Thinking in Patterns: using multi-voxel pattern analyses to find neural correlates of moral judgment in neurotypical and ASD populations

Jorie Koster-Hale[1], James Dungan[2], Rebecca Saxe[1], Liane Young[2]

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge MA 02139
[2]Department of Psychology, Boston College, Chestnut Hill, MA 02467

Abstract

Intentional harms are typically judged to be less forgivable than accidental harms. This difference depends on mental state reasoning (i.e., reasoning about beliefs and intentions), supported by a group of brain regions, the 'theory of mind' network. Prior research has found that (i) interfering with activity in this network can shift moral judgments away from reliance on mental state information, and (ii) high-functioning individuals with Autism Spectrum Disorder (ASD) rely significantly less on mental state information to make moral judgments than matched neurotypical (NT) participants. Across three experiments, we find using multi-voxel pattern analysis (MVPA) that, in NT adults, (i) one key region of the ToM network, the RTPJ, shows reliable and distinct spatial patterns of responses across voxels for intentional versus accidental harms, and (ii) individual differences in this neural pattern predict individual differences in moral judgment. By contrast, (iii) in ASD adults, the difference between intentional and accidental harms is not encoded in the voxel-wise pattern in the RTPJ or any other region, and (iv) higher symptom severity scores are predictive of diminished pattern discriminability. We conclude that MVPA can detect features of mental state representations and that these features are behaviorally and clinically relevant.

Keywords: morality, harms, theory of mind, autism, fMRI, multivoxel pattern analysis (MVPA)

Introduction

Intentional harms are usually judged to be morally worse than the same harms caused by accident (Cushman, 2008; Knobe, 2005; Malle & Knobe, 1997; Piaget, 1965; Singer et al, 2004; Young & Saxe, 2011). The capacity to distinguish between intentional and accidental harms depends on the capacity to represent another person's mental states, a cognitive function associated with a specific and selective group of brain regions (the 'theory of mind network'). Prior research has revealed that moral judgments of harmful actions depend on one region in particular, the right temporo-parietal junction (RTPJ). For example, individual differences in moral judgments of accidental harms are correlated with RTPJ activity (Young & Saxe, 2009), and disrupting RTPJ activity interferes with these judgments (Young et al, 2010).

Recent evidence suggests that moral judgments may provide a sensitive measure of enduring impairments in ToM in high-functioning individuals with Autism Spectrum Disorders (ASD). Children with ASD are disproportionately impaired on tasks that require them to consider the beliefs and intentions of other people (Peterson et al, 2005; Baron-Cohen, 1995). Although children with ASD distinguish between moral and conventional transgressions (Blair et al, 1996), and between good actions and bad actions (Leslie et al, 2006), they are delayed in using information about innocent intentions to forgive accidents (Grant et al, 2005). In a recent study, Moran et al. (2011) found that high-functioning adults with ASD show the same pattern, delivering less forgiveness and more blame for accidents than neurotypical (NT) adults.

These findings suggest that the RTPJ should encode the difference between accidental and intentional harm. Puzzlingly, however, we find that the average mean signal – a standard way of measuring neural involvement in a task – in RTPJ and the other theory of mind regions is not different for intentional versus accidental harmful actions in NT participants. One possibility is that this key dimension can be detected, not in the magnitude of response across a brain region, but in the pattern of responses across voxels.

A complementary approach to traditional neuroimaging analyses (which rely on average response magnitude) is to look at the pattern of activity across voxels within a region, using a technique called multi-voxel pattern analysis. If a different tasks, stimulus categories, or stimulus features are processed by (partially) different subpopulations of neurons within a brain region, the difference may not be detectable in the region's average response, but may nevertheless produce systematic and distinct patterns of activity across neighboring voxels within the region (Normal et al, 2006; Haynes et al, 2006; Kriegeskorte & Bandetti, 2007). A key advantage of this technique is that these patterns can be used to 'decode' information from the neural response not otherwise detectable in the overall magnitude (e.g., object category in ventral temporal regions, Haxby et al 2001). Thus MVPA can reveal how stimulus categories are processed within a functional region (Peelen et al, 2006; Haynes & Rees, 2006).

Given the importance of intentions for moral judgments of accidental and intentional harms, we predicted that one or more brain regions in the ToM network would explicitly encode this feature of others' mental states, in neurotypical (NT) adults. That is, we predicted that (i) while participants read about a wide range of harmful acts, we would be able to decode whether the described harm was intentional or accidental based on the spatial pattern of activity within ToM brain regions. We tested this prediction in three experiments. We also investigated (ii) whether the robustness of the spatial pattern information within individuals would predict those individuals' moral judgments, and (iii) whether high-functioning adults with ASD, who make atypical moral judgments of accidental harms, would show correspondingly atypical patterns of neural activity.

Methods

Participants

Experiment 1: Sixteen right-handed members of the MIT community (aged 18-50, 7 women). **Experiment 2:** Eighteen

right-handed college undergraduate students (aged 18-25 years, 8 women). **Experiment 3:** Fourteen right-handed college undergraduate students (aged 18-25 years, 8 women).

Experiment 4: Twelve individuals diagnosed with Autism Spectrum Disorder (aged 25-43 years, 2 women). Participants in were recruited via advertisements placed with the Asperger's Association of New England. All participants were prescreened using the Autism Quotient questionnaire (AQ; Baron-Cohen et al., 2001). ASD participants then underwent both the Autism Diagnostic Observation Schedule (ADOS) (Lord et al 2000, 2002) and impression by a clinician trained in both ADOS administration and diagnosis of ASD. All ASD participants received a diagnosis of an Autism Spectrum Disorder based on their social ADOS score (6.2 ± 0.6), communication ADOS score (3.5 ± 0.4), and total ADOS score (9.6 ± 0.8) and on clinical impression based upon the diagnostic criteria of the DSM-IV (APA; 2000). The matched NT (Exp. 1) and ASD (Exp. 4) groups did not differ in age (NT (mean ± SEM)=27.1 ± 2.3; ASD=31.8 ± 2.1; $t(28)=1.4$, $p > 0.17$] or IQ [NT: 118.1 ± 2.8; ASD: 121.0 ± 3.8; $t(27)=0.60$, $p > 0.55$). No participant had higher language than social scores, suggesting no specific language deficits.

All participants participated for payment, were native English speakers, had normal or corrected-to-normal vision and gave written informed consent in accordance with the requirements of Institutional Review Board at MIT. Data from Experiments 2 and 3 have previously been published, analyzing the magnitude but not the pattern of response in each region, in Young et al. (2008) and Young & Saxe (2009).

fMRI Protocol and Task

Experiment 1 & 4: Participants were scanned while reading 60 stories: 12 intentional harm violations, 12 accidental harm violations, 24 stories with other types of moral violations, and 12 neutral scenarios. Stories were presented in the second person, using present tense, and displayed in four cumulative segments: 1. Background (6s), 2. Action (4s), 3. Outcome (4s), 4. Intent: Good (Accidental Harm) or Bad (Intentional Harm) (4s). In the scanner, after each story, participants made moral judgments of the action from "not at all morally wrong" (1) to "very morally wrong" (4), using a button press. For example stimuli, see http://mit.edu/jorie/www/CogSci2012/Koster-Hale_CogSci2012_supp.pdf.

Stories were presented in a pseudorandom order; condition order was counterbalanced across runs and subjects, and no condition was immediately repeated. Participants never saw both intentional and accidental versions of the same scenario. Word count was matched across conditions. Ten stories were presented in each 5.5 min run; the total experiment, six runs, lasted 33.2 min. Rest blocks of 10 s were interleaved between each story. Stories were projected onto a screen via Matlab 5.0 running on an Apple MacBook Pro in 40-point white font.

Experiments 2 & 3: Participants were scanned while reading 48 stories. Experiment 2 included 12 intentional harms, 12 accidental harms, and 24 non-harm stories. All harms were physical harms, resulting in someone's death. Stories were presented in cumulative segments: 1. Background (6s) 2. Foreshadow (6s, only in Experiment 2), 3. Intent: Good (Accidental Harm) or Bad (Intentional Harm) (6 s), 4. Outcome (6s). Half of the stories in each run were presented with foreshadow before intent; the order was reversed in the other half. After each story, participants made moral judgments of the action on a 3-point scale, from "forbidden (1) to "permissible" (3), using a button press.

Experiment 3 included 8 intentional harms, 8 accidental harms, and 32 other non-harm stories. Participants delivered a non-moral judgment, answering a true/false question about the content of the final sentence.

Stories were counter balanced and matched as in Experiment 1. Rest blocks of 14 s were interleaved between each story. Stories were projected onto a screen via Matlab 5.0 running on an Apple G4 laptop in 24-point white font.

Theory of Mind Localizer task: In all four experiments, participants also saw 4 runs of a theory of mind localizer task, contrasting stories requiring inferences about mental state representations (e.g., thoughts, beliefs) versus physical representations (e.g., maps, signs, photographs), which are similar in their meta-representational and logical complexity but differ in whether the reader is building a representation of someone else's mental state. See Saxe & Kanwisher (2003) and Dodell-Feder et al (2010) for further discussion; stimuli and presentation from Saxe & Kanwisher 2003, Exp. 2.

Acquisition and Preprocessing

FMRI data were collected in a 3T Siemens scanner at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, using a 12-channel head coil. Using standard echoplanar imaging procedures, we acquired blood oxygen level dependent (BOLD) data in 26 near axial slices using 3x3x4 mm voxels (TR=2 s, TE=40 ms, flip angle=90◦). To allow for steady state magnetization, the first 4 seconds of each run were excluded.

Data processing and analysis were performed using SPM8 (Experiments 1 & 4) and SPM2 (Experiments 2 & 3) and custom software. The data were motion corrected, realigned, normalized onto a common brain space (MNI template), spatially smoothed using a Gaussian filter (FWHM 5 mm kernel) and subjected to a high-pass filter (128 Hz).

fMRI Analysis

All fMRI data were modeled using a boxcar regressor, convolved with a standard hemodynamic response function (HRF). The general linear model was used to analyze the BOLD data from each subject, as a function of condition. The model included nuisance covariates for run effects, global mean signal, and an intercept term. A slow event-related design was used. An event was defined as a single story, the event onset was defined by the onset of text on screen, and offset as the end of the story presentation.

Functional Localizer: Individual ROIs Based on prior research, functional regions of interest (ROIs) were defined in right and left temporo-parietal junction (RTPJ, LTPJ), medial precuneus (PC), and dorsal medial prefrontal cortex (DMPFC), for each participant. Using a pre-defined

hypothesis space for each ROI, each subject's contrast image (Belief > Photo) was masked and the peak voxel that occurred in a cluster of 10 or more voxels significant at $p < 0.001$ was selected. All voxels contiguous with the peak voxel, individually significant at $p < 0.001$, within a 9mm radius, were defined as the ROI.

Group ROIs: To compare NT participants from Experiment 1 and ASD participants from Experiment 4, we also identified independent group-level ROIs, using data from a previous set of theory of mind localizers (n=477 NT participants, 260 women). Pattern analyses for both groups were then conducted within these group ROIs. Selecting the same voxels across participants and group ensured that any differences found between the NT and ASD group are due to differences in the pattern itself, rather than any differences in the ROI selection method, across populations.

Within-ROI Magnitude Analysis: We measured the response to each condition in each ROI. The percent signal change (PSC) relative to baseline was calculated for each time point in each condition, averaging across all voxels in the ROI and across all blocks in the condition, where PSC(t) =100 × (average BOLD magnitude for condition (t) – average BOLD magnitude for fixation) / average BOLD magnitude for fixation. We averaged the PSC across the entire presentation – offset 6s from presentation time to account for hemodynamic lag – to get a single PSC for each condition, in each ROI, in each participant (Poldrack, 2006).

Within-ROI Pattern Analysis: In all experiments, we conducted within-ROI pattern analyses. Following Haxby et al. (2001), each participant's data were divided into even and odd runs ('partitions') and then the mean response (beta value) of every voxel in the ROI was calculated for each condition. The "pattern" of response was the vector of beta values across voxels within the ROI. To determine the within-condition correlation, the pattern in one (e.g., even) partition was compared to the pattern for the same condition in the opposite (e.g., odd) partition; to determine the across-condition correlations the pattern was compared to the opposite condition, across partitions.

For each individual, an index of classification was calculated for each condition pair as the z-scored within-condition correlation minus the z-scored across-condition correlation. A region successfully classified a category of stimuli if, across individuals, the within-condition correlation was higher than the across-condition correlation, using a Student's T complementary cumulative distribution function.

Results

Localizer

Replicating many studies using a similar functional localizer task (e.g., Saxe & Kanwisher, 2003), we localized four theory of mind brain regions showing greater activation for false belief stories compared to false photograph stories in the majority of participants (uncorrected, $p < 0.001$, $k > 10$): **Exp 1-3 (NT):** RTPJ 46/48, LTPJ, 44/48, PC 47/48, DMPFC 41/68; **Exp 4 (ASD):** RTPJ (12/12 participants), LTPJ (12/12), PC (11/12) and DMPFC (5/12).

Behavioral Results

Experiment 1: Participants judged intentional harms (3.31 ± 0.10) to be worse than accidental harms (1.62 ± 0.11; $t(14) = 15.1$, $p<0.0001$), both of which were judged to be worse than neutral stories (1.03 ± .02, $t(14)=18.1$, $p<0.001$).

Experiment 2: Replicating the results in Experiment 1, participants judged intentional harms (2.9 ± .03) to be worse than accidental harms (1.9 ± .11; $t(12)=8.24$, $p<0.0001$).

Experiment 4: Behavioral data were available for only 7 participants with ASD (remaining data were lost due to a coding error and due to theft of experimental equipment). When making moral judgments, ASD participants, like NT participants from Experiment 1, ASD participants judged intentional harms (3.5 ± 0.12) to be worse than accidental harms (1.85±0.21; $t(6)=8.9$; $p <0.0001$), both of which were judged to be worse than neutral stories (1.09 ± .04; $t(6) = 13.9$; $p<0.0001$).

Group Comparison: A mixed effects ANOVA crossing Group (NT in Exp 1, ASD in Exp 4) by Condition (neutral, accidental, intentional) yielded a main effect of condition and no interaction ($F(2,36) = 284.4$, $p < 0.0001$). Post-hoc t-tests revealed that ASD adults assign more blame for accidental harms than NT adults ($t(19)=1.7$, $p<0.05$), but there was no difference between NT and ASD judgments of intentional harms (Moran et al., 2011).

fMRI - Magnitude Analysis

Experiment 1-4 In all three experiments with NT adults, and in the final experiment with ASD adults, all four ROIs (RTPJ, LTPJ, PC, DMPFC) showed a higher BOLD response for moral violations than for neutral acts. However, none of the regions showed a significant difference between accidental and intentional harms (all $p>.2$).

fMRI - Pattern Analysis

Experiment 1 *Moral vs. Neutral:* Multi-voxel pattern analyses revealed reliably distinct patterns of neural activity for moral violations versus neutral acts in all four ROIs: RTPJ, LTPJ, PC, DMPFC. The pattern generated by stories within a category (moral or neural) were more correlated with other stories in the same category compared to stories in the opposite category (RTPJ: across condition correlation=0.90 (.1), within condition correlation=1.16(.1), $t(15)=2.6$, $p=0.02$; LTPJ: across=1.5(.07), within=1.6(.06), $t(14)=2.3$ $p=0.019$; PC: across=0.86(0.13) within=1.2(0.12), $t(14)=3.1$, $p=0.005$; DMPFC: across=1.1(0.13), within=1.2(0.11), $t(12)=1.9$, $p=0.04$). Group ROIs yielded the same results.

Accidental vs. Intentional: In only the RTPJ, the pattern of response distinguished between accidental and intentional harms (across condition correlation=0.91(.1), within condition correlation=1.08(.11), $t(15)=2.6$, $p=0.01$). No other regions showed distinct patterns of response to intentional versus accidental harms (all correlation differences < 0.1, all $p > 0.1$). Group ROIs yielded the same results, (**Figure 1-A**).

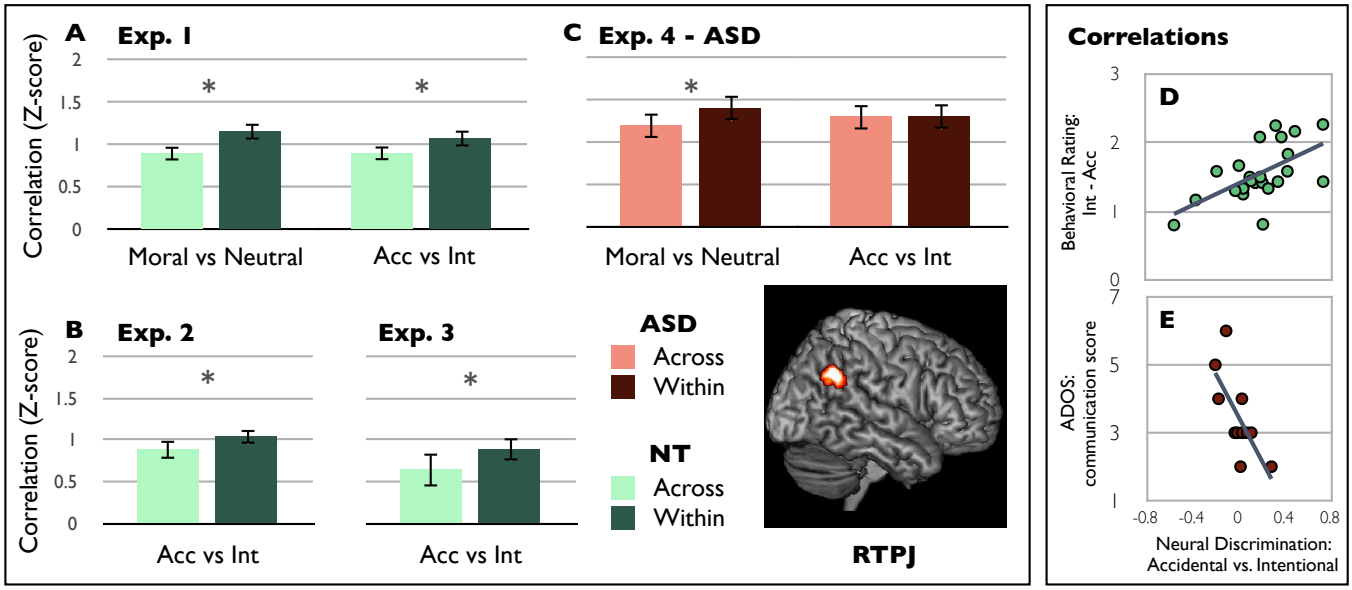

Figure 1: (**A**) MVPA results from Experiment 1: neurotypical adults (n=16) show pattern discrimination for moral vs. neutral actions, and for accidental vs. intentional harms in RTPJ, (**B**) a finding replicated across Experiments 2 and 3 (n=18,14). (**C**) Adults with ASD (n=12) show neutral discrimination of moral vs. neutral actions in RTPJ, but not of intentional vs. accidental harms. (**D**) Individual differences in pattern classification predict individual differences in behavior in NT adults (n=23). (**E**) In ASD adults (n=12), symptom severity is negatively correlated with neural discrimination of accidental and intentional harms.

Experiments 2 & 3: Experiments 2 and 3 replicate Experiment 1. In RTPJ, but no other region, MVPA analyses revealed that accidental harm and intentional harm elicited reliably distinct neural patterns (Exp 2 RTPJ: across condition correlation=0.76(.13), within=1.1(.11), t(15)=2.6, p=0.01; Exp 3 RTPJ: across=0.65(.21), within=0.89(.14), t(13)=1.8, p=0.04; all other regions: correlation differences <0.1, p>0.1).

Combining Experiments 1, 2, & 3: Pooling the data across all three experiments allowed us to increase our power to detect results in neural regions beyond RTPJ. Again, MVPA analyses revealed distinct neural patterns for accidental and intentional harms in RTPJ (RTPJ: across=0.78(0.09), within=1.0(0.08), t(45)=3.9, p=0.0002) and no other region (all differences < 0.1, p > 0.1, Figure 2); repeated measures ANOVAs crossing Region by Pattern yielded a significant interaction between RTPJ and each of the three other regions (all F > 8, all p < 0.007), and no interactions between any of the other regions (however, see Smith et al., 2011 for caution in interpreting differences in discriminability across regions).

Behavioral and Neural Correlation: In Experiments 1 and 2, NT participants provided moral judgments of each scenario in the scanner, allowing us to determine whether behavioral responses were related to the spatial pattern of the neural response in RTPJ or any other region. For each participant, we calculated the difference between moral judgments for intentional versus accidental harms. We tested whether this difference score was correlated, across participants, with the index of classification in each region (intentional vs. accidental, within-condition correlation minus across-condition correlation). In both experiments we found that only in the RTPJ was the difference between intentional and accidental harms in individuals' moral judgments correlated with the neural classification index (Exp 1: $r^2(12)=0.38$; p=0.03; Exp 2: $r^2(11)=0.40$; p=0.03). The correlation was also significant after combining the data from both experiments ($r^2(23)=0.35$, p=0.003; **Figure 1-D**).

Experiment 4 (ASD) *Moral vs. Neutral:* As in NT controls, pattern analyses revealed a separation in the pattern of response for moral violations and neutral scenarios in ASD adults. Using individual ROIs, significant discrimination was found in RTPJ and LTPJ (RTPJ: across=1.2(0.18), within=1.4 (0.16), t(11)=3.2, p=0.005; LTPJ: across=1.3(0.12), within=1.5(0.11), t(11)=2.3, p=0.02; see Figure 1); there was a trend in the same direction in the PC. DMPFC was found in only 5 of 12 individuals so we did not have sufficient power to test for pattern discriminability. However, using group ROIs, we found significant discrimination in DMPFC as well (across=1(0.12), within=1.2(0.10), t(11)=3, p=0.007).

Accidental vs. Intentional: Pattern analysis within both group and individual ROIs revealed no pattern discrimination between accidental and intentional harms in any ToM region, in participants with ASD (all differences < 0.1, p > 0.1).

Group Comparison *Moral vs. Neutral:* A Group (ASD, NT) x Pattern (within, across) ANOVA revealed that NT and ASD participants show equally robust neural discrimination in response to moral violations versus neutral actions in their RTPJ, with a main effect of Pattern (F(2,51) = 12.4, p = .002), no effect of Group (F(1,51)=1.07, p=.3), and no interaction (F (2,51) = .13, p = .7).

Accidental vs. Intentional: In contrast, NT participants discriminated between accidental and intentional harms to a greater extent than ASD participants, reflected in a significant

Group x Pattern interaction ($F(2,51) = 5.1$, $p = .03$), and no main effects (**Figure 1-C**).

Symptom Severity and Neural Correlation: In ASD participants, we found no significant correlation between neural pattern and behavior in any region. However, we found, in RTPJ and no other region, a significant inverse correlation with symptom severity: individuals with higher ADOS (Lord et al, 2000, 2002) scores showed less neural discriminability between intentional and accidental harms ($r^2(12)=0.51$, $p=0.01$, **Figure 1-E**). No symptom severity score correlated with moral versus neutral discrimination.

Discussion and Conclusion

Moral Judgments: Neurotypical adults

A central aim of this study was to ask whether the difference between accidental and intentional harms could be decoded from the pattern of response within theory of mind brain regions. Across three experiments with neurotypical (NT) adults, using different stimuli, paradigms, and participants, we found converging results: stories about intentional versus accidental harms elicited spatially distinct patterns of response within the right temporo-parietal junction (RTPJ). Moreover, this neural response mirrored behavioral judgments: individuals who showed more distinct patterns in the RTPJ also made a larger distinction between intentional and accidental harms in their moral judgments.

The convergence across experiments provides strong evidence that intentional and accidental harms can be discriminated, using MVPA, in RTPJ. Designed to test a series of separate questions, the three experiments differed in the story content, voice of the narrative (2nd or 3rd person), the order of information provided, the length of the stories, the number of stories per condition, and the participants' explicit task. Perhaps most importantly, the information indicating that the harmful action was accidental or intentional was provided by different cues. In Experiment 1, the same mental state content (e.g., your cousin's allergy to peanuts) was described as known or unknown (e.g., "you had no idea" vs. "you definitely knew"). By contrast, in Experiments 2 and 3, sentences with the same syntax and mental state verbs were used to describe beliefs with different content (e.g., "Steve believes the ground beef is safe / rotton"). Nevertheless, the spatial pattern of response was reliable and distinct for intentional versus accidental harms, and only in the right TPJ. The generalizability of the pattern discriminability indicates that, rather than being driven by specific stimulus features or task demands, the discriminable neural patterns reflect an underlying distinction in the representation of accidental and intentional hams.

In Experiments 1 and 2, participants made moral judgments in the scanner. In both experiments, individuals differed in the amount of blame they assign to accidental harm, some weighing intent more strongly (and thus were more forgiving) and some weighing outcome more strongly (and thus were more condemning), (Young & Saxe, 2009). These individual differences in moral judgment were predicted by individual differences in pattern discriminability in the RTPJ. While Experiment 1 used a blameworthiness scale ("How much blame should you get?") and Experiment 2 used a permissibility scale ("How permissible was Steve's action?"), we found the same result in both studies: individuals who showed more sensitivity to the dimension of intent in their neural pattern – those who processed accidental and intentional harms most differently in their RTPJ – were also those who showed the most forgiveness to characters who accidentally harmed someone.

As in prior work, the average magnitude of RTPJ response did not distinguish between intentional and accidental harms. This observation fits in a larger pattern emerging in the literature: while the RTPJ is selective for the cognitive process of mental state reasoning – and not, for example, generic attentional processes (Scholz et al., 2008; Young et al., 2010b; see also Decety and Lamm, 2007) – the average RTPJ response is unaffected by changes in the specific features of mental states, such as whether beliefs are true or false (Jenkins and Mitchell, 2009; Young, et al, 2010b), justified or unjustified (Young et al., 2010c), positively or negatively valenced (Kliemann et al., 2008), plausible or crazy (Young et al., 2010b), "constrained" or "open-ended" (Jenkins and Mitchell, 2009), attributed to friends or enemies (Bruneau & Saxe in press), or first-order or higher-order (Koster-Hale & Saxe, 2011).

These findings left open the question of whether and how the RTPJ or any other neural substrate encoded specific mental state features, like the dimension of intent. A key contribution of the current study then is to reveal that the dimension of intent is encoded in the voxel-wise pattern of the RTPJ, and specifically for the evaluation of harm.

Moral Judgments: ASD adults

A group of high functioning adults with Autism Spectrum Disorders showed a different response profile: the response of the RTPJ showed a reliable spatial pattern of response across moral stories, compared to neutral stories, but did not distinguish between intentional and accidental harms. Moreover, we found that symptom severity on the ADOS was predictive of decreased neural discrimination in RTPJ: those individuals with more severe diagnoses showed a less distinct neural response to accidental and intentional harms. Thus, the neural pattern mirrored the behavioral performance previously observed in participants with ASD (Moran et al 2011): compared to NT controls, participants with ASD judged accidental and intentional harms to be more similar in moral permissibility (though note that, in the current sample, the group by condition interaction was not replicated, likely due to lack of power).

One possible mechanism of reduced pattern information in ASD might be more noisy or heterogeneous neural responses. However, both the strong discrimination between moral violations and neutral stories, and the high overall pattern correlations speak against this alternative. Rather ASD participants seem to show a less sensitive neural response: accidental and intentional harms appear to be processed by the same neural sub-populations within the RTPJ.

Multivoxel pattern analysis may therefore be a successful way of measuring behaviorally-relevant neural differences in ASD. Note that due to the demands of the task and scanning environment, the ASD participants in this study (as in previous task-oriented neuroimaging studies) are extremely high functioning, which may limit the generalizability of the results to lower-functioning individuals. Nevertheless, the individuals in the current study do have disproportionate difficulties with social interaction and communication, and the current results may provide a window on the neural mechanism underlying these difficulties.

Conclusion

In summary, MVPA allows us to determine (i) that features of mental state representations that are not observable in the mean neural signal, including the behaviorally relevant difference between accidental and intentional harms, are encoded in the *pattern* of neural activity; (ii) that these mental state features elicit both stable and distinct patterns of neural activity in RTPJ, a region implicated in mental state reasoning; (iii) that individual differences in neural discrimination predict individual differences in moral judgment; and (iv) that atypical behavioral patterns in ASD are reflected in atypical neural patterns, which (v) are more atypical with increasing symptom severity.

Acknowledgments

This material is based upon work supported by the Simons Foundation, the National Science Foundation under Grant 095518, a John Merck Scholars Grant, and a National Science Foundation Graduate Research Fellowship, Grant 0645960.

References

Baron-Cohen S. (1995) Mindblindness: an essay on autism & theory of mind. MIT Press, Cambridge MA.

Blair J. (1996) Brief Report: Morality in the Autistic Child. *Journal of Autism & Developmental Disorders*. 26(5):571-579.

Castelli F, Frith C, Happé F, & Frith U. (2002) Autism, Asperger syndrome & brain mechanisms for the attribution of mental states to animated shapes. *Brain*. 125:1839-49

Cushman, F. (2008). Crime & Punishment: Distinguishing the roles of causal & intentional analysis in moral judgment. *Cognition*.

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

Dodell-Feder, D., Koster-Hale, J, Bedny M, & Saxe, RR. (2011), fMRI item analysis in a theory of mind task, *NeuroImage.*

Grant CM, Boucher J, Riggs KJ, & Grayson A. (2005). Moral understanding in children with autism. *Autism* 9(3):317-331.

Lord C, Rutter M, DiLavore PC, Risi S (2002) Autism Diagnostic Observation Schedule (Western Psychological Services, Los Angeles)

Happé F, Ronald A, & Plomin R. (2006) Time to give up on a single explanation for autism. *Nature Neuroscience* 9(10):1218-1220.

Haxby JV, Gobbini MI, Furey ML, Ishai, A, Schouten JL, & Pietrini P. (2001). Distributed & Overlapping Representations of Faces & Objects in Ventral Temporal Cortex. *Science* 293(5539):2425-30.

Haynes, JD & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7), pp.523–534.

Jenkins, A. C., & Mitchell, J. P. (2009). Mentalizing under uncertainty: Dissociated neural responses to ambiguous & unambiguous mental state inferences. *Cerebral Cortex*, 20(2), 404–410.

Kennedy DP & Courchesne E. (2008). Functional abnormalities of the default network during self- & other-reflection in autism. *Social Cognitive & Affective Neuroscience* 3(2): 177-190.

Koster-Hale, J & Saxe, R. R. (2010). theory of mind brain regions are sensitive to the content, not the structural complexity, of belief attributions *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.*

Knobe, J. (2005). theory of mind & Moral Cognition. *Trends in Cognitive Sciences*, 9, 357-359.

Kriegeskorte N & Bandettini P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage.*

Leslie A, Mallon R, & DiCorcia J. (2006). Transgressors, victims, & cry babies: Is basic moral judgment spared in autism? *Social Neuroscience* 1(3-4):270-283.

Lombardo MV, Chakrabarti B, Bullmore ET, MRC AIMS Consortium, & Baron-Cohen S. (2011). Specialization of right temporo-parietal junction for mentalizing & its relation to social impairments in autism. *NeuroImage* 56(3):1832-8.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental & Social Psychology*, 33, 101–121.

Mason RA, Williams DL, Kana RK, Minshew N, & Just MA. (2008). theory of mind disruption & recruitment of the right hemisphere during narrative comprehension in autism. *Neuropsychologia* 46(1).

Moran J, Young L, Saxe R, Lee SM, O'Young D, Mavros P, & Gabrieli J. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *In PNAS* 108(7):2688-92.

Peelen, M.V., Wiggett, A.J. & Downing, P.E., (2006). Patterns of fMRI Activity Dissociate Overlapping Functional Brain Areas that Respond to Biological Motion. *Neuron*, 49(6), pp.815–822.

Peterson C, Wellman H, & Liu D. (2005). Steps in theory-of-mind development for children with deafness or autism. *Child Development* 76(2):502-517.

Piaget, J. (1965). The moral judgment of the child. New York: Free Press.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *Neuroimage*, 19(4), 1835-1842.

Scholz, J., Triantafyllous, C., Whitfield-Gabrieli, S., Brown, E. N., & Saxe, R. (2009). Distinct regions of the right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLoS ONE,* 4(3), 1-7.

Silani G, Bird G, Brindley R, Singer T, Frith C, & Frith U. (2007) Levels of emotional awareness & autism: An fMRI study. *Social Neuroscience* 3(2):97-112.

Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J., & Frith, C. D. (2004). Brain responses to the acquired moral status of faces. *Neuron.*

Tesink CMJY, Buitelaar JK, Petersson KM, van der Gaag RJ, Kan CC, Tendolkar I, & Hagoort P. (2009). Neural correlates of pragmatic language comprehension in autism spectrum disorders. *Brain* 132(7).

Wang AT, Lee SS, Sigman M, & Dapretto M. (2006). Neural basis of irony comprehension in children with autism: the role of prosody & context. *Brain* 129(4):932-43.

Young, L., Saxe, R. (2008). The neural basis of belief encoding & integration in moral judgment. NeuroImage, 40, 1912-1920

Young, L., & Saxe, R. (2009). An FMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience* 21(7), 1396-1405.

Young, L., & Saxe, R. (2009). Innocent intentions: a correlation between forgiveness for accidental harm & neural activity. *Neuropsychologia*.

Young, L., Camprodon, J., 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. *PNAS*, 107, 6753-6758

Young, L., Dodell-Feder, D., & Saxe, R. (2010). What gets the attention of the temporo-parietal junction? An FMRI investigation of attention & theory of mind. *Neuropsychologia.*

Young, L., Nichols, S., Saxe, R. (2010). Investigating the Neural & Cognitive Basis of Moral Luck: It's Not What You Do but What you Know. Review of Philosophy & Psychology, 1, 333-349.

Young, L., Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*,120, 202-214.