Decoding moral judgments from neural representations of intentions

Jorie Koster-Haleab, Rebecca Saxeac, James Dunganb, and Liane L. Youngb

abMcGovern 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 temporoparietal 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 temporoparietal 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...
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 ± 0.3) were rated as more blameworthy than accidental harms [z = -25.0, P < 0.001], and both were rated more blameworthy than neutral acts [z = 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 ± 0.5) to be worse than accidental harms [z = 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 [z = -25.0, P < 0.001], and both intentional and accidental harms were rated worse than neutral acts [z = 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 (DMPC); 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 DMPC (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(15) = 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 ±
ROI was found in only 7 of 16 participants. Across neutral acts in ASD. Using individual ROIs, we found significant patterns of activity for harmful vs. intentional harms. Error bars indicate SEM.

Behavioral and neural correlation. In experiment 1 and 2, NT participants provided moral judgments of each scenario in the scanner, allowing us to determine whether behavioral responses were related to neural pattern. In both experiments, we found that in only the RTPJ, the difference between intentional and accidental harms in individuals’ moral judgments was correlated with the neural classification index [within-condition correlation minus across-condition correlation] predict differences in behavioral moral judgment (ratings of intentional versus accidental) in NT adults in RTPJ and LTPJ (n = 29, experiments 1 and 2), such that adults with more discriminable patterns are more forgiving of accidental harms relative to intentional harms. Error bars indicate SEM.

Whole brain searchlight analysis. Combining across all NT participants (n = 53), we found only one small region that discriminated (P < 0.001, voxelwise, uncorrected) between accidental and intentional harms in the fusiform gyrus [peak voxel Montreal Neurological Institute (MNI) coordinates: (30, −54, −14); Fig. S2]. We then median-split participants based on the difference in their moral judgments between intentional and accidental harms (experiments 1 and 2). In participants showing a larger behavioral effect (n = 15), the only region that discriminated between intentional and accidental harms was RTPJ [peak voxel MNI coordinates: (52, −60, 28); Fig. S2]. In the remaining participants (n = 15), no significant voxels were found.

Experiment 4 (ASD). Harm vs. neutral. As in NT controls, pattern analyses revealed distinct patterns of activity for harmful vs. neutral acts in ASD. Using individual ROIs, we found significant discrimination in RTPJ, LTPJ, and PC (all t > 1.2, P < 0.03; Fig. 2). The effect was nonsignificant in DMPFC [within = 1.2 ± 0.11, across = 1.2 ± 0.12, r(6) = 0.21, P = 0.42], possibly because an ROI was found in only 7 of 16 participants.

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 between participants’ moral judgments of accidental vs. intentional harms and pattern discrimination in RTPJ [r(9) = −0.54, P = 0.09]. Note that this effect is in the reverse direction of the correlation observed in NT adults. There was no correlation with behavior in any other region (all r < 0.2, P > 0.5). There was also no correlation with symptom severity [Autism Diagnostic Observation Schedule (ADOS) score] in any region.

Whole brain searchlight (ASD). No regions discriminated between accidental and intentional harms (P < 0.001, voxelwise, uncorrected).

Group comparison. Harm vs. neutral. A group (ASD, NT) by pattern (within, across) ANOVA revealed that NT and ASD participants show strong and equally robust neural discrimination in response to moral violations vs. neutral actions in their RTPJ, LTPJ, and PC, with a main effect of pattern (all F > 13, P < 0.0008) and no effect of group and no interaction (all F < 1.7, P > 0.2) in all three ROIs (SI fMRI Results: Voxelwise Pattern, Group Comparison: Harm vs Neutral). There were no significant effects in DMPFC [pattern: F(1,23) = 2.7, P = 0.1; group: F(1,23) = 0.14, P = 0.7; interaction: F(1,23) = 0.7, P = 0.4].

Accidental vs. intentional. In the RTPJ, accidental and intentional harms were more discriminable in NT than ASD participants, reflected in a significant group by pattern interaction [F(2,36) = 4.9, P = 0.03], with no main effect of pattern [F(1,36) = 1.7, P = 0.2] or group [F(1,36) = 0.71, P = 0.4]. The same interaction was observed in one-to-one matched subsets of participants [F(1,28) = 5.9, P = 0.02; SI Results: Pairwise Matched Subsets of ASD and NT]. There was no group by pattern interaction in any other region (all F < 0.9, all P > 0.3).

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).

Discussion.
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.

MVP A Discriminates Accidental and Intentional Harms in RTPJ of NT Adults. The current results suggest that the RTPJ contains a specific representation that distinguishes intentional from accidental harms. This representation was apparent in reliable but distinct spatial patterns of activity. These results extend this method to high-level cognition and abstract stimulus features (12,26–28). In particular, we provide evidence of a feature of mental state reasoning explicitly represented in a theory of mind brain region.

The convergence across experiments provides strong evidence that intentional and accidental harms can be discriminated, using MVP A, 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 non-harmful 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 distinguishing between true and false beliefs, or between negative and neutral intentions in the context of non-harmful 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.

Pattern Discrimination of Intentional vs. Accidental Harm Is Correlated with Moral Judgment. The distinctness of the spatial patterns in the RTPJ was consistent with the idea that 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 permisibility scale (“How necessary 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 inten- tional 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 permisibility 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 made judgments about 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 very high-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 infer harm (43). 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 not require the time 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 hyporesponsive [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 hyporesponsivity (51–53). Because increases in magnitude may reflect either successful representation of a stimulus, or effortful but ineffective processing, differences in the magnitude of response between groups can be difficult to interpret.

MVPa may therefore offer a sensitive tool for measuring neural activity in RTPJ in ASD that is 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 predictable 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 the representation in ASD of the population route of if-then scenarios, a feature 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 Association of New England. All participants were prescreened using the Autism Quotient questionnaire (AQ) (57). ASD participants (mean AQ = 32.6) scored significantly higher on the AQ than the NT participants from experiments 1 and 4 (mean AQ = 6.4, P < 0.0001; uncorrected). 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(52) = 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. 53). 12 described harm caused accidentally (e.g., kicking someone who is sleeping), 12 described harm caused accidentally (e.g., kicking someone who is sleeping), 12 described neutral actions (e.g., eating lunch), and 24 stories describing disgusting but not harmful actions (e.g., smelling 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. 53.

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), foreground (+4 s), intent (+4 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, Experiments 2 and 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. Data were preprocessed using the customized and extensively validated program 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, and ROI analyses 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).

Supporting Information

Koster-Hale et al. 10.1073/pnas.1207992110

SI Methods

Theory of Mind Localizer Task. All subjects participated in a theory of mind localizer task, contrasting stories requiring inferences about mental state representations (e.g., thoughts, beliefs) vs. physical representations (e.g., maps, signs, photographs). Mental state stimuli vs. physical representation stimuli were similar in their metarepresentational and logical complexity; the key difference was that for the mental state stories, the reader had to build a representation of someone else’s mental state (1, 2). Stimuli and experimental design are available at http://saxelab.mit.edu/supero6.php. This experiment was modeled treating each trial as a block, with a boxcar lasting 14 s, which was the whole period from the initial presentation of the story to the end of the question presentation. β values were estimated, in each voxel, for stories describing mental states (belief) or physical representations (photo). Then a simple contrast map was produced in each subject, identifying voxels responding more to belief than photo stories at P < 0.001 uncorrected for multiple comparisons.

Experiments 1 and 4. Participants were scanned while reading 60 stories (for samples, see Fig. S3). Stories were presented in the second person, using present tense. Each participant read 12 stories describing someone harming another individual caused accidentally, 12 describing someone harming another individual intentionally, 12 neutral actions (described below), and 24 stories describing disgusting but not harmful actions (e.g., consensual incest, drinking blood, smearing feces on one’s own face; not analyzed here).

Each story was displayed in four cumulative segments (i.e., earlier segments remained on the screen as later segments were added). Cumulative segments partially reduce the variability in reading time across participants and items by slowing down faster participants and shorter items, without cutting off longer segments or slower readers.

For accidental and intentional harms, the three segments of the stories were identical: information about the background (6 s), action (+4 s), and harmful outcome (+4 s). Harmful outcomes included both physical harms (kicking someone in the face, feeding them an allergen) and psychological harms (insulting or humiliating someone). The only information that distinguished between these two conditions was the final sentence, describing the extent (+4 s), which was innocent in the accidental harm condition (e.g., “you did not see,” “you did not know,” that the act would cause harm) and negative in the intentional harm condition (e.g., “you saw,” “you realized,” that the act would cause harm). The accidental and intentional sentences were matched pairwise by changing the fewest possible words in the intent segment, thus preserving length (see Fig. S3 for examples; mean = 14 words, SD = 1.9, range = 9–18). Across subjects, each scenario (i.e., background + action + outcome) was equally likely to occur in the intentional or accidental condition. Each participant saw either the intentional or the accidental version of each scenario but not both.

For the neutral actions, a separate set of scenarios described a neutral background and action (e.g., giving a class presentation, eating lunch). The outcome was mildly positive (the teacher liked the graph) or negative (a little spilled ketchup), and the intent described either knowledge or no knowledge of the outcome (e.g., “you did/did not realize” that you spilled ketchup). Because there were only 12 neutral stories in total, we used the averaged response for all neutral stories in the analyses.

In the scanner, 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. Participants were given 4 s to respond before the screen was blanked. Thus, the whole trial lasted 22 s. A 10-s rest block (blank screen) occurred after each trial.

Stories were presented in a pseudorandom order, such that no condition was immediately repeated, each condition was equally likely to occur in each position in the run, and each condition was equally likely to follow each other condition. To avoid being presented in each 5.5-min run; the total experiment, six runs, lasted 33.2 min. Stories were projected onto a screen via Matlab 5.0 running on an Apple MacBook Pro in 40-point white font.

Experiment 2. Participants were scanned while reading 48 stories. Stories were presented in the third person, using past tense. Each participant read stories describing 12 intentional harms, 12 accidental harms, 12 failed attempts to harm, and 12 neutral actions (for samples, see Fig. S4; for the complete list, see ref. 3). The focus of this article is the accidental and intentional harm stories. All harms were physical harms, resulting in someone’s death or serious injury.

Stories were presented in four cumulative segments: background (6 s); foreshadow (+6 s); belief (+6 s); and outcome (+6 s). Half of the stories in each run were presented with foreshadow before intent; the other half were presented with intent before foreshadow. The background, foreshadow, and outcome segments of intentional and accidental harms were identical. The belief segments described the content of the character’s belief (e.g., “Leah believes that the child is about to dive into shallow water and break his neck”) and the reason for that belief (e.g., “Because of a warning sign at the side of the pool”). The difference between conditions was created by manipulating the content of the belief. In the context of the foreshadow (“The child is about to dive into the shallow end and smash his head very hard”), the action (“Leah walks by, without saying anything to the child”), and the negative outcome, these beliefs determined whether the harm (“The child dives in and breaks his neck”) was caused intentionally or accidentally.

After each story, participants made moral judgments (“How morally permissible was X’s action?”) of the action on a three-point scale, from “forbidden” (1) to “permissible” (3), using a button press. Participants were given 4 s to respond before the screen was blanked. All trials were followed by a 14-s rest block. Stories were projected onto a screen via Matlab 5.0 running on an Apple G4 laptop in 24-point white font. Behavioral data were available for 10 participants. The methods and materials from experiment 2 have been published in ref. 3. Young and Saxe (3) reported results from 17 participants; data from 1 participant were lost, leaving 16 participants.

Experiment 3. Participants were scanned while reading 48 stories. Stories were presented in the third person, using present tense. Each participant read stories describing 8 intentional harms, 8 accidental harms, 8 failed attempts to harm, 8 neutral actions, and 16 actions for which the outcome was unknown (a morally irrelevant fact was presented in place of the outcome; for samples, see Fig. S5 and ref. 4, experiment 2). Stories were presented in cumulative segments: background (6 s), belief (+6 s), and fact (+6 s). The focus of this study is the accidental and intentional harm stories; the background and fact segments were identical for these conditions. Belief segments described the content of
the character’s belief (e.g., “Steve believes the ground beef is safe to eat”) and the reason for that belief (e.g., “because of the expiration date.”). The difference between conditions was created by manipulating the content of the belief. In the context of the fact (e.g., “The meat has some invisible but deadly bacteria”), these beliefs led to either accidental or intentional harm, although the harmful outcome itself (i.e., the children becoming ill) was only implied and not explicitly stated. After each story, participants answered a true/false question about the fact segment (e.g., “The meat is unsafe for consumption: True/False”).

Participants were given 4 s to respond before the screen was blanked. All stories were followed by a 14-s rest block. Stories were projected onto a screen via Matlab 5.0 running on an Apple G4 laptop in 24-point white font. The methods and materials from experiment 3 have been published in ref. 4.

Whole Brain Pattern Analysis: Searchlight. In this analysis, rather than using a predefined region of interest (ROI), a Gaussian kernel (14-mm full-width half-maximum, corresponding approximately to the size of the functional ROIs) is moved iteratively across the brain. Using the same logic as the ROI-based multivoxel pattern analysis (MVPA), we compute the spatial correlation, in each kernel, of the neural response (i.e., β) within condition and across conditions, to transform the correlations using Fisher’s Z and subtract the across-condition from the within-condition correlation to create an index of classification. Thus, for each voxel, we obtain an index of how well the spatial pattern of response in the local region (i.e., the area centered on that voxel) can distinguish between the two conditions of interest. The use of a Gaussian kernel smoothly deemphasizes the influence of voxels at increasing distances from the reference voxel (5). We created whole brain maps of the index of classification for each subject. These individual correlation maps were subjected to a second-level analysis using a one-sample t test (thresholded at P < 0.001, voxelwise, uncorrected). Note that because each voxel’s discrimination index reflects the spatial pattern in the surrounding region, the map of spatial discrimination is highly smooth, reducing the effective number of comparisons.

This classification procedure implements a simple linear decoder. Linear decoding, while in principle less flexible and less powerful than nonlinear decoding is preferable both theoretically and empirically. A nonlinear classifier can decode nearly any arbitrary feature contained implicitly within an ROI, reflecting properties of the pattern analysis algorithm rather than the brain, which makes successful classification largely uninformative (6–10). Moreover, linear codes have been argued to be more neurally plausible, reflecting the information made available to the next layer of neurons (7, 11–14).

Within-ROI Magnitude Analysis. We measured the response to each condition in each ROI. Baseline response in each region was the average blood oxygen level dependent (BOLD) response, in that region, at all time points when there was no stimulus on the screen, excluding the first 4 s after the offset of each stimulus (to allow the hemodynamic response to decay). The percent signal change (PSC) at time t = 100 t [average BOLD magnitude for condition at time t − average BOLD magnitude for fixation]/average BOLD magnitude for fixation. We averaged the PSC across the entire presentation (offset 6 s from presentation time to account for hemodynamic lag) to estimate a single PSC for each condition in each ROI in each participant (15). Using custom software to visualize the percent signal change in a region, we extract the full time course in the region (i.e., set of voxels), and then create an event-related average, by aligning and averaging the raw BOLD response at each time point after the trial onset per condition (rather than extracting the β value for the whole block). Other software packages (e.g., SPM) calculate the baseline response as the mean over the run, which is undesirable because it will overestimate the response at baseline in a region with a strong positive response to most blocks and will underestimate the response at baseline in a region that deactivates during most blocks.

In addition, because the stimuli in the current experiment were presented cumulatively, for experiments 1 and 4, we separately analyzed the PSC for two phases of the trial: (i) the initial phase (background + action + outcome, first 14 s), in which intentional and accidental harms could not be distinguished, and (ii) the final phase (intent + decision, final 8 s), after intentional and accidental harms were distinguished.

SI Behavioral Results: Results from Raw Behavioral Responses. In the main text, we report behavioral data and statistics that have been z-scored to account for variability in the participants’ use of the rating scales. Here we provide the same analyses, using the untransformed button press responses.

Experiment 1. From no blame at all (1) to very much blame (4), participants judged intentional harms (3.4 ± 0.09) to be more blameworthy than accidental harms [1.5 ± 0.08; t(19) = 19; P < 0.0001], both of which were judged to be worse than neutral stories [1.1 ± 0.03; t(19) = 30; P < 0.0001 and t(19) = 6.9; P < 0.0001].

Experiment 2. Replicating the results in experiment 1, on a scale of forbidden (1) to permissible (3), participants judged intentional harms (1.1 ± 0.04) as less permissible than accidental harms [2.2 ± 0.11; t(9) = 11; P < 0.001].

Experiment 3. Participants in experiment 3 did not make moral judgments of the scenarios, but instead answered true/false questions about the content of the fact. Reaction time (RT) was analyzed using a one-way repeated-measures ANOVA, revealing a main effect of condition [F(1,13) = 4.56; P < 0.02]. Post hoc t tests revealed that RTs for answering factual questions following nonharm stories (mean RT = 2.4 s) were not different from RTs following accidental harms [mean 2.3 s; t(13) = 1.65; P > 0.12] or intentional harms [mean 2.5 s; t(13) = 1.49; P > 0.16], but that responses following intentional harms were slower than responses following accidental harms [t(13) = 2.87; P < 0.01).

Experiment 4. When making moral judgments, autism spectrum disorder (ASD) participants, like neurotypical (NT) participants from experiment 1, judged intentional harms (3.4 ± 0.18) to be worse than accidental harms [1.9 ± 0.17; t(10) = 7.5; P < 0.0001], both of which were judged to be worse than neutral stories [1.2 ± 0.06; t(10) = 10.7; P = 8.3e−07 and t(10) = 4.3; P = 0.0016].

Group Comparison. A mixed-effects ANOVA crossing group (NT in experiment 1 and ASD in experiment 4) by condition (accidental and intentional raw behavioral ratings) yielded a main effect of condition [F(1,29) = 381.1; P < 0.0001], with no effect of group [F(1,29) = 2.7; P = 0.1] and no interaction [F(1,29) = 1.1; P = 0.3]. The planned comparison t test revealed that ASD adults assigned more blame for accidental harms than NT adults [t(14.2) = 1.7; P = 0.05; one-tailed] (16).

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 right temporoparietal junction (RTPJ) or any other region. For each participant, we calculated the difference in raw moral judgments
for intentional vs. 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 [experiment 1: r(17) = 0.55; P = 0.016; experiment 2: r(8) = 0.71; P = 0.02]. The correlation between neural pattern and behavior remained significant after combining the data from both experiments, using the moral judgments converted to the same scale [r(27) = 0.56; P = 0.002]. No other regions showed a significant correlation with behavioral judgments, although there was a strong trend in the left temporo-parietal junction [LTPJ; r(27) = 0.35; P = 0.06].

**Results: Theory of Mind Localizer ROIs**

Four ROIs were identified for each subject, based on the contrast “belief > photo” thresholded at P < 0.001. For experiments 1–3 (NT), these were as follows: RTPJ (52/53), LTPJ (51/53), precuneus (PC; 51/53), and dorsal medial prefrontal cortex (DMPFC; 43/53). For experiment 4 (ASD), these were as follows: RTPJ (16/16), LTPJ (15/16), PC (16/16), and DMPFC (7/16) (Table S1). Table S1 reports the average position of the peak and the average SD of the ROI size for each ROI for each experiment.

**SI fMRI Results: Voxelwise Pattern**

**Experiment 1: Harm vs. Neutral.** Multivoxel pattern analyses revealed reliably distinct patterns of neural activity for harmful (intentional, accidental) vs. neutral acts in three of four ROIs: RTPJ, LTPJ, and PC [RTPJ: within = 1.3 ± 0.10, across = 1.1 ± 0.10, t(21) = 3.2, P = 0.002; LTPJ: within = 1.6 ± 0.05, across = 1.4 ± 0.06, t(21) = 3.3, P = 0.002; PC: within = 1.1 ± 0.09, across = 0.86 ± 0.1, t(21) = 3.4, P = 0.001]. The DMPFC showed the same trend [DMPFC: within = 1.2 ± 0.09, across = 1.0 ± 0.09, t(17) = 1.7, P = 0.056].

**Experiment 4 (ASD). Harm vs. neutral.** As in NT controls, pattern analyses revealed a separation in the pattern of response for harmful vs. neutral acts in ASD. Using individual ROIs, significant discrimination was found in RTPJ, LTPJ, and PC [RTPJ: within = 1.4 ± 0.13, across = 1.2 ± 0.16, t(15) = 4.1, P = 0.0005; LTPJ: within = 1.5 ± 0.10, across = 1.3 ± 0.16, t(14) = 2.1, P = 0.027; PC: within = 1.2 ± 0.14, across = 1.1 ± 0.14, t(14) = 1.9, P = 0.04; Fig. 2]. The effect was nonsignificant in DMPFC [within = 1.2 ± 0.11, across = 1.2 ± 0.12, t(6) = 0.21, P = 0.42], possibly because an ROI for DMPFC was defined in only 7 of 16 participants.

**Accidental vs. Intentional.** Accidental and intentional harms did not elicit distinct patterns in any theory of mind ROI in ASD [RTPJ: within = 1.3 ± 0.11, across = 1.3 ± 0.12, t(15) = 1.1, P = 0.86; LTPJ: within = 1.4 ± 0.12, across = 1.4 ± 0.14, t(14) = 1.1, P = 0.86; PC: within = 1.2 ± 0.14, across = 1.2 ± 0.15, t(14) = 1, P = 0.16; DMPFC: within = 1 ± 0.15, across = 1.1 ± 0.08, t(6) = 0.59, P = 0.71].

**Group Comparison: Harm vs. Neutral.** A group (ASD, NT) × pattern (within, across) ANOVA revealed that NT and ASD participants show strong and equally robust neural discrimination in response to moral violations vs. neutral actions in their RTPJ, LTPJ, and PC, with a main effect of pattern (within > across), no effect of group, and no interaction in all three ROIs, RTPJ: pattern [F(1,36) = 25.9, P < 0.0001], group [F(1,36) = 0.5, P = 0.5], interaction [F(1,36) = 0.9, P = 0.3]; LTPJ: pattern [F(1,35) = 13.3, P = 0.0008], group [F(1,35) = 1.2, P = 0.3], interaction [F(1,35) = 0.3, P = 0.6]; PC: pattern [F(1,35) = 15.1, P = 0.0004], group [F(1,35) = 1.7, P = 0.2], interaction [F(1,35) = 1.18, P = 0.3]. There were no significant effects in DMPFC [pattern: F(1,23) = 2.7, P = 0.1; group: F(1,23) = 0.14, P = 0.7; interaction: F(1,23) = 0.7, P = 0.4].

**SI fMRI Results: Response Magnitude**

See refs. 3 and 4 for analysis of the response magnitude in experiments 2 and 3.

**Experiment 1 (NT).** Averaged over the whole trial, harmful actions elicited a higher response than neutral acts in all four ROIs (RTPJ, LTPJ, PC, and DMPFC; all t > 4.6, P < 0.0003).

In the initial phase only (background, action, outcome), we compared the response to harms vs. neutral acts. In all four ROIs, harmful actions elicited a higher response than neutral acts (all t > 4, P < 0.001). In the final phase only, after the intention had been revealed, in NT adults, the response in RTPJ 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 trend in the same direction [accidental: 0.15 ± 0.04, intentional: 0.08 ± 0.05; t(21) = 1.82, P = 0.08]. There was no difference in PC or DMPFC (both t < 0.2, P > 0.3).

**Experiment 4 (ASD).** Averaged over the whole trial, harmful actions elicited a higher response than neutral acts in all four ROIs (RTPJ, LTPJ, PC, and DMPFC; all t > 3, P < 0.02). In the initial phase only (background, action, outcome), we compared the response to harms vs. neutral acts. In all four ROIs, harmful actions elicited a higher response than neutral acts (all t > 3.5, P < 0.0012). In the final phase only, after the intention had been revealed, in ASD adults, no region showed a significant difference between accidental and intentional harms (all t < 1, P > 0.3).

**Group Comparison.** We directly compared the magnitude of responses across groups (ASD, NT) in three ways. First, we compared the response to all harms vs. all neutral acts, averaged across the whole trial. No region showed a main effect of group or a group by condition interaction (all F < 2.6, all P > 0.12).

Second, we compared the response to all harms vs. all neutral acts during just the initial phase (background, action, and outcome) of each trial. In PC, the response trended toward being overall higher in the ASD group; an ANOVA crossing group (ASD, NT) by condition (harms, neutral) revealed a marginal main effect of group [F(1,35) = 3.37, P = 0.08]. There were no other main effects of group or any group by condition interactions (all F < 1.7, P > 0.2).

Finally, we compared the response to accidental vs. intentional harms specifically during the final phase of the trial. A group (NT, ASD) by condition (accidental, intentional) ANOVA revealed 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 < 1.3, P > 0.25). Fig. S1 shows the PSC for both groups.

**Independent Effects of Magnitude and Pattern.** In NT adults, we observed differences between accidental and intentional harms in the pattern of response over the whole trial and in the magnitude response in the final phase. Specifically in the RTPJ, both of these effects contrasted with the pattern observed in ASD adults, with significant condition by group interactions. When interpreting these results, it is important to know whether pattern and magnitude are two different ways of measuring the same effect, or

*Raw data: Behavioral data (key presses) from experiments 1 and 2 were translated to the same scale for comparison. Specifically, responses from experiment 2 were inverted to make r(106, r(21) = 0.64; the top of the scale for all experiments by subtracting each response from 4 (e.g., a rating of 1 becomes a rating of 3) and then scaling linearly from a three-point scale to a four-point scale.*
converging evidence of two different aspects of atypical RTPJ function in ASD. To address this question, we asked whether these two effects were correlated across individuals. We found that these two effects were not correlated across either NT ($t(20) = 0.11, P = 0.62$) or ASD [$\beta(14) = -0.16, P = 0.55$], suggesting that the effects are sensitive to different aspects of RTPJ function. Moreover, our analysis techniques are sensitive to different features of the data: because correlations are not sensitive to the magnitude of response to intentional vs. accidental harms [accidental - 0.39 ± 0.05, intentional 1.17 ± 0.03; ASD: accidental - 0.23 ± 0.07, intentional 1.01 ± 0.09]. In these subjects, a mixed effects ANOVA crossing group (NT, ASD) by condition (accidental, intentional z-scored ratings) yielded a main effect of condition [$F(1,22) = 304, P < 0.0001$] and a marginal group by condition interaction [$F(1,22) = 3.93, P = 0.059$]. A planned comparison ($t(22) = 1.6, P = 0.06$, one-tailed).

**Behavioral Results.** Behavioral data were available for 11 ASD and 13 NT participants from the matched subsets (NT: accidental - 0.39 ± 0.05, intentional 1.17 ± 0.03; ASD: accidental - 0.23 ± 0.07, intentional 1.01 ± 0.09). In these subjects, a mixed effects ANOVA crossing group (NT, ASD) by condition (accidental, intentional z-scored ratings) yielded a main effect of condition [$F(1,22) = 304, P < 0.0001$] and a marginal group by condition interaction [$F(1,22) = 3.93, P = 0.059$]. A planned comparison ($t(22) = 1.6, P = 0.06$, one-tailed).

**Motion and Artifact.** The subsets of NT and ASD participants did not differ in the number of motion artifacts per run (NT: 4.3 ± 1.2; ASD: 3.6 ± 1.3; $t(27.9) = 0.35, P = 0.7$), in the number of global signal outliers per run (NT: 2.5 ± 0.4; ASD: 2.7 ± 0.3; $t(26.1) = 0.50, P = 0.6$), or in the total vector translation (NT: 0.25 ± 0.02; ASD: 0.23 ± 0.02; $t(28) = 0.73, P = 0.47$).

**Voxelwise Pattern Results. Harm vs. neutral.** As in the full sample, the neural pattern in the RTPJ of both NT and ASD participants discriminated harms from neutral actions (NT: within 1.3 ± 0.11, across 1.2 ± 0.13, $t(14) = 2.0, P = 0.03$; ASD: within 1.5 ± 0.14, across 1.2 ± 0.16, $t(14) = 3.7, P = 0.001$). A group (ASD, NT) by pattern (within, across) ANOVA revealed that matched NT and ASD participants show equally robust neural discrimination in response to harmful vs. neutral actions in their RTPJ, with a main effect of pattern [$F(1,28) = 16.7, P = 0.0003$], no effect of group [$F(1,28) = 0.3, P = 0.5$], and no interaction [$F(1,28) = 1.7, P = 0.2$].

**Accidental vs. intentional.** As in the full sample, the neural pattern in the RTPJ of NT participants discriminated accidental from intentional harms (NT: within 1.1 ± 0.14, across 1.3 ± 0.13, $t(14) = 1.1, P = 0.8$). Matched NT participants discriminated between accidental and intentional harms to a greater extent than ASD participants, reflected in a significant group by pattern interaction [$F(1,28) = 5.9, P = 0.02$], with no main effect of group [$F(1,28) = 1.93, P = 0.18$] or pattern [$F(1,28) = 1.53, P = 0.23$].

**Behavioral and Neural Correlation.** The correlation of pattern discrimination in RTPJ with behavioral responses was significantly larger in NT ($r(11) = 0.56; P = 0.047$) than in ASD participants ($r(9) = 0.54, P = 0.08; z = 2.6, P = 0.009$). Note that the marginal effect in ASD is in the reverse direction: stronger neural discrimination in individuals with more similar behavioral moral judgments.

**Results: ASD Symptom Severity**
Some prior work has found correlations between symptom severity and neural information, either in mean signal (18) or in pattern discriminability (19). However, in this data set, we found no significant correlation between neural pattern and ADOS symptom severity scores (20) in any region (RTPJ: $r(16) = 0.33, P = 0.22$; LTPJ: $r(15) = 0.05, P = 0.85$; PC: $r(15) = 0.14, P = 0.61$; DMFPC: $r(7) = 0.04, P = 0.93$).

**Results: Searchlight MVPA**
In the main text, we mostly focus on the comparison of the full sample of NT ($n = 23$) and ASD ($n = 16$) participants. Here we report the results of the key analyses when conducted in a subset of each group that were matched in pairs. These analyses find the same key pattern of results; importantly, the group by condition interaction in the neural pattern of RTPJ remains significant.

**Participants.** Of the 16 ASD participants, we were able to create one-to-one matches based on sex, age, and IQ for 15 participants, resulting in two groups (NT and ASD), both $n = 15$: 13 male and 2 female. These pairs are matched in age [NT (mean ± SD) = 30 ± 10 y; ASD = 30 ± 8 y; maximum difference between pairs, 7 y; average absolute difference, 3.4 y; $t(26.5) = 0.26, P = 0.8$] and IQ [NT mean = 121 ± 11.8; ASD = 120 ± 12.7; maximum difference, 12 points; average absolute difference, 4.2 points; $t(27.1) = 0.32, P = 0.7$]. As in the full sample, ASD participants scored significantly higher than NT participants on the Autism Quotient questionnaire [AQ: NT (mean ± SD) = 18.5 ± 6.6 y; ASD = 32.5 ± 7.2 y; $t(20.2) = 4.8, P < 0.0001$].
that discriminated \( P < 0.001 \), voxelwise, uncorrected) between accidental and intentional harms, combining across all NT participants \( n = 53 \), experiments 1–3. This unpredicted region is in the fusiform gyrus [peak voxel Montreal Neurological Institute (MNI) coordinates: 30,−54;−14; Fig. S2]. We then median-split participants based on the difference in their moral judgments between intentional and accidental harms (experiments 1 and 2).

In participants showing a larger behavioral effect \( \eta = 15 \), mean z-scored difference in moral judgment = 1.7), the only region that discriminated between intentional and accidental harms was in RTPJ, shown in the lower panel peak voxel MNI coordinates: 52,−60,28; Fig. S2). In the remaining participants \( n = 15 \), mean difference = 1.1), no voxels passed the threshold.

### SI Results: Decoding True vs. False Beliefs?

One open question about the current results is whether the stimulus feature that we successfully decoded is actually accidental vs. intentional harm or another feature of the beliefs in the scenarios that is confounded with this distinction. Two features are confounded with accidental vs. intentional: (i) false beliefs/ignorance vs. true beliefs/knowledge and (ii) good/neutral vs. bad intentions. In all of our stimuli, accidental harms were actions based on a false belief or ignorance about the outcome of the act and a neutral or good intention, whereas intentional harms were actions based on a true belief or foreknowledge about the outcome of the act and a negative intention.

To test which of these features is represented by the patterns of neural response, we analyzed additional data from experiments 2 and 3. In these experiments, in addition to accidental and intentional harms, participants read about failed attempts to harm (i.e., neutral outcome, false belief, negative intention) and neutral actions (i.e., neutral outcome, true belief, neutral intention). Like the comparison between accidental and intentional harms, this contrast thus holds the outcome constant and manipulates whether the belief is true or false and whether the intention is good/neutral or bad.

We found that even when combining data from experiments 2 and 3 (for maximum power, \( n = 30 \)), the patterns in response to failed attempts and neutral actions were not different in RTPJ \( \text{[within = 0.74(0.10), across = 0.78(0.08),} \), \( \eta(29) = −0.8, P = 0.8 \). These results suggest that the pattern distinction observed in RTPJ for accidental vs. intentional harms, in the same participants, is not just due to the difference between true and false beliefs or whether the intention is good/neutral or bad.

---

Fig. S2. (Upper) Whole brain searchlight analysis in NT adults for accidental vs. intentional harms ($n = 53$; $P < 0.001$, uncorrected). Sagittal slice, showing region in left fusiform gyrus and axial slices 40–110. (Lower) Whole brain searchlight analysis in behaviorally sensitive NT adults for accidental vs. intentional harms ($n = 15$; $P < 0.001$, uncorrected). Surface showing region in anterior RTPJ and slices 40–110. Participants were median-split based on the difference in their moral judgments between intentional and accidental harms (experiments 1 and 2). RTPJ is the only region that discriminated intentional and accidental harms in the group, with the largest difference in their ratings of accidental and intentional harms.
<table>
<thead>
<tr>
<th>Scenario</th>
<th>Background</th>
<th>Action</th>
<th>Outcome</th>
<th>Intent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accidental Harm</strong></td>
<td>You are at lunch during school. A classmate approaches you and asks you to show him some moves you learned recently in your martial arts class. You tell him to stand back.</td>
<td>You get ready and perform the martial arts kick.</td>
<td>Your classmate is standing too close, and you end up kicking him in the face. You could not see that your classmate was standing too close, so you kicked.</td>
<td>You could see that your classmate was standing too close, but you kicked anyway.</td>
</tr>
<tr>
<td><strong>Intentional Harm</strong></td>
<td>You are in English class, preparing for the AP test at the end of the year. Your teacher passes out a sample essay that you are to discuss openly and honestly.</td>
<td>You suggest the essay must have been written by a third-grader.</td>
<td>The student who wrote the essay is in your class listening to your critique. The essay was typed, so you completely didn’t realize who had written it.</td>
<td>The essay was handwritten, so you realized right away who had written it.</td>
</tr>
<tr>
<td><strong>Accidental Harm</strong></td>
<td>You and your friend are in the park roller-skating. You skate ahead and sit down behind a tree. You try to get comfortable, but there is a large stick in your way.</td>
<td>You toss the stick aside to make yourself comfortable, and it lands on the park path.</td>
<td>Your friend skates over the stick, and breaks his ankle. You could not see that your friend was about to skate by, so you tossed the stick.</td>
<td>You saw that your friend was about to skate by, and you still tossed the stick.</td>
</tr>
<tr>
<td><strong>Intentional Harm</strong></td>
<td>You are very hungry and decide to go to your favorite fast-food restaurant for lunch. You look through the many choices on the menu and decide to get the chicken sandwich.</td>
<td>You get your food and hungrily eat the whole thing.</td>
<td>While you eat, you spill a little bit of ketchup on your pants. You realize this when it happens and clean it off.</td>
<td></td>
</tr>
<tr>
<td><strong>Neutral</strong></td>
<td>You and a friend are about to give a big presentation in class. You both have been working on it for many weeks and are completely prepared.</td>
<td>You two show up for class and give the presentation.</td>
<td>The teacher enjoys it, especially the graphs you included. You did not know the teacher would enjoy the graphs.</td>
<td></td>
</tr>
</tbody>
</table>

**Timing and length**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Mean</th>
<th>Std</th>
<th>Mean</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accidental Harm</td>
<td>6 sec</td>
<td>0.7 sec</td>
<td>6 sec</td>
<td>0.7 sec</td>
</tr>
<tr>
<td>Intentional Harm</td>
<td>6 sec</td>
<td>0.7 sec</td>
<td>6 sec</td>
<td>0.7 sec</td>
</tr>
</tbody>
</table>

**Fig. S3.** More example stories from experiments 1 and 4. Background information was provided to set the scene and contextualize the protagonist’s action. Outcome described the negative outcome (i.e., harm) of the action. Intent information stated whether the protagonist did not know about the likely negative outcome (accidental harm) or did know about it (intentional harm). Differences between conditions are highlighted in bold text. Neutral stories described actions and outcomes with no moral valence, and the intent information described knowing, or not knowing, about the neutral outcomes. After the story, participants judged how much blame was deserved (an additional 4 s).

**Fig. S4.** Example stimulus from experiment 2. Background information was provided to set the scene and contextualize the protagonist’s action. Foreshadow information foreshadowed whether the action resulted in a negative or neutral outcome. Intent information stated whether the protagonist believed his action would result in a neutral (accidental harm) or negative (intentional harm) outcome. In half of the stimuli, counterbalanced across participants, belief information was presented first and foreshadow information was presented second; in the other half, foreshadow information was presented first and belief information was presented second. Outcome revealed the final effect of the action. Changes between conditions are highlighted in bold text. Each part of the story was presented cumulatively in 6-s blocks. For more examples, see ref. 3.
Fig. S5. Example stimulus from experiment 3. Background information was provided to set the scene and contextualize the protagonist’s action. Intent information stated whether the protagonist believed his action would result in a neutral (accidental harm) or negative (intentional harm) outcome. Outcome revealed the final effect of the action. Changes across conditions are highlighted in bold text. After each item, participants made true/false judgments of factual statements (e.g., “The pond water is fine to wade in”). Sentences were presented cumulatively in 6-s blocks (4).

<table>
<thead>
<tr>
<th>Region</th>
<th>Experiment 1</th>
<th>Experiment 2</th>
<th>Experiment 3</th>
<th>Experiment 4 (ASD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average peak</td>
<td>Average peak</td>
<td>Average peak</td>
<td>Average peak</td>
</tr>
<tr>
<td></td>
<td>size voxel</td>
<td>size voxel</td>
<td>size voxel</td>
<td>size voxel</td>
</tr>
<tr>
<td>RTPJ</td>
<td>260 (17)</td>
<td>151 (18)</td>
<td>106 (22)</td>
<td>194 (24)</td>
</tr>
<tr>
<td></td>
<td>54</td>
<td>56</td>
<td>53</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>−54</td>
<td>−56</td>
<td>−55</td>
<td>−55</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>22</td>
<td>21</td>
<td>24</td>
</tr>
<tr>
<td>LTPJ</td>
<td>144 (26)</td>
<td>121 (21)</td>
<td>178 (23)</td>
<td>156 (35)</td>
</tr>
<tr>
<td></td>
<td>−51</td>
<td>−50</td>
<td>−54</td>
<td>−51</td>
</tr>
<tr>
<td></td>
<td>−56</td>
<td>−63</td>
<td>−58</td>
<td>−60</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>26</td>
<td>19</td>
<td>26</td>
</tr>
<tr>
<td>PC</td>
<td>253 (25)</td>
<td>183 (25)</td>
<td>192 (20)</td>
<td>260 (23)</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>−1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>−58</td>
<td>−58</td>
<td>−58</td>
<td>−54</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>39</td>
<td>37</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>29</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>DMPFC</td>
<td>142 (23)</td>
<td>87 (17)</td>
<td>130 (19)</td>
<td>175 (55)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>−2</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>44</td>
<td>58</td>
<td>54</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>29</td>
<td>36</td>
<td>31</td>
</tr>
</tbody>
</table>