.001	-0.01	0.005	0.05
	(0.05)	(0.05)	(0.05)	(0.05)	(0.07)
AIC	15114.8	6095.5	3074.3	3058.6	7203.8
BIC	15203.4	6162.4	3122.2	3106.5	7270.6
N (Observation)	11823	5912	2952	2960	5911
N (Participant)	47	47	47	47	47

1022

1023 Note: <sup>a</sup> We standardized these variables for the analyses.

1024 <sup>b</sup> These variables were added as covariates only when the regressor “Association” (and its interaction)

1025 was not in the regression model, as the regressor “payoff for oneself” qualitatively co-varied with

1026 “Association”, which might cause the collinear issue.

1027 Reference levels were set as follows: Group = NC, Audience = private, Association = good cause

1028 (charity). Table also shows goodness-of-fit statistics: AIC = Akaike Information Criterion, BIC = Bayesian

1029 Information Criterion. Significance: <sup>†</sup>p < 0.06, <sup>\*</sup>p < 0.05, <sup>\*\*</sup>p < 0.01, <sup>\*\*\*</sup>p < 0.001.

1030 **Table 4 Within-subject RSA results in TPJ using valid trials<sup>a</sup>**

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

1031

1032 Note: <sup>a</sup>We excluded trials that did not reach the behavioral criterion (i.e., those with a decision time [DT]  
 1033 shorter than 200ms or longer than mean ± 3\*SD of that individual) or fMRI criterion (all trials in a run with  
 1034 an excessive head motion: ASD: > 5mm; HC: > 3mm). <sup>b</sup>These masks were spheres with a radius of  
 1035 10mm centering on the MNI coordinates based on a recent meta-analysis involving the mentalizing  
 1036 process (peak MNI coordinates: left TPJ/pSTS: -53/-59/20; right TPJ/pSTS: 56/-56/18).

1037 <sup>c</sup>We added the standardized age as the covariates to the regression, using the ImPerm package.

1038 \*\*\* These effects are significantly higher than 0 (i.e., one-sample T-test with 5000 permutations; p<sub>permutation</sub>  
 1039 < 0.001).

---

1040 Abbreviations: l: left, r: right, TPJ: temporoparietal junction; ASD: autism spectrum disorder, HC: healthy  
1041 control.

1042 **Table 5 Within-subject RSA results in TPJ using all 256 trials**

			Spearman's rho (mean ± SD)		P <sub>permutation</sub>	P <sub>permutation</sub> <sup>b</sup>
			ASD	HC		
Neurosynth	ITPJ	Audience	0.023 ± 0.016 <sup>***</sup>	0.023 ± 0.014 <sup>***</sup>	0.967	0.538
		Moral context	0.117 ± 0.047 <sup>***</sup>	0.144 ± 0.050 <sup>***</sup>	0.063	0.094
		Payoff for oneself	-0.004 ± 0.009	-0.002 ± 0.006	0.299	0.358
		Payoff for association	-0.003 ± 0.006	-0.003 ± 0.006	0.932	0.919
	rTPJ	Audience	0.028 ± 0.022 <sup>***</sup>	0.022 ± 0.010 <sup>***</sup>	0.203	0.169
		<b>Moral context</b>	<b>0.098 ± 0.040<sup>***</sup></b>	<b>0.146 ± 0.070<sup>***</sup></b>	<b>0.009</b>	<b>0.027</b>
		Payoff for oneself	-0.003 ± 0.009	-0.003 ± 0.007	0.854	0.387
		Payoff for association	-0.003 ± 0.009	-0.005 ± 0.008	0.386	0.667
Meta-Analysis <sup>a</sup>	ITPJ	Audience	0.018 ± 0.015 <sup>***</sup>	0.021 ± 0.016 <sup>***</sup>	0.579	0.850
		Moral context	0.097 ± 0.058 <sup>***</sup>	0.096 ± 0.058 <sup>***</sup>	0.972	0.914
		Payoff for oneself	-0.001 ± 0.008	-0.00004 ± 0.008	0.721	0.745
		Payoff for association	-0.003 ± 0.005	-0.004 ± 0.007	0.775	0.871
	rTPJ	Audience	0.021 ± 0.019 <sup>***</sup>	0.018 ± 0.013 <sup>***</sup>	0.613	0.451
		<b>Moral context</b>	<b>0.067 ± 0.034<sup>***</sup></b>	<b>0.111 ± 0.064<sup>***</sup></b>	<b>0.006</b>	<b>&lt; 0.001</b>
		Payoff for oneself	-0.001 ± 0.008	-0.001 ± 0.008	0.990	0.528
		Payoff for association	-0.002 ± 0.006	-0.003 ± 0.006	0.796	0.689

1043

1044 Note: *Post-hoc* 2 (group) × 4 (cognitive RDM) mixed ANOVA on the Fisher r-to-z transformed Spearman's  
 1045 rho revealed a strong interaction between group and cognitive RDM only in rTPJ regardless of the way  
 1046 we defined the ROI (the parcellation-based ROI: F(3,126) = 6.59, *p* < 0.001; the coordinate-based ROI:  
 1047 F(3,126) = 7.37, *p* < 0.001), which was not true in ITPJ (the parcellation-based ROI: F(3,126) = 3.00, *p* =  
 1048 0.033; the coordinate-based ROI: F(3,126) = 0.03, *p* = 0.994) after controlling for the age difference,  
 1049 which further confirmed that the specific between-group effect in representing information of Moral  
 1050 Context was unique in rTPJ.

---

1051 <sup>a</sup> These masks were spheres with a radius of 10mm centering on the MNI coordinates based on a recent  
1052 meta-analysis involving the mentalizing process (peak MNI coordinates: left TPJ/pSTS: -53/-59/20; right  
1053 TPJ/pSTS: 56/-56/18).

1054 <sup>b</sup> We added the standardized age as the covariates to the regression, using the ImPerm package.

1055 \*\*\* These effects are significantly higher than 0 (i.e., one-sample T-test with 5000 permutations;  $p_{\text{permutation}}$   
1056 < 0.001).

1057 Abbreviations: l: left, r: right, TPJ: temporoparietal junction; ASD: autism spectrum disorder, HC: healthy  
1058 control.

1059 **Table 6 Supplementary univariate GLM results<sup>a</sup>**

Brain Region	Hemisphere	Cluster Size	MNI			BA	T-value	p(cl-FWE)
			x	y	z			
<b>ASD</b>								
<b>Public &gt; Private</b>								
Cingulate Gyrus/ Corpus Callosum	L	96	-16	8	30		4.99	0.162
<b>HC &gt; ASD</b>								
<b>Private &gt; Public</b>								
Cingulate Gyrus/ Corpus Callosum	L	71	-10	2	30		4.33	0.371
<b>Good &gt; Bad</b>								
Prec/SOG	L	95	-18	-58	32	7	4.84	0.229
IPL/PoCG	L	56	-38	-32	44	40	3.59	0.525

1060

1061 Note: <sup>a</sup>We excluded trials that did not reach the behavioral criterion (i.e., those with a decision time  
 1062 shorter than 200ms or longer than mean  $\pm$  3\*SD of that individual) or fMRI criterion (all trials in a run with  
 1063 an excessive headmotion).

1064 Regions shown here met an uncorrected voxel-level threshold of  $p < 0.001$  with  $k = 50$ . Coordinates  
 1065 shown here were based on Montreal Neurological Institute (MNI) coordinate system. Abbreviations: ASD:  
 1066 autism spectrum disorder, HC: healthy control; L: left, R: right, B: bilateral, BA: Brodmann Area; cl-FWE:  
 1067 cluster-level Family-Wise Error (corrected); CC: corpus callosum, CG: cingulate gyrus, IPL: inferior  
 1068 parietal lobule, PoCG: post-central gyrus, Prec: precuneus, SOG: superior occipital gyrus.