OXFORD

Social Cognitive and Affective Neuroscience, 2018, 460–470

doi: 10.1093/scan/nsy029
Advance Access Publication Date: 28 April 2018
Original article

Neural substrates for moral judgments of psychological versus physical harm

Lily Tsoi,[1] James A. Dungan,[2] Aleksandr Chakroff,[3] and Liane L. Young[1]

[1]Department of Psychology, Boston College, Chestnut Hill, MA 02467, USA, [2]Center for Decision Research, University of Chicago, Chicago, IL 60637, USA, and [3]Charlie Finance, San Fransisco, CA, USA

Correspondence should be addressed to Lily Tsoi, Department of Psychology, Boston College, 140 Commonwealth Ave, McGuinn 300, Chestnut Hill, MA 02467, USA. E-mail: lily.tsoi@bc.edu

Abstract

While we may think about harm as primarily being about physical injury, harm can also take the form of negative psychological impact. Using functional magnetic resonance imaging, we examined the extent to which moral judgments of physical and psychological harms are processed similarly, focusing on brain regions implicated in mental state reasoning or theory of mind, a key cognitive process for moral judgment. First, univariate analyses reveal item-specific features that lead to greater recruitment of theory of mind regions for psychological harm versus physical harm. Second, multivariate pattern analyses reveal sensitivity to the psychological/physical distinction in two regions implicated in theory of mind: the right temporoparietal junction and the precuneus. Third, we find no reliable differences between neurotypical adults and adults with autism spectrum disorder with regard to neural activity related to theory of mind during moral evaluations of psychological and physical harm. Altogether, these results reveal neural sensitivity to the distinction between psychological harm and physical harm.

Key words: morality; theory of mind; autism; fMRI

Introduction

Concerns about harm—about protecting the life and well-being of others—make up one of the primary types of moral concerns that people have (Shweder *et al.*, 1997; Graham *et al.*, 2011). Some researchers theorize that all of morality can be broadly understood in terms of an agent causing either real or perceived harm to a patient (Gray *et al.*, 2012, 2014). If asked to generate different examples of harmful acts, shooting someone or punching someone may be among the more immediately accessible examples. Indeed, prior work on moral judgments of harm has primarily focused on harms causing physical damage and has largely neglected harms causing psychological damage. When it comes to evaluating harms, as in the case of moral judgment, the distinction between physical and psychological may be meaningful.

Whether moral judgments or the cognitive processes that support moral judgments are sensitive to this physical/psychological distinction is the primary question of the present work. First, it is unclear whether people judge physical harms and psychological harms similarly. People may in general judge physical harms more harshly than psychological harms given that physical harms are typically associated with more easily discernable traces when compared to psychological harms. In order to test this first question, we explicitly test for differences in behavioral ratings. Second, cognitive processes involved in evaluating harms may differ for physical and psychological harms. Prior work has revealed contributions of emotional and cognitive processes in moral sensitivity and moral evaluation of harm (Bzdok *et al.*, 2012; Decety *et al.*, 2012). One difference between physical harms and psychological harms may be that psychological harms elicit greater consideration of people's

Received: 19 July 2017; **Revised:** 22 March 2018; **Accepted:** 16 April 2018

mental experience (e.g. victim's mortification, panic) than do physical harms. Alternatively, because mental experiences arising from physical harms must be inferred rather than stated as may be the case for psychological harms, physical harms may actually elicit more mental state reasoning. Indeed, researchers have proposed that mental state reasoning may have evolved primarily for processing behaviors in terms of mental state inferences as opposed to processing explicit information about mental states (Dungan *et al.*, 2016). Here, we use functional magnetic resonance imaging (fMRI) to test these two hypotheses.

The use of neuroimaging methods to test these hypotheses is feasible, given neuroscientific literature revealing a network of brain regions that support the capacity to attribute, infer and reason about mental states, a capacity often referred to as theory of mind (ToM) (Fletcher *et al.*, 1995; Gallagher *et al.*, 2000; Saxe & Kanwisher, 2003; Gobbini *et al.*, 2007; Molenberghs *et al.*, 2016b; Schurz *et al.*, 2014). This network includes bilateral temporoparietal junction (TPJ), precuneus and dorsomedial prefrontal cortex (dmPFC). Critically, ToM has been found to play a role in moral judgment (for a review, see Young & Tsoi, 2013). Indeed, not only are people sensitive to characteristics of perpetrators who cause harm (e.g. group membership; Molenberghs *et al.*, 2016a), but they also often take into consideration the beliefs and intentions of the perpetrator when evaluating an action: Generally, people judge harms caused intentionally as worse than harm caused accidentally. This pattern is found across development (Decety *et al.*, 2012) and across many different societies (although the extent to which intentions influence judgments may differ; Barrett *et al.*, 2016). In fact, people consistently judge attempted harm (e.g. wanting to shoot a person but failing to do so) as morally worse than accidental harm (e.g. wanting to shoot a duck but missing and shooting a person instead) although in the case of attempted harm no actual harm occurs (Young *et al.*, 2007). This pattern is also reflected neurally, with greater activity in ToM regions for attempted harms than for intended harms, accidental harms, or neutral acts (Young, Cushman *et al.*, 2007). Other work extending this line of inquiry has consequently revealed a causal role of ToM in moral judgment by showing that transiently disrupting activity in the right TPJ, a key node in the ToM network, leads people to rely less on the actor's mental states (i.e. leading them to judge attempted harms as more morally permissible) (Young *et al.*, 2010). Similarly, adults with autism spectrum disorder (ASD), a neurodevelopmental disorder characterized by impairments in social interactions, have been found to rely less on mental states and more on the outcome of the action when making moral judgments (Moran *et al.*, 2011). If psychological harms do elicit ToM to a greater extent than physical harms, then activity in ToM regions may be greater for psychological harms than for physical harms. Moreover, we may expect to see diminished neural differences between psychological and physical harms among individuals with ASD when compared to neurotypical individuals.

While the physical/psychological distinction has not been directly explored with regard to moral judgment, it has been widely investigated by social psychologists and neuroscientists studying pain. In particular, extensive prior work has focused on social or psychological pain—for example, negative experiences associated with social exclusion, isolation, or loss (from a failed relationship or death of a loved one) (Eisenberger, 2012). People readily make associations between psychological pain and physical pain (e.g. using terms like *broken* hearts; *hurt* feelings) and linguistic associations, such as the examples provided, can be found around the world (MacDonald & Leary, 2005).

While negative social experience, often termed as "social pain", differs from physical pain in that it does not share the component of physical pain that codes for the localization, quality and intensity of the pain (sensory component), social pain has been revealed to overlap with the component of physical pain that codes for the unpleasantness or distress of pain (affective component) (Eisenberger, 2015; but see Kross *et al.*, 2011). There appears to be shared neural circuitry for physical and psychological pain: The dorsal anterior cingulate cortex (dACC) and the anterior insula supporting the affective component of physical pain are also recruited by negative social experiences, including but not limited to being socially excluded (Eisenberger *et al.*, 2003), viewing images related to rejection (Kross *et al.*, 2007) and being negatively evaluated (Eisenberger *et al.*, 2011). Altogether, these results suggest that physical and social pain share some common psychological processes aside from mere metaphorical similarity (Eisenberger, 2015). Intriguingly, recent work using multivariate pattern analyses (MVPA) reveals the ability to distinguish physical pain and psychological pain from spatial patterns of neural activity (Woo *et al.*, 2014), leading to the possibility that the physical/psychological distinction may be encoded in distributed patterns of neural activity and not detectable in overall response magnitudes measured in conventional univariate analyses. We test this possibility directly in the domain of moral judgment.

The present fMRI study has two goals: (1) to compare psychological harms with physical harms and (2) to examine whether individuals with ASD process psychological harms versus physical harms differently from neurotypical individuals. Univariate analyses and multivariate analyses were used to examine neural differences between psychological harms and physical harms across the whole brain. Given prior work revealing a role for ToM in moral judgment, we also conducted region of interest (ROI) or ROI-based univariate and multivariate analyses to examine the specific role of brain regions implicated in ToM for processing psychological and physical harms.

Materials and methods

This study presents new analyses of previously published data (Koster-Hale *et al.*, 2013; Chakroff *et al.*, 2016; Wasserman *et al.*, 2017). Here, we focus on examining neural distinctions between moral judgments of physical harm and moral judgments of psychological harm using univariate analyses and multivariate pattern analyses as well as comparing these results across two different groups: a neurotypical group and an ASD group.

Participants

Two groups of people participated in the study: one group of neurotypical adults (NT) and one group of adults with autism spectrum disorder (ASD). The NT group consisted of 25 adults from the Boston area between the ages of 18 and 50 ($M = 28.56$, $SD = 10.10$; 7 women). The ASD group consisted of 16 adults between the ages of 20 and 46 ($M = 31.13$, $SD = 8.21$; 2 women). ASD participants were recruited using advertisements placed with the Asperger's Association of New England.

Both groups were first prescreened using the Autism Quotient questionnaire (AQ; Baron-Cohen *et al.*, 2001). The ASD group scored significantly higher on the AQ than the NT group ($M_{NT} = 17.33$, $SD_{NT} = 5.88$, $M_{ASD} = 32.64$, $SD_{ASD} = 6.96$; $t(27) = 6.42$, $p < 0.001$). The ASD participants also underwent the Autism

Diagnostic Observation Schedule (ADOS; Lord *et al.*, 2000) and impression by a clinician trained in both ADOS administration and diagnosis of ASD. All ASD participants received a diagnosis of ASD based on their total ADOS score (criterion ≥ 7; $M = 9.50$, $SD = 2.68$) and clinical impression based on the diagnostic criteria of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (APA, 2000). Both NT and ASD groups did not differ in age ($M_{NT} = 28.56$, $M_{ASD} = 31.13$; $t(39) = 0.85$, $p = 0.40$) or IQ ($M_{NT} = 117.52$; $M_{ASD} = 119.80$; $t(38) = 0.49$, $p = 0.63$).

All participants were right-handed native English speakers with normal or corrected-to-normal vision. They all provided written informed consent and were paid for participating in the study. The study was approved the Institutional Review Board at the Massachusetts Institute of Technology.

None of the participants exhibited excessive in-scanner movement (> 3 mm within-run displacement).

Experimental task

Participants were scanned while reading 60 vignettes in the second-person point of view (i.e. "You buy spinach for your grandmother. You use it to make her a large salad"; see Supplementary Material for a full list of vignettes). Twenty-four vignettes involved intentional and accidental harm violations. Of these 24 vignettes, 12 scenarios involved physical harm (e.g. giving someone food poisoning) and 12 involved psychological harm (e.g. exposing someone to their object of phobia). Twelve vignettes involved neutral acts (e.g. stepping into a puddle). The remaining vignettes involved different types of purity violations (e.g. smearing feces on one's own face [pathogen] vs. having sex with a sibling [incest]); we will not be reporting on these items as they have been reported elsewhere (Chakroff *et al.*, 2016; Wasserman *et al.*, 2017).

Each scenario contained five segments: background (6 s), action (4 s), outcome (4 s), intent (4 s) and judgment (4 s). During the judgment segment, participants made moral judgments of the action (i.e. "Judge how morally wrong your behavior was") on a scale from 1 (not at all) to 4 (very) using a response box. The scale was always presented with 1 on the left of the screen and 4 on the right of the screen. Ten scenarios were presented in each 5.5-min run; there were a total of six runs, which lasted 33.2 min. Scenarios were presented in a pseudorandom order, with the order of conditions counterbalanced across runs and across participants.

Acquisition and preprocessing of fMRI

The fMRI data were collected using a 12-channel head coil in a 3 T Siemens scanner at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at the Massachusetts Institute of Technology. Data were acquired in 26 3×3×4 mm near-axial slices using standard echoplanar imaging procedures (TR (repetition time) = 2 s, TE (echo time) = 40 ms, flip angle = 90°). The first 4 s of each run were excluded to allow for steady state magnetization. Data processing and analysis were performed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm) and custom software. The data were motion-corrected, realigned, normalized onto a common brain space (Montreal Neurological Institute, MNI, template), spatially smoothed using a Gaussian filter (full-width half-maximum = 8 mm kernel) and high-pass filtered (128 s).

Data analysis

We provide an outline of our analysis plan: First, we conducted analyses of in-scanner responses to examine whether judgments of moral permissibility differed across conditions (physical harm, psychological harm and neutral acts) and across groups (NT vs. ASD). We then conducted whole-brain and ROI univariate analyses to examine response magnitudes across our conditions. These two analyses served different functions: Whole-brain analyses were conducted to reveal involvement of any region in the brain in processing psychological and physical harms. ROI analyses, on the other hand, allowed us to examine neural activity across conditions within ToM regions defined using an independent functional localizer task; these analyses allowed us to directly test our hypotheses regarding the role of ToM in moral judgments of psychological and physical harm. Notably, the ROI analyses were conducted by taking into account by-participant and by-item variance, allowing us to make inferences that can generalize past the specific sample of participants tested as well as the specific items we used in the present study (Judd *et al.*, 2012). We also conducted whole-brain and ROI multivariate analyses to examine whether the physical/psychological dimension of harm is a feature encoded in spatial patterns of activity across the brain and specifically within ToM regions, respectively. These analyses focused on activity for the duration of the entire trial (though components of the trial are analyzed separately and are reported in Supplementary Material). Our main analyses included comparisons between physical and psychological harm, between psychological harm and neutral acts and between physical harm and neutral acts. Other analyses, which are referenced throughout the paper, are reported in Supplementary Material.

Behavioral analyses. Behavioral analyses were conducted in R (version 3.3.3); scripts are available on GitHub (https://github.com/tsoices/psych-phys-harm).

Ratings were analyzed using cumulative link mixed models (clmm) with an ordinal response term (from a scale of 1–4). Mixed models were run using the package "ordinal" (Christensen, 2015). We were primarily interested in understanding whether ratings differed across conditions for the NT group and whether differences across conditions differed for the NT and ASD groups. Our full model included the following predictors: condition (physical harm vs. psychological harm vs. neutral act) and group (NT vs. ASD). We also examined the two-way interaction between condition and group. Participant and item were included as random effects, and we fit an intercept for each participant and for each item, allowing the intercept to vary across individuals and items. To assess the importance of our predictors of interest, we performed likelihood ratio tests (LRTs) to test whether the model including a given predictor would provide a better fit to the data than a model without that term.

Reaction times were analyzed using linear mixed effect models the package 'lme4' (Bates *et al.*, 2015) in R, with the same predictors and random effects as the analyses for ratings above.

Whole-brain analyses. Whole-brain analyses were conducted using SPM8; anatomical labels for peak coordinates were retrieved using SPM Anatomy Toolbox v2 (Eickhoff *et al.*, 2005), and results were visualized using the xjView toolbox (http://www.alivelearn.net/xjview).

Analyses of response magnitudes. Preprocessed images were analyzed using a general linear model (GLM) framework. The experiment had a slow event-related design, which was modeled using

boxcar regressors convolved with a standard hemodynamic response function (HRF). An event was defined as a single vignette (22 s), and its onset was defined by the onset of the background component of each vignette (results from models of events defined as just the outcome component or just the intent component are included in Supplementary Material). The GLM also included six motion parameters as nuisance regressors.

Beta values were estimated in each voxel for all 10 conditions: 2 (intent: intentional vs. accidental) × 5 (content: psychological harm, physical harm, pathogen, incest, neutral). Contrast maps for the following contrasts were produced for each participant: (1) *psychological harm* > *physical harm*, (2) *physical harm* > *psychological harm*, (3) *psychological harm* > *neutral act*, and (4) *physical harm* > *neutral act*.

For each contrast, participants' images were used to perform group-level analyses. To correct for multiple comparisons, images from the group-level analyses were subjected to a voxel-wise threshold of $P < 0.001$ (uncorrected) and a cluster extent threshold ensuring $q < 0.05$ (false discovery rate (FDR)-corrected).

Analyses of spatial patterns. fMRI time courses for all voxels were extracted from unsmoothed images and high-pass filtered with a 128-s cutoff. The signals were mean-centered to normalize intensity differences among runs. Time courses were mapped to conditions following a GLM framework. Regressors for the conditions of interest (psychological harm and physical harm) per run were constructed by convolving the onset of each trial with the canonical HRF. If the height of the regressor at each time point was greater than the mean height of the regressor, then the time point was assigned to the condition. Time points were labeled as either a 'psychological harm' class or a 'physical harm' class.

A searchlight approach was used: A three-voxel radius sphere was moved throughout the brain, centering on each voxel. In each searchlight sphere, a binary classification (psychological harm or physical harm) was performed using a Gaussian Naïve Bayes classifier. For validating the classification, a leave-one-run-out procedure was used, wherein data from one of the runs were reserved for testing, and the remaining data were used for training (a total of 6-fold cross-validations). Participants' output searchlight images were used to perform group-level analyses. The group maps were thresholded using a voxel-wise threshold of $P < 0.001$ (uncorrected), after which we corrected for multiple comparisons by using a cluster extent threshold ensuring $q = 0.05$ (FDR-corrected).

ROI analyses.

Defining functional ROIs. A ToM localizer task (Saxe & Kanwisher, 2003; Dodell-Feder *et al*., 2011) was used to functionally define the following regions implicated in ToM: the rTPJ, lTPJ, precuneus and dmPFC. Details about the task can be found in Supplementary Material.

Analyses of response magnitudes. The BOLD response over baseline to each condition was calculated for each ROI. Baseline response in each ROI was calculated as the average response in that ROI at all time points during the resting period, excluding the first 6 s after the offset of each stimulus (to allow the hemodynamic response to decay). The percent signal change (PSC) relative to baseline was calculated for each time point in each condition, averaging across all voxels in the ROI, where PSC (at time t) = 100 × [(average magnitude response for condition at time t – average magnitude response for fixation)/average magnitude response for fixation]. The PSC was averaged across the entire trial (11 TRs or 22 s; offset 6 s from presentation time to adjust for hemodynamic lag) to estimate a single PSC for each condition in each ROI for each participant. Linear mixed models were run using the package 'lme4' (Bates *et al*., 2015) in R. We were primarily interested in understanding whether the PSC for ROIs differed across conditions in the NT group, and whether differences across conditions differed for the NT and ASD groups. Our full model included the following predictors: condition (physical harm vs. psychological harm vs. neutral act) and group (NT vs. ASD). We also examined the two-way interaction between condition and group. Participant and item were included as random effects, and we fit an intercept for each participant and for each item, allowing the intercept to vary across individuals and items. To assess the importance of our predictors of interest, we performed LRTs to test whether the model with a given predictor would provide a better fit to the data than a model without that term.

Analyses of spatial patterns. A similar procedure to the searchlight procedure was used, except that instead of a searchlight sphere, ROIs were used. An accuracy score averaged across training/testing set combinations was computed for each ROI and every individual. Permutation tests were also conducted for each ROI using a fold-wise permutation scheme with 1000 iterations in which psychological and physical labels were randomly shuffled.

Results

Behavioral results

Ratings of moral wrongness were made on a scale from 1 (not at all) to 4 (very). The interaction between condition and group was not significant (LRT, $\chi^2(2) = 2.067$, $P = 0.36$), suggesting that differences in moral judgments across the conditions did not differ for the NT and ASD groups (Figure 1). However, there was a main effect of condition (LRT, $\chi^2(2) = 75.98$, $P < 0.001$): Pairwise contrasts revealed that physical and psychological harms were both judged as more wrong than neutral acts (physical: $z = 13.163$, $P < 0.001$; psychological: $z = 13.432$, $P < 0.001$) but no different from each other ($z = 2.102$, $P = 0.90$), even when accounting for by-participant and by-item variance.

We also examined reaction times for in-scanner ratings: There was an interaction between condition and group (LRT, $\chi^2(2) = 6.51$, $P = 0.039$). Further examination within each group revealed an effect of condition for the NT group (physical: $M = 1.572$, $SD = 0.547$, psychological: $M = 1.574$, $SD = 0.515$, neutral: $M = 1.289$, $SD = 0.543$; LRT, $\chi^2(2) = 7.89$, $P = 0.019$) but not for the ASD group (physical: $M = 1.285$, $SD = 0.524$, psychological: $M = 1.348$, $SD = 0.453$, neutral: $M = 1.284$, $SD = 0.427$; LRT, $\chi^2(2) = 0.42$, $P = 0.81$). That is, reaction times in the NT group were similar for physical and psychological harms ($t(20) = 0.03$, $P = 1.00$), but both differed from neutral acts (physical vs. neutral: $t(20) = 4.5$, $P < 0.001$; psychological vs. neutral: $t(20) = 3.2$, $P = 0.004$). Meanwhile, reaction times in the ASD group were similar across all three conditions (Ps > 0.05).

Neural results

Regions recruited for psychological and physical harm. We compared psychological harms with neutral acts and physical harms with neutral acts across the whole brain (Table 1; Figure 2). For the NT group, the following regions were recruited more for psychological harms than for neutral acts (psychological harm > neutral act): the left inferior temporal gyrus, precuneus, left superior medial gyrus, left angular gyrus, right

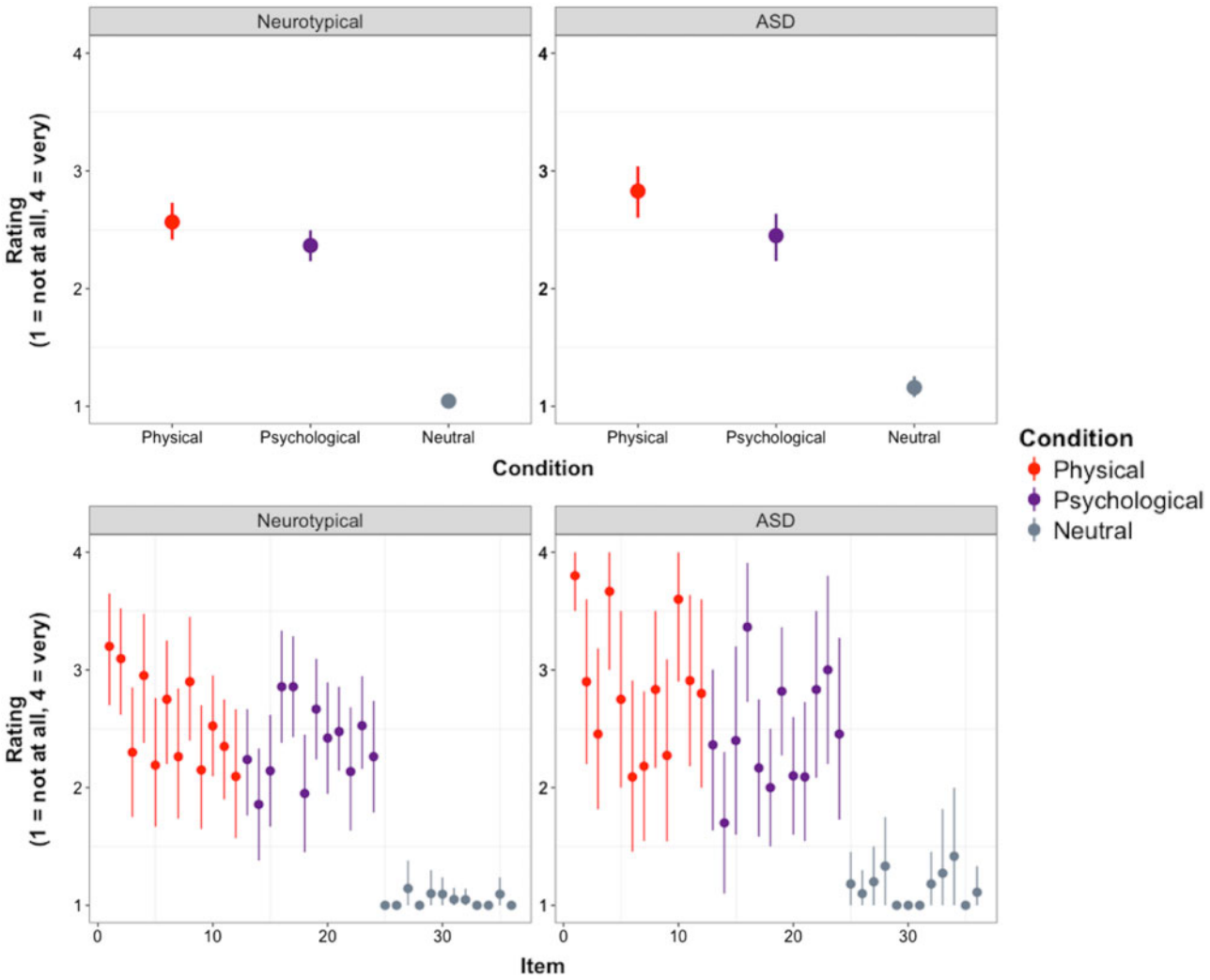

Fig. 1. In-scanner ratings of moral wrongness across conditions (top) and across items (bottom) for each group. Error bars denote 95% confidence interval (CI).

superior temporal gyrus and bilateral caudate nucleus; several of these regions overlapped with the ToM network as elicited by the localizer task. One region was recruited more for physical harms than for neutral acts (physical harm > neutral act): the right superior temporal gyrus; this region had a small overlap with the rTPJ, a node in the ToM network as elicited by the localizer task. A similar extent of overlap between activity for each contrast and the ToM network was found when we modeled just the portion of the trial in which participants received direct information about the harm (outcome component) as opposed to the entire trial as we have done here (Supplementary Material). When we included reaction time as a control regressor, we found overlap between activity for the psychological harm > neutral act contrast and the ToM network but no overlap in activity for the physical harm > neutral act contrast and the ToM network (Supplementary Material). Overall, these results show a consistent pattern, whereby psychological harms relative to neutral acts elicit activity in regions implicated in ToM.

Moreover, no differences between the NT group and ASD group were found at the whole-brain level (see results for the ASD group in Supplementary Material). That is, independent-samples *t* test comparing the NT group with the ASD group revealed no group differences at the whole-brain level for *psychological harm* > *neutral act* or for *physical harm* > *neutral act*.

Regions preferentially recruited for psychological harm vs. physical harm and vice versa. We examined, at the whole-brain level, regions that were recruited more for psychological harm versus physical harm and vice versa (Table 2; Figure 2). For the NT group, the contrast psychological harm > physical harm revealed clusters with peak coordinates in the left precuneus, right superior medial gyrus, left middle temporal gyrus and right angular gyrus; these regions overlapped with the precuneus, dmPFC and rTPJ as elicited by the ToM localizer. On the other hand, the contrast physical harm > psychological harm revealed no significant clusters; only when we use a more lenient threshold (voxel-wise, $P < 0.001$, $k = 10$) do we see activation of the left inferior frontal gyrus and right calcarine gyrus. A similar extent of overlap between the psychological harm > physical harm contrast and the ToM network was found when we modeled just the portion of the trial in which participants received direct information about the harm (outcome component) as opposed to the entire trial as we have done here (Supplementary Material). When we included reaction time as a control regressor, the precuneus was the only region that

overlapped between the psychological harm > physical harm contrast and the ToM network (although we do see other ToM regions when we use a more lenient threshold; see Supplementary Material).

Of the regions that were revealed for the contrast *psychological harm > physical harm*, one region was recruited more for the NT group than the ASD group: the anterior cingulate cortex (see results for the ASD group in Supplementary Material). However, this region did not appear when we ran the same analysis based on the model of just the outcome component (Supplementary Material) nor did it appear when we included reaction time as a regressor in the model used here (Supplementary Material).

We additionally examined whether there was an interaction between harm type and intent: that is, whether the difference between psychological and physical harm differed for harm caused intentionally versus accidentally. We did not find any

Table 1. Regions showing greater activation in the NT group for psychological harm *vs*. neutral acts (top) and physical harm vs. neutral acts (bottom)

Region name	MNI coordinates x	y	z	t value	# of voxels
Psychological > neutral: NT					
L inferior temporal gyrus	**−54**	**−1**	**−32**	**7.00**	**237**
L medial temporal pole	−51	8	−35	6.65	
L inferior temporal gyrus	−54	−13	−29	5.44	
Precuneus	**0**	**−58**	**40**	**6.41**	**213**
Precuneus	0	−55	31	6.09	
L superior medial gyrus	**−9**	**32**	**61**	**6.27**	**128**
R superior frontal gyrus	18	26	61	5.69	
L posterior medial frontal	−6	23	64	5.22	
L angular gyrus	**−54**	**−58**	**31**	**6.14**	**316**
L supramarginal gyrus	−63	−49	28	5.71	
Area PGp	−48	−73	34	5.48	
L superior medial gyrus	**−3**	**59**	**28**	**5.92**	**292**
L anterior cingulate cortex	−6	47	13	5.35	
L middle frontal gyrus	−27	53	28	4.97	
R superior temporal gyrus	**48**	**−58**	**22**	**5.80**	**144**
R superior temporal gyrus	63	−52	22	5.66	
R middle occipital gyrus	57	−64	25	4.92	
R caudate nucleus	**12**	**8**	**13**	**5.36**	**53**
N/A	12	5	4	4.53	
L caudate nucleus	**−9**	**8**	**4**	**5.20**	**45**
L caudate nucleus	−15	5	16	4.92	
L putamen	−18	8	7	3.92	
Physical > neutral: NT					
R superior temporal gyrus	**60**	**−58**	**22**	**4.95**	**55**
R middle temporal gyrus	66	−49	13	4.82	
R middle temporal gyrus	51	−55	19	3.91	

Note: Regions in bold and plain denote peak and subpeak coordinates, respectively. All regions survived cluster-level correction (FDR, $q < 0.05$).

Table 2. Regions recruited for psychological *vs*. physical harm and vice versa by group

Region name	MNI coordinates x	y	z	t value	# of voxels
Psychological > physical: NT					
L precuneus	**−3**	**−55**	**31**	**7.12**	**210**
R superior medial gyrus	**3**	**56**	**22**	**6.31**	**272**
L mid orbital gyrus	−3	47	−11	6.13	
L superior medial gyrus	−9	62	31	4.58	
L middle temporal gyrus	**−57**	**−10**	**−23**	**5.85**	**58**
R angular gyrus	**45**	**−64**	**31**	**4.77**	**38**
Psychological > physical: ASD					
R precuneus	**6**	**−58**	**31**	**7.46**	**263**
L precuneus	−12	−49	40	6.76	
L precuneus	−3	−61	25	6.35	
L middle temporal gyrus	**−63**	**−7**	**−23**	**5.56**	**49**
L middle temporal gyrus	−57	−1	−26	5.11	
L middle temporal gyrus	−66	−10	−14	5.10	
R supramarginal gyrus	**51**	**−31**	**25**	**5.24**	**42**
Psychological > physical: NT > ASD					
Anterior cingulate cortex	**0**	**47**	**13**	**4.54**	**41**
Anterior cingulate cortex	0	44	4	3.86	
Physical > psychological: NT					
none					
Physical > psychological: ASD					
none					
Physical > psychological: NT > ASD					
none					

Note: Regions in bold and plain denote peak and subpeak coordinates, respectively. All regions survived cluster-level correction (FDR, $q < 0.05$).

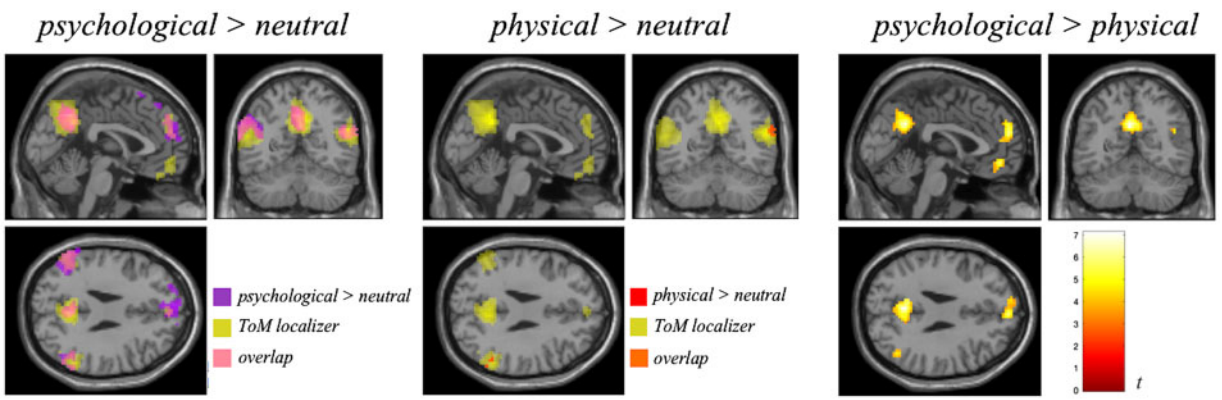

Fig. 2. Results of whole-brain univariate analyses for the neurotypical (NT) group. Left: Overlap in neural substrates for theory of mind and processing of psychological harm *vs* neutral acts. Center: Overlap in neutral substrates for theory of mind and processing of physical harm *vs* neutral acts. Right: Regions showing preferential activation for psychological *vs* physical harms; no regions showed preferential activation for physical harms *vs* psychological harms. For all images, cluster-level correction (FDR, $q < 0.05$) was applied. Images viewed at $x = 0$, $y = -58$, $z = 28$.

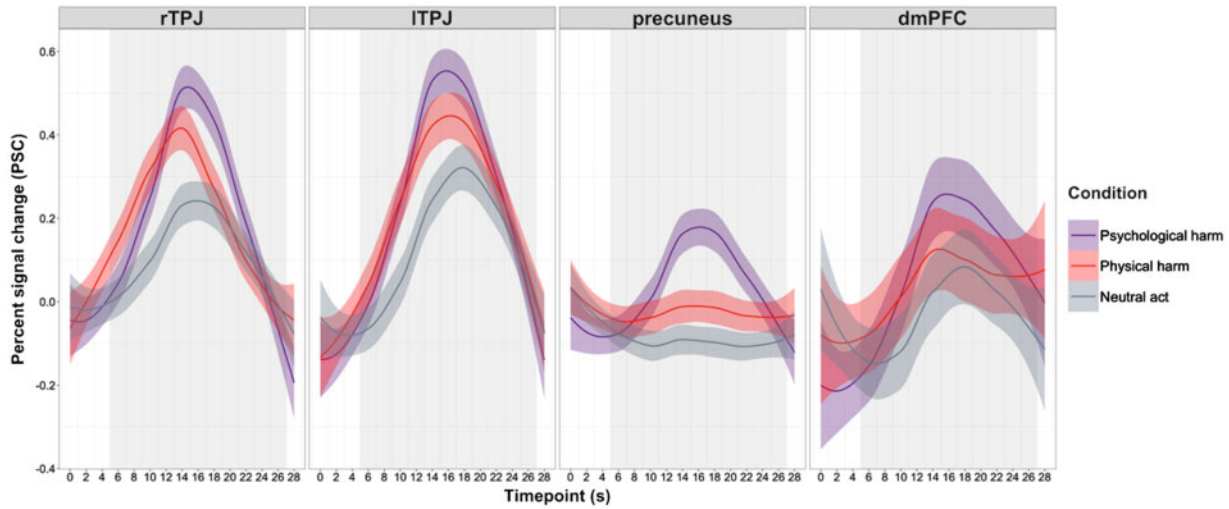

Fig. 3. Time courses for each condition across regions of interest for the neurotypical (NT) group. Shaded areas (6–26 s) denote time windows during which a trial is presented, adjusted for hemodynamic lag. Ribbons denote 95% confidence interval (CI).

regions sensitive to this interaction, though we that acknowledge the analysis may have been underpowered (Supplementary Material).

The role of ToM regions in processing psychological harm and physical harm. To examine the specific role of ToM regions in processing psychological harm, physical harm and neutral acts for the NT group, we conducted ROI-based univariate analyses (Figure 3). We used a linear mixed model to compare the percent signal change over the entire time window within ROIs across the three conditions while taking into account by-participant and by-item variance. LRTs revealed no significant interaction between condition and ROI ($\chi^2(6) = 0.965$, $P = 0.99$). Importantly, there was a significant main effect of condition ($\chi^2(2) = 8.42$, $P = 0.015$), and pairwise comparisons (with degrees of freedom approximated using the Satterthwaite method and P values adjusted using the Tukey method for correction of multiple comparisons) revealed a significant difference between psychological harms vs. neutral acts ($t(36.89) = 2.979$, $P = 0.014$), no significant difference between physical harms vs. neutral acts ($t(36.89) = 2.106$, $P = 0.10$) and no significant difference between psychological harms and physical harms ($t(36.89) = 0.874$, $P = 0.66$). Contrasts performed for each individual ROI also revealed no significant difference between psychological harms and physical harms in any of the ROIs (Ps > 0.05). Moreover, adding a random slope to the model allowing the effect of condition to vary across participants did not improve the model ($\chi^2(5) = 0.552$, $P = 0.99$), suggesting that the effect of condition did not significantly vary across participants. Overall, these results provide evidence for the general pattern of condition effects across ToM ROIs: greater PSC for psychological harms vs. neutral acts and no difference in PSC for psychological harms vs. physical harms. Further analyses separating the time course into each individual trial component (i.e. background; action; outcome; intent; judgment) revealed no differences between psychological harms and physical harms for most components with the exception of the outcome component (see Supplementary Material). Including reaction time as a regressor did not affect the results.

Moreover, no differences between the NT group and ASD group were found at the ROI level (see results for the ASD group in Supplementary Material). That is, there was no significant interaction between condition and group nor a main effect of group for any ToM ROI (Ps > 0.05).

Information separating psychological harm from physical harm can be decoded from spatial patterns of activity in the brain. We examined, using a searchlight procedure, regions for which the spatial patterns of activity for psychological harm and physical harm can be accurately classified above-chance level (50%). For the NT group, above-chance classification was found for clusters with peak coordinates in the left precuneus, superior medial gyrus, anterior cingulate cortex and left middle temporal gyrus (Table 3). The use of a different classifier produced similar results (Supplementary Material).

Additionally, we examined whether the spatial patterns of neural activity for psychological harm and physical harm could be accurately classified above chance level (50%) in any ToM ROI (Figure 4). As expected, permutation tests using shuffled condition labels showed at-chance classification for all ROIs (Ps > 0.05), whereas experimental tests using true condition labels revealed above-chance classification accuracies for the rTPJ and precuneus (rTPJ: $t(23) = 2.287$, $P = 0.016$; precuneus: $t(24) = 4.403$, $P < 0.001$) and at-chance classification accuracies for the lTPJ and dmPFC (lTPJ: $t(23) = 1.148$, $P = 0.131$; dmPFC: $t(14) = 0.142$, $P = 0.445$). Comparisons between the experimental tests and permutation tests resulted in significantly greater differences for the experimental tests versus the permutation tests for both the rTPJ and the precuneus (rTPJ: $t(23) = 2.249$, $P = 0.017$; precuneus: $t(24) = 4.389$, $P < 0.001$). These results suggest that the rTPJ and precuneus are able to distinguish, in their voxel-wise patterns of activity, differences between psychological and physical harm.

We also compared the searchlight results of the NT group to that of the ASD group (Table 3); no cluster emerged even we used a more lenient threshold (voxel-wise $P < 0.001$, $k = 10$). Moreover, we compared the two groups' classification accuracies for each ROI (Figure 4). There was no significant difference in mean classification accuracy between the NT and ASD groups for any ToM ROI (Ps > 0.05) (see results of the ASD group in Supplementary Material).

Discussion

When judging the extent to which an act is morally wrong, people tend to heavily rely on the mental states of the actor: whether they think the man *intended* to poison his friend or whether they think the woman *wanted* to injure her cousin. Much work on adult moral psychology has focused on physical harms (e.g. trolley dilemmas; Greene *et al.*, 2001), although in reality harm does not always take the form of physical injury. In this study, we investigated whether people engage in mental state reasoning to a similar extent for physically and psychologically harmful actions by exploring differences in neural processing of both.

We first examined the extent to which neural substrates for processing psychological harm and physical harm overlapped. Whole-brain analyses revealed that while no region was preferentially recruited for physical harm over psychological harm (after cluster-level correction), regions overlapping with the ToM network (i.e. rTPJ, precuneus and dmPFC) were recruited preferentially for psychological over physical harm. These results suggest that people engage in ToM to a greater extent during evaluations of psychological harms when compared to physical harms. This finding is intriguing, since we found no differences in rating: psychological harms were not judged as more or less morally wrong than physical harms. One potential explanation for the asymmetry could be that the specific items we used in the psychological harm condition required greater consideration of people's minds than the items we used in the physical harm condition. To test this idea, we collected, from a separate sample of people, ratings of the extent to which an item made them think about thoughts and desires; indeed, ratings were significantly higher for items in the psychological harm condition than in the physical harm condition ($t(16.299) = 2.719$, $P = 0.015$). This explanation appears plausible given that after taking into account by-participant and by-item variance (which allows us to make inferences that can generalize past the sample of participants we used as well as the items we used in our study), we found no significant difference in response magnitude between psychological and physical harms in any ToM ROI when averaging over the entire time course.

Additionally, our ROI-based MVPA revealed significantly above-chance classification accuracies for only the rTPJ and the precuneus. These results suggest that information separating the two harms are encoded in two regions of the mentalizing or ToM network. This finding raises a question: What features are being captured in spatial patterns of neural activity in the rTPJ

Table 3. Searchlight results

Region name	MNI coordinates x	y	z	*t* value	# of voxels
NT					
L precuneus	**−6**	**−61**	**34**	**5.21**	**163**
L precuneus	−9	−52	25	5.06	
R precuneus	12	−52	37	4.14	
Superior medial gyrus	**0**	**53**	**37**	**4.93**	**91**
R superior medial gyrus	6	53	16	4.34	
L superior medial gyrus	−3	53	16	4.31	
Anterior cingulate cortex	**0**	**44**	**4**	**4.64**	**62**
R superior medial gyrus	6	50	7	4.56	
L anterior cingulate cortex	−3	50	−2	4.37	
L middle temporal gyrus	**−48**	**−64**	**7**	**4.64**	**32**
L middle temporal gyrus	−51	−55	13	3.88	
ASD					
Precuneus	**0**	**−64**	**25**	**6.79**	**254**
Calcarine gyrus	0	−61	13	6.37	
R precuneus	3	−61	43	5.27	
Mid-cingulate cortex	**0**	**−34**	**34**	**5.04**	**42**
L mid-cingulate cortex	−3	−25	37	4.61	
Mid-cingulate cortex	0	−13	34	4.45	
NT > ASD					
none					

Note: Regions in bold and plain denote peak and subpeak coordinates, respectively. All regions survived cluster-level correction (FDR, $q < 0.05$).

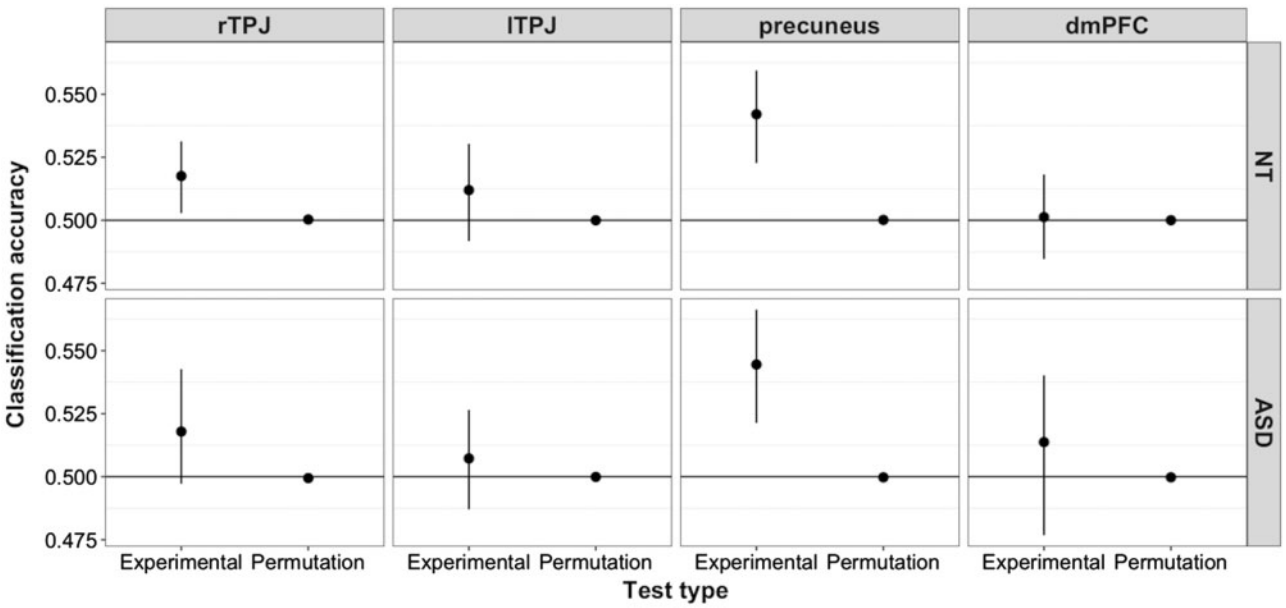

Fig. 4. Classification accuracies by regions of interest, test type, and group (neurotypical [NT]: top; autism spectrum disorder [ASD]: bottom). Chance is at 50%. Error bars denote 95% confidence interval (CI).

and precuneus? As an exploratory approach to this question, we fed the text of the items in the study to a computational linguistic tool (Coh-Metrix) with the aim of potentially uncovering linguistic features that could account for the differences between items in the two conditions. While there were no significant differences between the number of words (Welch two-sample *t* test: $t(32) = 1.3$, $P = 0.2$) or readability of the items (Welch two-sample *t*-test: $t(45) = 0.028$, $P = 1$) across the psychological harm and physical harm conditions, items in the psychological harm condition scored lower in word concreteness (the extent to which content words are concrete vs. abstract; Welch two-sample *t* test: $t(42) = 2.7$, $P = 0.01$). One possibility may be that the rTPJ and precuneus are encoding differences in the amount of effort required to reason about the two types of harm; more effort may be needed to reason about psychological harms than concrete physical harms. We explored this possibility by examining whether reaction times were longer for psychological harms than for physical harms. We did not find any evidence that this was the case: Reaction times did not differ across psychological and physical harm across all participants ($t(31) = 0.49$, $P = 0.60$) nor across participants within each group (NT group: $t(20) = 0.03$, $P = 1$; ASD group: $t(10) = 0.90$, $P = 0.4$). However, because participants were not instructed to make a response as quickly as they could, these null results are difficult to interpret. Future studies may directly test this hypothesis.

Not too surprisingly, most regions that showed preferential recruitment for psychological harms over physical harms in our whole-brain univariate analyses were also regions that showed above-chance classification accuracies in our multivariate searchlight analyses. These independent analyses converge to support the idea that the precuneus, superior medial gyrus and middle temporal gyrus are able to distinguish—in both their response magnitude and spatial patterns of activity—between psychological harms and physical harms.

Finally, while a large focus of this study was on cognitive and neural processing of psychological and physical harms among neurotypical adults, we also examined how these processes were similar/different in a clinical population characterized by difficulties with social interactions. Prior work has revealed no evidence of any difference between neurotypical and ASD groups regarding recruitment of theory of mind regions during false belief tasks (Dufour *et al.*, 2013); similarly, the present study showed no reliable group difference in recruitment of theory of mind regions for moral judgments of psychological harms and physical harms. However, we note that a lack of effect could be due to our small sample size for the ASD group. Nevertheless, given academic and clinical interests in understanding the social cognitive capacities that are and are not affected by ASD, we provide our results here and in Supplementary Material.

Limitations

We acknowledge that while our MVPA classification accuracies were at above-chance levels, the effect sizes were modest. Classification accuracies around this level may reflect difficulty with classification of more abstract features; indeed, classification accuracies are similar across many studies examining socially relevant, higher level features (e.g. Chiu *et al.*, 2011; Kaul *et al.*, 2011; Ratner *et al.*, 2013; Anzellotti *et al.*, 2014; Tsoi, Dungan, Waytz, & Young, 2016). As more researchers start using MVPA to address socially- and morally relevant questions, we may be able to better characterize specific features that lead to lower classification accuracies.

Relatedly, we were concerned that our MVPA results could be explained by potential confounds such as motor differences elicited during the experiment in response to psychological harm and physical harm. However, there were no statistically significant differences in behavioral ratings or reaction times for psychological and physical harms, making it unlikely that our MVPA results reflect motor differences across the two conditions. While we tried to minimize differences across psychological and physical harms (for instance, by ensuring that there was no difference in the number of words presented across conditions), we recognize the possibility that the MVPA results may nevertheless reflect differences in some aspect of the stimuli unrelated to our desired manipulation of the physical/psychological dimension.

Conclusion

In short, this study has revealed sensitivity within regions implicated in theory of mind to the physical/psychological dimension of harm, a finding that contributes to our growing understanding of social and moral judgment. This line of work has implications for the law as well: While tort law recognizes both physical and psychological harm, the legal system typically devalues psychological harm in relation to physical harm (Bornstein & Schwartz, 2009; Vallano, 2013). More research in this area may help highlight and address discrepancies between how the legal system and how people (potential jurors) treat and differentiate between psychological harm and physical harm (Eggen & Laury, 2011).

Acknowledgements

The authors would like to thank Emily Wasserman, Jordan Theriault and other members of the Morality Lab for feedback and comments.

Funding

This study was supported by National Science Foundation Graduate Research Fellowships awarded to L.T. and J.D. (#1258923) and the Alfred P. Sloan Foundation.

Supplementary data

Supplementary data are available at *SCAN* online.

Conflict of interest. None declared.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (text rev.). 4th edn.

Anzellotti, S., Fairhall, S.L., Caramazza, A. (2014). Decoding representations of face identity that are tolerant to rotation. *Cerebral Cortex*, **24**(8), 1988–95.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., Clubley, E. (2001). The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, **31**(1), 5–17.

Barrett, H.C., Bolyanatz, A., Crittenden, A.N., *et al.* (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, **113**(17), 4688–93.

Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1–48.

Bornstein, B.H., Schwartz, S.L. (2009). Injured body, injured mind: Dealing with damages for psychological harm. *The Jury Expert*, **21**(3), 32–9.

Bzdok, D., Schilbach, L., Vogeley, K., *et al.* (2012). Parsing the neural correlates of moral cognition: aLE meta-analysis on morality, theory of mind, and empathy. *Brain Structure and Function*, **217**(4), 783–96.

Chakroff, A., Dungan, J., Koster-Hale, J., Brown, A., Saxe, R., Young, L. (2016). When minds matter for moral judgment: intent information is neurally encoded for harmful but not impure acts. *Social Cognitive and Affective Neuroscience*, **11**(3), 476–84.

Chiu, Y.-C., Esterman, M., Han, Y., Rosen, H., Yantis, S. (2011). Decoding task-based attentional modulation during face categorization. *Journal of Cognitive Neuroscience*, **23**(5), 1198–204.

Christensen, R.H.B. (2015). *ordinal—regression models for ordinal data*. R package version 2015.6-28. https://cran.r-project.org/package=ordinal.

Decety, J., Michalska, K.J., Kinzler, K.D. (2012). The contribution of emotion and cognition to moral sensitivity: a neurodevelopmental study. *Cerebral Cortex*, **22**(1), 209–20.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., Saxe, R. (2011). fMRI item analysis in a theory of mind task. *NeuroImage*, **55**(2), 705–12.

Dufour, N., Redcay, E., Young, L., *et al.* (2013). Similar brain activation during false belief tasks in a large sample of adults with and without autism. *PLoS One*, **8**(9), e75468.

Dungan, J.A., Stepanovic, M., Young, L. (2016). Theory of mind for processing unexpected events across contexts. *Social Cognitive and Affective Neuroscience*, **11**(8), 1183–92.

Eggen, J.M., Laury, E.J. (2011). Toward a neuroscience model of tort law: how functional neuroimaging will transform tort doctrine. *Columbia Science and Technology Law Review*, **13**(2012), 73.

Eickhoff, S.B., Stephan, K.E., Mohlberg, H., *et al.* (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, **25**(4), 1325–35.

Eisenberger, N.I. (2012). The pain of social disconnection: Examining the shared neural underpinnings of physical and social pain. *Nature Reviews Neuroscience*, **13**(6), 421–34.

Eisenberger, N.I. (2015). Social Pain and the brain: controversies, questions, and where to go from here. *Annual Review of Psychology*, **66**(1), 601–29.

Eisenberger, N.I., Inagaki, T.K., Muscatell, K.A., Byrne Haltom, K.E., Leary, M.R. (2011). The neural sociometer: brain mechanisms underlying state self-esteem. *Journal of Cognitive Neuroscience*, **23**(11), 3448–55.

Eisenberger, N.I., Lieberman, M.D., Williams, K.D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, **302**(5643), 290–2.

Fletcher, P.C., Happé, F., Frith, U., *et al.* (1995). Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition*, **57**(2), 109–28.

Gallagher, H.L., Happé, F., Brunswick, N., Fletcher, P.C., Frith, U., Frith, C.D. (2000). Reading the mind in cartoons and stories: an fMRI study of "theory of mind" in verbal and nonverbal tasks. *Neuropsychologia*, **38**(1), 11–21.

Gobbini, M.I., Koralek, A.C., Bryan, R.E., Montgomery, K.J., Haxby, J.V. (2007). Two takes on the social brain: a comparison of theory of mind tasks. *Journal of Cognitive Neuroscience*, **19**(11), 1803–14.

Graham, J., Nosek, B.A., Haidt, J., Iyer, R., Koleva, S., Ditto, P.H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, **101**(2), 366–85.

Gray, K., Schein, C., Ward, A.F. (2014). The myth of harmless wrongs in moral cognition: automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, **143**(4), 1600–15.

Gray, K., Young, L., Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychological Inquiry*, **23**(2), 101–24.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, **293**(5537), 2105–8.

Judd, C.M., Westfall, J., Kenny, D.A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, **103**(1), 54–69.

Kaul, C., Rees, G., Ishai, A. (2011). The gender of face stimuli is represented in multiple regions in the human brain. *Frontiers in Human Neuroscience*, **4**, 238.

Koster-Hale, J., Saxe, R., Dungan, J., Young, L.L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences*, **110**(14), 5648–53.

Kross, E., Berman, M.G., Mischel, W., Smith, E.E., Wager, T.D. (2011). Social rejection shares somatosensory representations with physical pain. *Proceedings of the National Academy of Sciences*, **108**(15), 6270–75.

Kross, E., Egner, T., Ochsner, K., Hirsch, J., Downey, G. (2007). Neural dynamics of rejection sensitivity. *Journal of Cognitive Neuroscience*, **19**(6), 945–56.

Lord, C., Risi, S., Lambrecht, L., *et al.* (2000). The autism diagnostic observation schedule—generic: a standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, **30**(3), 205–23.

MacDonald, G., Leary, M.R. (2005). Why does social exclusion hurt? The relationship between social and physical pain. *Psychological Bulletin*, **131**(2), 202–23.

Molenberghs, P., Gapp, J., Wang, B., Louis, W.R., Decety, J. (2016a). Increased moral sensitivity for outgroup perpetrators harming ingroup members. *Cerebral Cortex*, **26**(1), 225–33.

Molenberghs, P., Johnson, H., Henry, J.D., Mattingley, J.B. (2016b). Understanding the minds of others: a neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews*, **65**, 276–91.

Moran, J.M., Young, L.L., Saxe, R., *et al.* (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences*, **108**(7), 2688–92.

Ratner, K.G., Kaul, C., Van Bavel, J.J. (2013). Is race erased? Decoding race from patterns of neural activity when skin color is not diagnostic of group boundaries. *Social Cognitive and Affective Neuroscience*, **8**(7), 750–5.

Saxe, R., Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind.". *NeuroImage*, **19**(4), 1835–42.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, **42**, 9–34.

Shweder, R.A., Much, N.C., Mahapatra, M., Park, L. (1997). The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. In: Brandt A. M. Rozin P., editors. *Morality and Health* (pp. 119–169). Florence, KY: Taylor & Frances/Routledge.

Tsoi, L., Dungan, J., Waytz, A., Young, L. (2016). Distinct neural patterns of social cognition for cooperation versus competition. *NeuroImage*, **137**, 86–96.

Vallano, J.P. (2013). Psychological injuries and legal decision making in civil cases: what we know and what we do not know. *Psychological Injury and Law*, **6**(2), 99–112.

Wasserman, E.A., Chakroff, A., Saxe, R., Young, L. (2017). Illuminating the conceptual structure of the space of moral violations with searchlight representational similarity analysis. *NeuroImage*, **159**, 371–87.

Woo, C.-W., Koban, L., Kross, E., *et al*. (2014). Separate neural representations for physical pain and social rejection. *Nature Communications*, **5**, 5380.

Young, L., Cushman, F., Hauser, M., Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, **104**(20), 8235–40.

Young, L., Camprodon, J.A., Hauser, M., Pascual-Leone, A., Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, **107**(15), 6753–8.

Young, L., Tsoi, L. (2013). When mental states matter, when they don't, and what that means for morality: when mental states matter for morality. *Social and Personality Psychology Compass*, **7**(8), 585–604.