

MIT Open Access Articles

*Decoding moral judgments from
neural representations of intentions*

The MIT Faculty has made this article openly available. **Please share**
how this access benefits you. Your story matters.

Citation: Koster-Hale, J., R. Saxe, J. Dungan, and L. L. Young. "Decoding moral judgments from neural representations of intentions." *Proceedings of the National Academy of Sciences* 110, no. 14 (April 2, 2013): 5648-5653.

As Published: <http://dx.doi.org/10.1073/pnas.1207992110>

Publisher: National Academy of Sciences (U.S.)

Persistent URL: <http://hdl.handle.net/1721.1/81292>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of Use: Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.



Decoding moral judgments from neural representations of intentions

Jorie Koster-Hale^{a,1}, Rebecca Saxe^a, James Dungan^b, and Liane L. Young^b

^aMcGovern Institute for Brain Research and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139;

^bDepartment of Psychology, Boston College, Chestnut Hill, MA 02467

Edited by Robert Desimone, Massachusetts Institute of Technology, Cambridge, MA, and approved January 29, 2013 (received for review May 11, 2012)

Intentional harms are typically judged to be morally worse than accidental harms. Distinguishing between intentional harms and accidents depends on the capacity for mental state reasoning (i.e., reasoning about beliefs and intentions), which is supported by a group of brain regions including the right temporo-parietal junction (RTPJ). Prior research has found that interfering with activity in RTPJ can impair mental state reasoning for moral judgment and that high-functioning individuals with autism spectrum disorders make moral judgments based less on intent information than neurotypical participants. Three experiments, using multivoxel pattern analysis, find that (i) in neurotypical adults, the RTPJ shows reliable and distinct spatial patterns of responses across voxels for intentional vs. accidental harms, and (ii) individual differences in this neural pattern predict differences in participants' moral judgments. These effects are specific to RTPJ. By contrast, (iii) this distinction was absent in adults with autism spectrum disorders. We conclude that multivoxel pattern analysis can detect features of mental state representations (e.g., intent), and that the corresponding neural patterns are behaviorally and clinically relevant.

functional MRI | morality | theory of mind

Thinking about another's thoughts increases metabolic activity in a specific group of brain regions. These regions, which comprise the "theory of mind network," include the medial prefrontal cortex (MPFC), precuneus (PC), right superior temporal sulcus (RSTS), and bilateral temporal-parietal junction (TPJ). Although many studies have investigated the selectivity and domain specificity of these brain regions for theory of mind (1, 2), a distinct but fundamental question concerns the computational roles of these regions: which features of people's beliefs and intentions are represented, or made explicit, in these brain regions? Prior work has focused on where in the brain mental state reasoning occurs, whereas the present research builds on this work to investigate how neural populations encode these concepts.

A powerful approach for understanding neural representation in other domains has been to ask which features of a stimulus can be linearly decoded from a population of neurons. For example, in the ventral visual stream (involved in object recognition), low-level stimulus properties like line orientation and shading are linearly decodable from small populations of neurons in early visual areas (e.g., V1), whereas in higher-level regions, the identity of an object becomes linearly decodable and invariant across viewing conditions (3, 4). These results suggest that as information propagates through the ventral pathway, the neural response is reformatted to make features that are relevant to object identity more explicit to the next layer of neurons (3).

A decoding approach can be similarly applied to functional MRI (fMRI) data, using multivoxel pattern analysis (MVPA) to examine the spatial pattern of neural response within a brain region. If a distinction between cognitive tasks, stimulus categories, or stimulus features is coded in the population of neurons within a brain region, and if the subpopulations within the region are (at least partially) organized into spatial clusters or maps over cortex (5, 6), then the target distinction may be detectable in reliable spatial patterns of activity measurable with fMRI (7–9). MVPA has therefore been used to identify categories and

features that are represented within a single region (10–12) and to relate these representations to behavioral performance (7, 9, 13).

Compared with object recognition, much less is known about the cognitive and neural mechanisms that support theory of mind. However, linear separability of the neural response could serve as a diagnostic measure of the core features and local computations even in this abstract domain. We therefore asked whether the spatial pattern of response in theory of mind brain regions could be used to decode a feature that has previously been shown to be critical for theory of mind: whether an action was performed intentionally or accidentally.

The distinction between intentional and accidental acts is particularly salient in the case of moral cognition. Adults typically judge the same harmful act (e.g., putting poison in a drink, failing to help someone who is hurt, making an insensitive remark) to be more morally wrong and more deserving of punishment when committed intentionally vs. accidentally (14). These moral judgments depend on individuals' ability to consider another person's beliefs, intentions, and knowledge, and emerge relatively late in childhood, around age 6–7 y (15). Individuals with autism spectrum disorders (ASD), who are disproportionately impaired on tasks that require them to consider people's beliefs and intentions (16, 17), are also impaired in using information about an innocent intention to forgive someone for accidentally causing harm (18–20, but see ref. 21).

The right TPJ (RTPJ) is particularly implicated in these moral judgments. In prior research, increased RTPJ activation is related to greater consideration of mitigating intentions and more lenient punishment (22, 23); individual differences in the forgiveness of accidental harms are correlated with the magnitude of activity in the RTPJ at the time of the judgment (24); and interfering with activity in the RTPJ shifts moral judgments away from reliance on mental states (25).

Given the importance of intent for moral judgments of harms, we predicted that one or more of the brain regions in the theory of mind 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 range of harmful acts, we would be able to decode whether the harm was intentional or accidental based on the spatial pattern of activity within theory of mind brain regions. We tested this prediction in three experiments with NT adults. We also investigated (ii) whether the robustness of the spatial pattern within individuals would predict those individuals' moral judgments and (iii) whether, in a fourth experiment, high-functioning adults with ASD, who

Author contributions: J.K.-H., R.S., and L.L.Y. designed research; J.K.-H., J.D., and L.L.Y. performed research; J.K.-H. and R.S. contributed new reagents/analytic tools; J.K.-H. and R.S. analyzed data; and J.K.-H., R.S., and L.L.Y. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The neuroimaging data have been deposited with the National Database for Autism Research, <http://ndar.nih.gov> (accession nos. 8415, 8418–8435, 8437–8439, and 8442–8446).

¹To whom correspondence should be addressed. E-mail: jorie@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1207992110/-DCSupplemental.

	Background	Action	Outcome	Intent
Accidental	Your family is over for dinner. You wish to show off your culinary skills. For one of the dishes, adding peanuts will really bring out the flavor.	You grind up some peanuts, add them to that dish, and serve everyone.	Your cousin, one of your dinner guests, is severely allergic to peanuts.	You had absolutely no idea about your cousin's allergy when you added the peanuts.
Intentional				You knew about your cousin's peanut allergy when you added the peanuts to the dish.

Fig. 1. Example stimulus from experiments 1 and 4.

make atypical moral judgments of accidental harms, would show atypical patterns of neural activity in pattern or magnitude.

In all four experiments, participants in the scanner read short narratives in which someone caused harm to another individual, intentionally or accidentally (Fig. 1), as well as narratives involving no harm. Participants in experiments 1 and 2 made a moral judgment about the action. Participants in experiment 3 made true/false judgments about facts from the narratives. In experiment 4, high-functioning adults with ASD read and made moral judgments about the same narratives as in experiment 1.

Results

Behavioral Results. Because participants used the scales in different ways (e.g., some using largely 2–4 and others using largely 1–3), we z-scored the behavioral data. For analyses of untransformed data, see *SI Behavioral Results, Raw Behavioral Responses*. In experiment 1, intentional harms (1.2 ± 0.03) were rated as more blameworthy than accidental harms [-0.38 ± 0.04 ; $t(19) = -25.0$, $P < 0.001$], and both were rated more blameworthy than neutral acts [-0.8 ± 0.03 ; $t(19) = 50$, $P < 0.001$ and $t(19) = 6.9$, $P < 0.001$]. In experiment 2, replicating the results in experiment 1, participants judged intentional harms (1.0 ± 0.05) to be worse than accidental harms [-0.53 ± 0.11 ; $t(9) = 16.0$, $P < 0.001$]. In experiment 3, participants did not make moral judgments of the scenarios. For analyses of reaction times, see *SI Behavioral Results, Experiment 3*. In experiment 4, when making moral judgments, ASD participants, like NT participants from experiment 1, judged intentional harms (1.0 ± 0.09) more blameworthy than accidental harms [-0.23 ± 0.08 ; $t(10) = 8.0$, $P < 0.0001$], and both intentional and accidental harms were rated worse than neutral acts [-0.77 ± 0.08 ; $t(10) = 11.5$, $P < 0.0001$ and $t(10) = 4.3$, $P = 0.0015$].

Group Comparison. A mixed-effects ANOVA crossing group (NT in experiment 1 and ASD in experiment 4) by condition (accidental, intentional, z-scored ratings) yielded a main effect of condition [$F(1,29) = 446.9$, $P < 0.0001$] and a group by condition interaction [$F(1,29) = 4.7$, $P = 0.03$]. A planned comparison t test (19) revealed that ASD adults assigned more blame for accidental harms than NT adults [$t(29) = 1.9$; $P = 0.03$, one-tailed]. An additional post hoc t test revealed that ASD adults also assigned less blame to intentional actions [$t(29) = 2.1$, $P = 0.04$].

Motion and Artifact Analysis Results. There was no difference in total motion between NT (experiment 1, mean = 0.24 mm/run) and ASD participants [mean = 0.23 mm/run, $t(37) = 0.24$, $P = 0.81$] or in the number of outliers per run [NT: 3.7 ± 0.86 ; ASD: 3.4 ± 1.2 ; $t(37) = 0.2$, $P = 0.8$].

fMRI Results: Functional Localizer. Replicating studies using a similar functional localizer task (2), we localized four theory of mind brain regions showing greater activation for mental state stories (e.g., describing false beliefs) compared with physical state stories (e.g., describing outdated physical representations; $P < 0.001$, $k > 10$) in the majority of individual participants (Table S1). All subsequent analyses are conducted using individually defined regions of interest (ROIs).

fMRI Results: Response Magnitude. Experiments 1 and 4. Averaged over the whole trial, harmful actions elicited a higher response than neutral acts in all four ROIs [RTPJ, left TPJ (LTPJ), PC, dorsal medial prefrontal cortex (DMPFC); all $t > 4.6$, $P < 0.0003$]. In the final 8 s of the trial, after the intention had been revealed, the response in the RTPJ of NT adults was higher for accidental than intentional harms [mean percent signal change from rest, accidental: 0.1 ± 0.04 , intentional: 0.01 ± 0.04 , $t(21) = 3.59$, $P = 0.002$]. LTPJ showed a similar trend [$t(21) = 1.82$, $P = 0.08$]; there was no difference in the other two regions (all $t < 1.5$, all $P > 0.1$). In adults with ASD, there was no difference between accidental and intentional harms in any region (all $t < 1$, $P > 0.3$). In a group (NT, ASD) by condition (accidental, intentional) ANOVA, there was a group by condition interaction in RTPJ [$F(1,36) = 5.37$, $P = 0.03$]. No other effects of group or group by condition interactions were significant (all $F < 2$, all $P > 0.1$; Fig. S1).

fMRI Results: Voxelwise Pattern. Experiment 1. Harm vs. neutral. Multivoxel pattern analyses revealed reliably distinct patterns of neural activity for harmful (intentional and accidental) vs. neutral acts in three of four ROIs: RTPJ, LTPJ, and PC (all $t > 3.2$, $P < 0.002$) and a trend in DMPFC (Fig. 2); the pattern generated by stories in one category (i.e., harmful or neutral) was more correlated with the pattern from other stories in the same category than in the opposite category. All correlations are Fisher Z transformed to allow statistical comparisons with parametric tests.

Experiment 1. Accidental vs. intentional. Only in RTPJ did the pattern of activity distinguish between accidental and intentional harms [within = 1.2 ± 0.12 , across = 1.1 ± 0.12 , $t(21) = 2.2$, $P = 0.02$]. No other regions showed sensitivity to intent (all correlation differences < 0.02 , all $P > 0.3$; Fig. 2).

Experiments 2 and 3. Experiments 2 and 3 replicate experiment 1. In only the RTPJ did MVPA reveal reliably distinct neural patterns for intentional and accidental harms [experiment 2 RTPJ: within = 1.1 ± 0.13 , across = 0.91 ± 0.10 , $t(15) = 2.6$, $P = 0.01$; experiment 3 RTPJ: within = 0.42 ± 0.10 , across = 0.24 ± 0.10 , $t(13) = 2$, $P = 0.034$; all other regions: correlation differences < 0.1 , $P > 0.1$; Fig. 2].

Combining experiments 1–3. Pooling the data across all three experiments increased our power to detect results in neural regions beyond RTPJ. Again, MVPA revealed distinct patterns for accidental and intentional harms in RTPJ [within = $0.96 \pm$

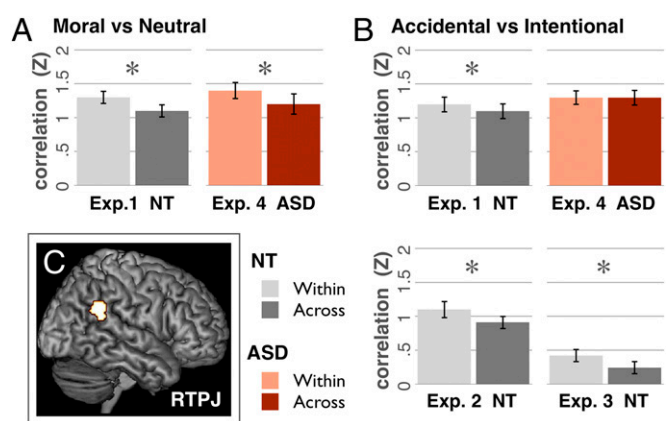
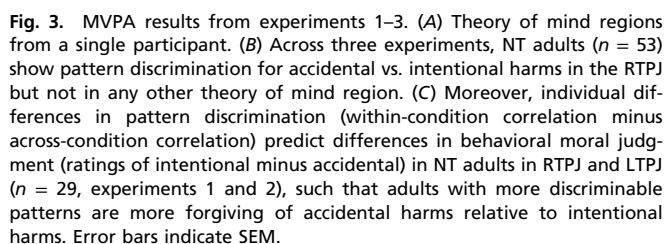


Fig. 2. MVPA results from experiments 1–4. (A) NT ($n = 23$) and ASD adults ($n = 16$) show pattern discrimination in the RTPJ for moral vs. neutral actions, with higher within-condition correlations than across-condition correlations. (B) NT adults show pattern discrimination for accidental vs. intentional harms in RTPJ, a finding replicated across experiments 2 and 3 ($n = 16, 14$), but adults with ASD do not show discrimination of intent. High overall correlation in ASD, combined with matched motion parameters for both groups, suggest that this is not due to noise but rather a stereotyped response across both accidental and intentional harms in adults with ASD. (C) RTPJ in a single participant. Error bars indicate SEM.



Accidental vs. intentional. Accidental and intentional harms did not elicit distinct patterns in any theory of mind ROI in ASD [all $t < 1.1$, $P > 0.16$; Fig. 2; *SI fMRI Results: Voxelwise Pattern, Experiment 4 (ASD)*]. There was a marginal negative correlation

The correlation of pattern discrimination in RTPJ with behavioral responses was significantly larger in NT than in ASD participants (full sample, difference of correlations, $z = 3.0$, $P = 0.003$; matched subsets, $z = 2.6$, $P = 0.009$).

A central aim of this study was to ask whether the difference between accidental and intentional harms could be decoded from the pattern of neural response within theory of mind brain regions. Across three experiments, using different stimuli, paradigms, and participants, we found converging results: in NT adults, stories about intentional vs. accidental harms elicited spatially distinct patterns of response within the RTPJ. Moreover, this neural response mirrored behavioral judgments: individuals who showed more distinct patterns in RTPJ also made a larger distinction between intentional and accidental harms in their moral judgments. Notably, in the fourth experiment, this pattern was absent in adults with ASD: we found no neural difference between accidental and intentional harms in pattern or mean signal, and behavioral judgments showed a marginal negative correlation with neural pattern.

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 story content, voice of the narrative (second or third person), tense (past or present), the severity of harm caused (mild in experiment 1, extreme in experiment 2, unspecified in experiment 3), the order and timing of the story segments, the number of stories per condition, and the participants' explicit task. Perhaps most importantly, the cues to intent were different across experiments. 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/rotten”). Nevertheless, the spatial pattern of response was reliable and distinct for intentional vs. accidental harms specifically in the RTPJ across all three experiments. The converging results across experiments suggests that, rather than being driven by superficial stimulus features or task demands, the distinct neural patterns reflect an underlying distinction in the representation of accidental and intentional harms.

The pattern difference found in the current work suggests that the distinction between intentional and accidental harm is encoded in the neural representation in RTPJ. Interestingly, the pattern in RTPJ did not distinguish between true and false beliefs or between negative and neutral intentions in the context of nonharmful acts (i.e., the difference between neutral acts and failed attempts to harm; *SI Results: Decoding True vs. False Beliefs?*), suggesting that the representations underlying the difference between accidental and intentional harm are relatively specific.

This evidence that one feature of mental states is explicitly represented in the neural pattern opens the door to many future studies. Many other features may also be decodable. For example, patterns of response in theory of mind brain regions may discriminate between attributing beliefs that are justified or unjustified (29), plausible or crazy (30), attributed to friends or enemies (31), constrained or open-ended (32), or first order or higher order (33). Important challenges for future research will include (i) determining the full set of mental state features that can be decoded from the RTPJ response using MVPA and (ii) identifying the features of social stimuli that can be decoded from other regions within and beyond the theory of mind network.

Pattern Discrimination of Intentional vs. Accidental Harm Is Correlated with Moral Judgment. The distinctness of the spatial patterns in the RTPJ was correlated with individuals’ moral judgments. NT adults differed in the amount of blame they assigned to accidental harms: some weighed intent more strongly (i.e., forgive more based on innocent intentions), whereas others weighed outcome more strongly [i.e., blamed more based on bad outcomes (34, 35)]. These differences in moral judgment were predicted by individual differences in neural pattern discriminability in the RTPJ and more weakly in the LTPJ. Although 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?”), the same result emerged in both studies: the individuals who encoded the difference between accidental and intentional harms most strongly in their RTPJ also showed the greatest difference in their moral judgments of these acts. These findings were corroborated by the whole-brain searchlight results: in the participants whose moral judgments were most sensitive to the difference between intentional and accidental harm, only a region within RTPJ discriminated between intentional and accidental harms.

Note that mental states (e.g., beliefs, intentions) represent just one of many inputs to moral judgment. Decisions about moral blame and permissibility depend on many features of the event, including the agent’s beliefs and desires (14), the severity of the harm (36), the agent’s prior record (37), the means of the harm (38, 39), and the external constraints on the agent [e.g., coercion, self-defense, (37, 40), and more (29)]. Thus, activity in the RTPJ reflects the representation of one input to moral judgment, rather than the judgment itself.

In addition to finding distinct neural patterns in RTPJ, we also found a difference in the magnitude of response when participants were reading about intentions and making moral judgments: participants showed a higher level of RTPJ activity for accidental harms relative to intentional harms. The current results converge with, and extend, prior reports that RTPJ is recruited in the face of mitigating mental state information such as innocent intent (22–24). We find the effect in response

magnitude is independent of the difference in neural pattern (*SI fMRI Results: Response Magnitude, Independent Effects of Magnitude and Pattern*), suggesting that these effects are sensitive to different aspects of RTPJ function. A higher magnitude of response to accidental harms suggests that NT adults use mental state reasoning in their moral judgments more when faced with relevant information about the mind of the perpetrator. The stable difference in neural pattern suggests that additionally the distinction between intentional and accidental harm is encoded in the neural representation in RTPJ.

High-Functioning Adults with ASD Show Atypical Neural Activity in RTPJ. In ASD adults, we found distinct patterns for harmful and neutral acts in all theory of mind regions, but no distinction between intentional and accidental harms in the RTPJ or any other region. Similarly, mean signal did not differentiate between accidental and intentional harms in any region. Finally, unlike in NT adults, behavioral responses showed a marginal negative correlation with neural discrimination in the RTPJ.

The results of the current study match the behavioral profile of high-functioning individuals with ASD. Individuals with ASD do not have impairments in moral judgment as a whole: children with ASD make typical distinctions between moral and conventional transgressions (41) and between good and bad actions (42). However, they are delayed in using information about innocent intentions to forgive accidents (18). Furthermore, even very high-functioning adults with ASD who pass traditional tests of understanding (false) beliefs neglect beliefs and intentions in their moral judgments compared with NT adults (19, 20). In the current sample, this effect was observed more strongly in z-scored behavioral data (*SI Behavioral Results: Raw Behavioral Responses, Group Comparison*) (21).

Two prior papers have used MVPA to study neural mechanisms of social cognition (although not specifically theory of mind) in ASD. Gilbert et al. (43) measured the magnitude and pattern of activity in MPFC while participants performed a task that did or did not elicit thinking about another person. The magnitude of response during the two tasks was equivalent in participants with ASD and controls, but participants with ASD showed a less reliable and distinct pattern of activation in the MPFC during the “person” task. Coutanche et al. (44) measured the pattern and magnitude of response in ventral visual areas while viewing faces vs. houses. Again, the magnitude of response in these regions was not different in typical adults and those with ASD, but individuals with ASD showed less discriminable patterns in response to faces, a social category, compared with houses. Disorganization of the pattern of activity for faces vs. houses was correlated with ASD symptom severity.

Broadly consistent with this prior work, the current results suggest that ASD affects the organization (i.e., pattern) of information in theory of mind brain regions. One concern might be that the reduced pattern information is merely a result of more noisy or heterogeneous neural responses in ASD (45). Our results do not favor this interpretation: rather than lower within-condition spatial correlations (i.e., noisier or more idiosyncratic responses), we found numerically higher within-condition correlations in participants with ASD. That is, participants with ASD seemed to show a reliable pattern of response in the RTPJ in response to all harmful acts, regardless of whether the act was intentional or accidental.

We also found a difference between groups in the magnitude of response. NT adults showed a higher response to accidental than intentional harms in the RTPJ, suggesting increased activity in the face of mitigating mental state information. In contrast, ASD adults showed equal activation to both types of stories, suggesting that accidental harms did not elicit more consideration of mental states than intentional harms. Given the independence of the observed differences in magnitude and pattern (*SI fMRI Results: Voxelwise Pattern, Independent Effects of Magnitude and Pattern*), our results provide converging evidence from two different types of analyses of atypical representations of intentional vs. accidental harms in the RTPJ of adults with ASD.

Prior studies investigating the magnitude of response in theory of mind regions of individuals with ASD have found inconsistent results. Some studies suggest that theory of mind regions are hypoactive [i.e., produce a smaller or less selective response (46, 47)], whereas other studies find no difference between ASD and NT individuals (43, 48), and still others find the opposite pattern, hyperactivation, in ASD (49–51). Studies using tasks that elicit spontaneous or implicit social processing may be more likely to find hypoactivation (51–53). Because increases in magnitude may reflect either successful representation of a stimulus, or effortful but ineffectual processing, differences in the magnitude of response between groups can be difficult to interpret.

MVPA may therefore offer a sensitive tool for measuring neural differences in ASD that are related to social impairments. Note, however, that due to the demands of the task and scanning environment, the present ASD participants [as in previous task-oriented neuroimaging studies (54–56)] are very high functioning, which may limit the generalizability of the results to lower-functioning individuals. Nevertheless, the individuals in the current study do experience disproportionate difficulties with social interaction and communication; therefore, the current results provide a window into the neural mechanism underlying these difficulties.

Conclusion

In summary, using MVPA across four experiments, we found that (i) the difference between accidental and intentional harms is linearly decodable from stable and distinct spatial patterns of neural activity in RTPJ; (ii) individual differences in neural discrimination in RTPJ predict individual differences in moral judgment; and (iii) these neural patterns are not detectable in high-functioning adults with ASD. Considerable neuroimaging work on theory of mind suggests that the RTPJ plays some role in thinking about others' thoughts; the current evidence suggests that one aspect of this role is to make explicit, in the population response of its neurons, features of beliefs that are most relevant for inference and decision-making: for example, encoding the intent behind a harmful act.

Methods

Participants. Experiment 1 included 23 right-handed members of the local community (age: 18–50 y, mean = 27 y; seven women). Experiment 2 included 16 right-handed college undergraduate students (age: 18–25 y; eight women). Experiment 3 included 14 right-handed college undergraduate students (age: 18–25 y; eight women). Experiment 4 included 16 individuals diagnosed with ASD (age: 20–46 y, mean = 31 y; two women). Participants in experiment 4 were recruited via advertisements placed with the Asperger's Association of New England. All participants were prescreened using the Autism Quotient questionnaire (AQ) (57). ASD participants (mean = 32.6) scored significantly higher on the AQ than the NT participants from experiment 1 [mean = 17.3; $t(25.5) = 6.4$, $P < 0.0001$].

ASD participants underwent both the ADOS (58, 59) and an impression by a clinician trained in both ADOS administration and diagnosis of ASD. All ASD participants received a diagnosis of ASD based on their social ADOS score (criterion ≥ 4 ; mean = 6.4), communication ADOS score (criterion ≥ 2 ; mean = 3.1), total ADOS score (criterion ≥ 7 ; mean = 9.5), and on a clinical impression based on the diagnostic criteria of the DSM-IV (60). The NT (experiment 1) and ASD (experiment 4) groups did not differ in age [NT mean = 27 y, ASD mean = 31 y, $t(35.26) = 1.32, P = 0.2$] or IQ [NT mean = 121; ASD = 120; $t(28.3) = 0.26, P = 0.8$]. Additionally, all analyses were run with one-to-one matched subsets of participants (both $n = 15$; *SI Results: Pairwise Matched Subsets of ASD and NT*).

All participants were native English speakers, had normal or corrected-to-normal vision, gave written informed consent in accordance with the requirements of Institutional Review Board at Massachusetts Institute of Technology (MIT), and received payment. Data from experiments 2 and 3 have previously been published in papers analyzing the magnitude but not the pattern of response in each region (35, 61).

fMRI Protocol and Task. *Experiments 1 and 4.* Participants were scanned while reading 60 stories told in second person (Fig. 1; Fig. S3): 12 described harm caused intentionally (e.g., knowingly kicking someone in the face), 12 described harm caused accidentally (e.g., kicking without seeing the person), 12 described neutral actions (e.g., eating lunch), and 24 stories describing disgusting but not harmful actions (e.g., smearing feces on one's own face;

not analyzed here). Stories were presented in four cumulative segments, describing the background (6 s), action (+4 s), outcome (+4 s), and intention (+4 s). After each story, participants made a moral judgment of the action ("How much blame should you get?") from "none at all" (1) to "very much" (4), using a button press. Behavioral data from five ASD and three NT participants were lost due to error ($n = 4$) or to theft of experimental equipment ($n = 4$). Ten stories were presented in each 5.5-min run; the total experiment, six runs, lasted 33.2 min. For more details, and more sample stimuli, see [SI Methods, Experiments 1 and 4](#), and [Fig. S3](#).

Experiments 2 and 3. Participants were scanned while reading 48 stories, both in 3rd person. Experiment 2 included 12 intentional harms, 12 accidental harms, and 24 actions that did not cause harm (Fig. S4). Stories were presented in four cumulative segments, describing the background (6 s), fore-shadow (+6 s), intent (+6 s), and outcome (+6 s). Participants made moral judgments of the action on a three-point scale from “forbidden” (1) to “permissible” (3) (61). Experiment 3 included 8 intentional harms, 8 accidental harms, 16 actions that did not cause harm, and 16 actions with no specified outcome (Fig. S5). Stories were presented in three cumulative segments, describing the background (6 s), intent (+6 s), and outcome (+6 s). Participants answered a true/false question about the content of the final sentence (35). For more details and sample stimuli, see *SI Methods, Experiment 2 and Experiment 3*, and Fig. S4 and S5.

Theory of Mind Localizer Task. Participants read verbal narratives about thoughts (belief) vs. about physical representations like photographs and maps (photo; *SI Methods, Theory of Mind Localizer Task*) (2).

Acquisition and Preprocessing. In all four experiments, fMRI data were collected in a 3-T 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, blood oxygen level dependent (BOLD) signal was acquired in 26 near axial slices using $3 \times 3 \times 4$ -mm voxels (repetition time (TR) = 2 s, echo time (TE) = 40 ms, flip angle = 90°). To allow for steady-state magnetization, the first 4 s of each run were excluded. Data processing and analysis were performed using Statistical Parametric Mapping 8 (experiments 1 and 4; <http://www.fil.ion.ucl.ac.uk/spm>) and 2 (experiments 2 and 3), and custom software. The data were motion corrected, realigned, normalized onto a common brain space (MNI template), spatially smoothed using a Gaussian filter (full-width half-maximum 5-mm kernel), and high-pass filtered (128 Hz).

Behavioral Analysis. For more information on behavioral analysis, see [SI Behavioral Results: Raw Behavioral Responses](#).

Motion and Artifact Analysis. For the NT (experiment 1) and ASD (experiment 4) participants, we calculated the total motion per run (sum of translation in three dimensions), and the number of time points that either (i) deviated from the global mean signal by more than 3 SD or (ii) included TR-to-TR motion of more than 2 mm.

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: beginning with the onset of text on screen and ending after the response prompt was removed.

Theory of mind localizer: Individual ROIs. Functional ROIs were defined in the RTPJ, LTJP, PC, and DMPCF. For each participant, we found the peak voxel in the contrast image for each region. ROIs were then defined as all voxels within a 9-mm radius of the peak voxel that passed threshold in the contrast image (belief > photo, $P < 0.001$, uncorrected, $k > 10$; [SI Results: Theory of Mind Localizer ROIs](#) and [Table S1](#)).

ROI pattern analysis. In all four experiments, we conducted within-ROI pattern analyses. Following Haxby et al. (62), each participant's data were divided into even and odd runs (partitions), and the mean response (β value) of every voxel in the ROI was calculated for each condition. The pattern of activity was defined as the vector of β values across voxels within the ROI. To calculate the within-condition correlation, the pattern in one (e.g., even) partition was compared with the pattern for the same condition in the opposite (e.g., odd) partition; to calculate the across-condition correlation, the pattern was compared with the opposite condition, across partitions. For each individual, an index of classification was calculated as the z-scored within-condition correlation minus the z-scored across-condition correlation. A region successfully classified a difference in conditions if, across individuals, the within-condition correlation was higher than the across-condition correlation.

relation, using a Student's *t* complementary cumulative distribution function. Note that in this procedure (unlike other machine learning approaches; e.g., support vector machines), differences in the spatial patterns across conditions are independent of differences in the average magnitude of response. This procedure implements a simple linear decoder, which is, although in principle less flexible and less powerful than nonlinear decoding, preferable both theoretically and empirically (3, 4, 9).

Whole Brain Pattern Analysis. For more information on whole brain pattern analysis, see *SI Methods, Whole Brain Pattern Analysis: Searchlight*.

- Aichhorn MM, et al. (2009) Temporo-parietal junction activity in theory-of-mind tasks: Falseness, beliefs, or attention. *J Cogn Neurosci* 21(6):1179–1192.
- 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.
- DiCarlo JJ, Zoccolan D, Rust NC (2012) How does the brain solve visual object recognition? *Neuron* 73(3):415–434.
- Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. *Nat Neurosci* 8(5):679–685.
- Dehaene S, Cohen L (2007) Cultural recycling of cortical maps. *Neuron* 56(2):384–398.
- Formisano E, et al. (2003) Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron* 40(4):859–869.
- Haynes J-D, Rees G (2006) Decoding mental states from brain activity in humans. *Nat Rev Neurosci* 7(7):523–534.
- Kriegeskorte N, Bandettini P (2007) Analyzing for information, not activation, to exploit high-resolution fMRI. *Neuroimage* 38(4):649–662.
- Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* 10(9):424–430.
- Kamitani Y, Tong F (2006) Decoding seen and attended motion directions from activity in the human visual cortex. *Curr Biol* 16(11):1096–1102.
- Mahon BZ, Caramazza A (2010) Judging semantic similarity: An event-related fMRI study with auditory word stimuli. *Neuroscience* 169(1):279–286.
- Peelen MV, Wiggett AJ, Downing PE (2006) Patterns of fMRI activity dissociate overlapping functional brain areas that respond to biological motion. *Neuron* 49(6):815–822.
- Raizada RDS, et al. (2010) Linking brain-wide multivoxel activation patterns to behaviour: Examples from language and math. *Neuroimage* 51(1):462–471.
- Cushman F (2008) Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108(2):353–380.
- Baird JA, Astington JW (2004) The role of mental state understanding in the development of moral cognition and moral action. *New Dir Child Adolesc Dev* 2004(103):37–49.
- Baron-Cohen S (1997) *Mindblindness: An essay on autism and theory of mind* (MIT Press, Cambridge, MA).
- Peterson CC, Wellman HM, Liu D (2005) Steps in theory-of-mind development for children with deafness or autism. *Child Dev* 76(2):502–517.
- Grant CM, Boucher J, Riggs KJ, Grayson A (2005) Moral understanding in children with autism. *Autism* 9(3):317–331.
- Moran JM, et al. (2011) Impaired theory of mind for moral judgment in high-functioning autism. *Proc Natl Acad Sci USA* 108(7):2688–2692.
- Yang D, Baillargeon R (2013) Difficulty in understanding social acting (but not false beliefs) mediates the link between autistic traits and ingroup relationships. *J Autism Dev Disord* 10.1007/s10803-013-1757-3.
- Baez S, et al. (2012) Integrating intention and context: Assessing social cognition in adults with Asperger syndrome. *Front Hum Neurosci* 6:302.
- Yamada M, et al. (2012) Neural circuits in the brain that are activated when mitigating criminal sentences. *Nat Commun* 3:759.
- Buckholz JW, et al. (2008) The neural correlates of third-party punishment. *Neuron* 60(5):930–940.
- Young LL, Saxe R (2009) Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia* 47(10):2065–2072.
- Young LL, Camprodon JA, 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. *Proc Natl Acad Sci USA* 107(15):6753–6758.
- Fedorenko E, Nieto-Castañón A, Kanwisher N (2012) Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia* 50(4):499–513.
- Hampton AN, O’Doherty JP (2007) Decoding the neural substrates of reward-related decision making with functional MRI. *Proc Natl Acad Sci USA* 104(4):1377–1382.
- Tusche A, Bode S, Haynes JD (2010) Neural responses to unattended products predict later consumer choices. *J Neurosci* 30(23):8024–8031.
- Young LL, Nichols S, Saxe R (2010) Investigating the neural and cognitive basis of moral luck: It’s not what you do but what you know. *Rev Philos Psychol* 1(3):333–349.
- Young LL, Dodell-Feder D, Saxe R (2010) What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia* 48(9):2658–2664.
- Bruneau EG, Pluta A, Saxe R (2012) Distinct roles of the ‘shared pain’ and ‘theory of mind’ networks in processing others’ emotional suffering. *Neuropsychologia* 50(2):219–231.
- Jenkins AC, Mitchell JP (2010) Mentalizing under uncertainty: Dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cereb Cortex* 20(2):404–410.
- Koster-Hale J, Saxe RR (2011) Theory of mind brain regions are sensitive to the content, not the structural complexity, of belief attributions. *Proceedings of the 33rd Annual Cognitive Science Society Conference*, eds Carlson L, Hoelscher C, Shipley TF (Cognitive Science Society, Austin, TX), pp 3356–3361.
- Nichols S, Ulatowski J (2007) Intuitions and individual differences: The Knobe effect revisited. *Mind Lang* 22(4):346–365.
- Young LL, Saxe R (2009) An fMRI investigation of spontaneous mental state inference for moral judgment. *J Cogn Neurosci* 21(7):1396–1405.
- Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD (2001) An fMRI investigation of emotional engagement in moral judgment. *Science* 293(5537):2105–2108.
- Woolfolk RL, Doris JM, Darley JM (2006) Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition* 100(2):283–301.
- Cushman F, Young LL, Hauser M (2006) The role of conscious reasoning and intuition in moral judgment: testing three principles of harm. *Psychol Sci* 17(12):1082–1089.
- Greene JD, Nystrom LE, Engell AD, Darley JM, Cohen JD (2004) The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44(2):389–400.
- Young LL, Phillips J (2011) The paradox of moral focus. *Cognition* 119(2):166–178.
- Blair RJ (1996) Brief report: Morality in the autistic child. *J Autism Dev Disord* 26(5):571–579.
- Leslie AM, Mallon R, DiCorcia JA (2006) Transgressors, victims, and cry babies: Is basic moral judgment spared in autism? *Soc Neurosci* 1(3-4):270–283.
- Gilbert SJ, Meuwese JDI, Towgood KJ, Frith CD, Burgess PW (2009) Abnormal functional specialization within medial prefrontal cortex in high-functioning autism: A multi-voxel similarity analysis. *Brain* 132(Pt 4):869–878.
- Coutanche MN, Thompson-Schill SL, Schultz RT (2011) Multi-voxel pattern analysis of fMRI data predicts clinical symptom severity. *Neuroimage* 57(1):113–123.
- Dinstein I, et al. (2010) Normal movement selectivity in autism. *Neuron* 66(3):461–469.
- Kennedy DP, Courchesne E (2008) Functional abnormalities of the default network during self- and other-reflection in autism. *Soc Cogn Affect Neurosci* 3(2):177–190.
- Lombardo MV, Chakrabarti B, Bullmore ET, Baron-Cohen S, MRC AIMS Consortium (2011) Specialization of right temporo-parietal junction for mentalizing and its relation to social impairments in autism. *Neuroimage* 56(3):1832–1838.
- Nieminen-von Wendt T, et al. (2003) Changes in cerebral blood flow in Asperger syndrome during theory of mind tasks presented by the auditory route. *Eur Child Adolesc Psychiatry* 12(4):178–189.
- Mason RA, Williams DL, Kana RK, Minshew N, Just MA (2008) Theory of Mind disruption and recruitment of the right hemisphere during narrative comprehension in autism. *Neuropsychologia* 46(1):269–280.
- Tesink CMJY, et al. (2009) Neural correlates of pragmatic language comprehension in autism spectrum disorders. *Brain* 132(Pt 7):1941–1952.
- Wang AT, Lee SS, Sigman M, Dapretto M (2006) Neural basis of irony comprehension in children with autism: The role of prosody and context. *Brain* 129(Pt 4):932–943.
- Adolphs R (2001) The neurobiology of social cognition. *Curr Opin Neurobiol* 11(2):231–239.
- Groen WB, et al. (2010) Semantic, factual, and social language comprehension in adolescents with autism: An fMRI study. *Cereb Cortex* 20(8):1937–1945.
- Brieber S, et al. (2010) Coherent motion processing in autism spectrum disorder (ASD): an fMRI study. *Neuropsychologia* 48(6):1644–1651.
- Koshino H, et al. (2008) fMRI investigation of working memory for faces in autism: Visual coding and underconnectivity with frontal areas. *Cereb Cortex* 18(2):289–300.
- Philip RCM, et al. (2010) Deficits in facial, body movement and vocal emotional processing in autism spectrum disorders. *Psychol Med* 40(11):1919–1929.
- Baron-Cohen S (2001) Theory of mind in normal development and autism. *Prisme* 34:174–183.
- Lord C, et al. (2000) The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism. *J Autism Dev Disord* 30(3):205–223.
- Joseph RM, Tager-Flusberg H, Lord C (2002) Cognitive profiles and social-communicative functioning in children with autism spectrum disorder. *J Child Psychol Psychiatry* 43(6):807–821.
- American Psychiatric Association, American Psychiatric Association Task Force on DSM-IV (2000) *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR* (American Psychiatric Association, Washington, DC).
- Young LL, Saxe R (2008) The neural basis of belief encoding and integration in moral judgment. *Neuroimage* 40(4):1912–1920.
- Haxby JV, et al. (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293(5539):2425–2430.