# Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments

Liane Young[a,1], Joan Albert Camprodon[b], Marc Hauser[c], Alvaro Pascual-Leone[b], and Rebecca Saxe[a]

[a]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; [b]Berenson–Allen Center for Noninvasive Brain Stimulation, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA 02215; and [c]Departments of Psychology and Human Evolutionary Biology, Harvard University, Cambridge, MA 02138

Edited* by Nancy G. Kanwisher, Massachusetts Institute of Technology, Cambridge, MA, and approved February 22, 2010 (received for review December 21, 2009)

**When we judge an action as morally right or wrong, we rely on our capacity to infer the actor's mental states (e.g., beliefs, intentions). Here, we test the hypothesis that the right temporoparietal junction (RTPJ), an area involved in mental state reasoning, is necessary for making moral judgments. In two experiments, we used transcranial magnetic stimulation (TMS) to disrupt neural activity in the RTPJ transiently before moral judgment (experiment 1, offline stimulation) and during moral judgment (experiment 2, online stimulation). In both experiments, TMS to the RTPJ led participants to rely less on the actor's mental states. A particularly striking effect occurred for attempted harms (e.g., actors who intended but failed to do harm): Relative to TMS to a control site, TMS to the RTPJ caused participants to judge attempted harms as less morally forbidden and more morally permissible. Thus, interfering with activity in the RTPJ disrupts the capacity to use mental states in moral judgment, especially in the case of attempted harms.**

functional MRI | morality | theory of mind

According to a basic tenet of criminal law, "the act does not make the person guilty unless the mind is also guilty." Like legal doctrine, mature moral judgment depends on the ability to reason about mental states. By contrast, young children's failure to reason fully and flexibly about mental states and, in particular, to integrate mental state information for moral judgment leads them to focus instead on the action's consequences (1–3).

The neural basis of mental state attribution in healthy adults has been investigated using functional MRI (fMRI), implicating a network of brain regions (4), including the medial prefrontal cortex, precuneus, and temporoparietal junction (TPJ). In particular, the right TPJ (RTPJ) shows increased metabolic activity whenever participants read about a person's beliefs in nonmoral (5–8) and moral (9) contexts. However, fMRI cannot identify whether activity in these regions is causally necessary for mental state attribution or, *a fortiori*, for moral judgment.

The current study used offline (experiment 1) and online (experiment 2) repetitive transcranial magnetic stimulation (TMS) to test the hypothesis that normal neural function in the RTPJ allows participants to represent a protagonist's beliefs for moral judgments. We hypothesized that disrupting RTPJ function should reduce the influence of those beliefs on moral judgments. To locate the RTPJ in each participant, we first carried out an fMRI scan, using a functional localizer for brain regions implicated in mental state attribution (7). In a subsequent session, we presented participants with moral scenarios in which (*i*) the protagonist acts on either a negative belief (e.g., that he or she will cause harm to another person) or a neutral belief and (*ii*) the protagonist either causes a negative outcome (e.g., harm to another person) or a neutral outcome (9, 10) (Fig. 1 and *SI Text*). We compared each participant's moral judgments following TMS to the RTPJ and TMS to a control brain region in right parietal cortex.

Experiment 1 used an offline TMS paradigm in which participants received TMS at 1 Hz for 25 min and then read and responded to a series of moral scenarios (Fig. 2, *Upper*). Experiment 2 provided a replication and extension of experiment 1 in a different group of participants. Specifically, to minimize the possibility that the effect of offline TMS might spread, over time, to regions beyond the target RTPJ, experiment 2 used an online paradigm in which participants received short bursts of TMS at 10 Hz for 500 ms, the onset of which was concurrent with the moral judgment for each scenario (Fig. 2, *Lower*). We predicted that in both experiments, TMS to the RTPJ would reduce the role of beliefs and, as a direct result, increase the role of outcomes in participants' moral judgments relative to (*a*) judgments made by the same participants following TMS to the control region and (*b*) judgments made by other participants who received no TMS at all (Fig. S1). Confirming this prediction would provide clear evidence for the causal role of the RTPJ in belief attribution and the essential role of belief attribution in moral judgment.

## Results

**Experiment 1.** To analyze the effect of TMS site on participants' moral judgments, we conducted a 2 (belief: neutral vs. negative) × 2 (outcome: neutral vs. negative) × 2 (TMS site: RTPJ vs. control) repeated measures ANOVA. Following TMS to the RTPJ, moral judgments were less influenced by the actor's beliefs than following TMS to the control site [interaction between belief and TMS site: $F(1,7) = 7.4$, $P = 0.03$, partial $h^2 = 0.51$; Fig. 3, *Upper*]. There were no other main effects or interactions involving TMS site or order of stimulation site, that is, whether the RTPJ or the control site was stimulated first. Also, there were no differences between judgments obtained in the control TMS condition and judgments obtained from a different group of participants with no TMS [e.g., belief by TMS (control TMS vs. no TMS) interaction: $F(1,28) = 0.06$, $P = 0.2$; pilot study described in *Materials and Methods*].

In addition, we conducted an item analysis using each of the 48 scenarios as the unit of analysis instead of each of the eight participants. Only one significant effect emerged: participants judged attempted harms (negative belief, neutral outcome) as more permissible following TMS to the RTPJ vs. the control site [independent samples $t$ test: $t(59) = 2.28$, $P = 0.03$].

Does TMS to the RTPJ only bias the content of moral judgments, or is the time taken to make a moral judgment also affected? There were no effects of TMS site on participants' reaction times

Author contributions: L.Y., J.A.C., M.H., A.P.-L., and R.S. designed research; L.Y. and J.A.C. performed research; L.Y. and R.S. analyzed data; and L.Y., J.A.C., M.H., A.P.-L., and R.S. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]To whom correspondence should be addressed. E-mail: lyoung@mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0914826107/DCSupplemental.

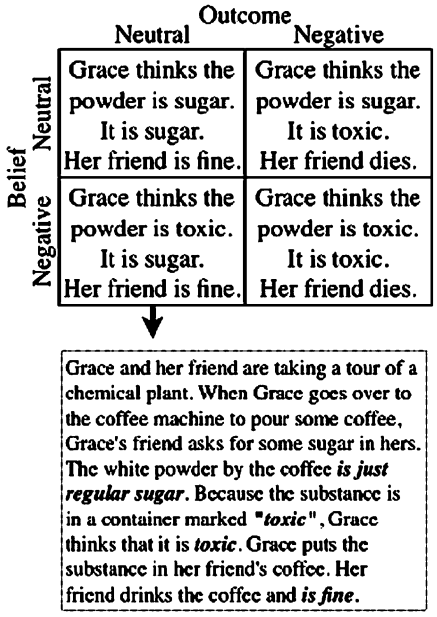


**Fig. 1.** Experimental stimuli and design. (*Upper*) Combination of belief (neutral vs. negative) and outcome (neutral vs. negative) factors yielded a $2 \times 2$ design with four conditions. (*Lower*) Text of a sample "attempted harm" scenario. Bold italicized sections indicate words that differed across conditions.

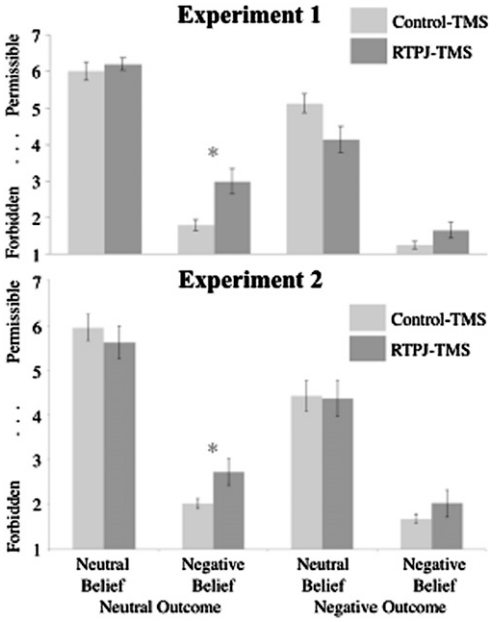


**Fig. 3.** Results for experiment 1 (*Upper*) and experiment 2 (*Lower*). Moral judgments were made on a seven-point scale. Light bars correspond to control TMS, and dark bars correspond to RTPJ TMS. Bars represent SEM. Moral judgments of attempted harm (negative belief, neutral outcome) are significantly different by TMS site (RTPJ vs. control; $*P < 0.05$).

[e.g., belief by TMS site interaction: $F(1,7) = 0.3$, $P = 0.6$]. The only significant effect on reaction times was a belief by outcome interaction [$F(1,7) = 9.6$, $P = 0.02$, partial $h^2 = 0.56$], which reflected the shorter reaction times for intentional harms than for the other conditions (intentional harm, 1.2 s; attempted harm, 1.6 s; accidental harm, 1.8 s; nonharm, 1.6 s). There was also no effect of TMS site on the variability of participants' judgments, as measured by the SD of judgments within a condition across participants [e.g., belief by TMS site interaction: $F(1,7) = 0.1, P = 0.8$].

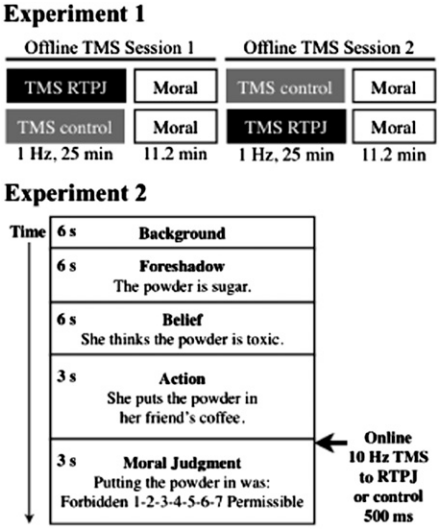


**Fig. 2.** Design for experiment 1 (*Upper*) and experiment 2 (*Lower*). Experiment 1 used an offline TMS paradigm in which participants received TMS at 1 Hz for 25 min and then read and responded to a series of moral scenarios. The order of TMS sessions, RTPJ first vs. control first, was counterbalanced across participants. Experiment 2 used an online TMS paradigm in which participants received TMS at 10 Hz for 500 ms. TMS onset was concurrent with onset of the moral judgment question for each story.

In sum, (*i*) moral judgments following TMS to the control site were no different from a no-TMS control and (*ii*) there was no evidence that TMS site affected the reaction time or variability of judgments in any condition. We therefore conclude that the selective bias in moral judgments induced by TMS to the RTPJ cannot be explained by differences in difficulty between conditions or by the effects of TMS on attention or task performance more generally.

The results of experiment 1 demonstrate that offline TMS to the RTPJ, in comparison to TMS to a nearby control brain region, disrupts participants' use of belief information in moral judgments. As a result, moral judgments appear to be more outcome-based rather than belief-based. Pairwise comparisons in the item analysis showed a pronounced effect for the case of attempted harms, in which the agent believes he or she will harm another but fails to do so. Disrupting RTPJ activity has the selective effect of causing participants to judge attempted harms as more morally permissible than they would normally.

There are two methodological issues that pose a challenge to the interpretation offered thus far. First, information about the potential outcome of the action was available to participants both implicitly before the belief (e.g., the white substance is poison) and explicitly after the belief (e.g., she puts the substance in her friend's coffee, and her friend dies). Offline TMS to the RTPJ may therefore have caused participants to attend to the information presented either most often or most recently, leading to a relative focus on outcomes. Second, offline TMS may have caused the suppression of neural function to spread to distant regions, possibly with some delay (11, 12), from the RTPJ to brain regions closely connected to it (13). Experiment 2 directly addressed these concerns by (*i*) modifying the stimuli to remove the repetition of the outcome information and (*ii*) using brief pulse trains of TMS concurrent with the onset of each moral scenario's question. Specifically, we shortened the train of stimulation (10 Hz for 500 ms) and reduced the time between stimulation and task (application of TMS online during participants' moral judgments), relying on the logic that the shorter the train of TMS and the

shorter the interval between TMS and behavioral testing, the less likely it is for effects on distant regions to be responsible for changes in moral judgments (14). In addition, this experiment allowed us to assess the robustness of our initial findings, using a different TMS paradigm, in a different group of participants.

**Experiment 2.** Disrupting activity in the RTPJ during the task showed a trend toward moral judgments that were less belief-based than judgments made in the TMS control condition [interaction between belief and TMS site: $F(1,11) = 4.6$, $P = 0.056$, partial $h^2 = 0.50$; Fig. 3, *Lower*]. There were no other main effects or interactions involving TMS site or order of stimulation site. Also, judgments obtained in the control TMS condition did not differ from judgments obtained without TMS [$n = 10$; belief by TMS interaction: $F(1,20) = 0.02$, $P = 0.9$]. An item analysis revealed that during TMS to the RTPJ, participants judged attempted harms as more permissible than the same scenarios presented during TMS to the control site [independent samples $t$ test: $t(81) = 2.11$, $P = 0.038$], paralleling the findings from experiment 1.

No effects or interactions involving TMS site were found for reaction time [e.g., belief by TMS site interaction: $F(1,10) = 0.9$, $P = 0.4$], although, overall, negative beliefs elicited shorter reaction times than neutral beliefs [neutral beliefs: 0.70 s; negative beliefs: 0.64 s, $F(1,10) = 21.2$, $P = 0.001$, partial $h^2 = 0.68$]. TMS also did not affect the variability of participants' judgments [e.g., belief by TMS site interaction: $F(1,7) = 0.1$, $P = 0.8$].

The full pattern of results of experiment 2 provides an overall replication of experiment 1 in different participants, using a temporally specific TMS protocol targeting the specific time of moral judgment. In addition, the stimuli were modified so that outcome and belief information was matched for frequency and recency. In both experiments, TMS to the RTPJ diminished the role of beliefs in participants' moral judgments, thereby creating a selective bias toward outcomes.

**Combined Analysis.** A combined analysis of data collected in both experiments 1 and 2 from a total of 20 participants allowed us (*i*) to detect any systematic differences between the two experiments and (*ii*) to take advantage of the increased power to detect any small but consistent effects. Specifically, we conducted a $2 \times 2 \times 2 \times 2 \times 2 \times 2$ ANOVA (belief × outcome × TMS site × order × experiment × gender) of participants' moral judgments. The only significant effects in this combined analysis were main effects of belief [$F(1,12) = 90.5$, $P < 0.001$, partial $h^2 = 0.88$], outcome [$F(1,12) = 110.9$, $P < 0.001$, partial $h^2 = 0.90$], a belief by outcome interaction [$F(1,12) = 5.6$, $P = 0.035$, partial $h^2 = 0.32$], and, critically, the same TMS site by belief interaction found in both experiments 1 and 2 [$F(1,12) = 7.6$, $P = 0.017$, partial $h^2 = 0.38$]. The experiments did not interact with any variable in this analysis. TMS site specifically affected judgments of attempted harms: TMS to the RTPJ vs. the control site resulted in participants' judging attempted harms as more permissible [independent samples $t$ test based on the item analysis: $t(87) = 3.6$, $P = 0.001$].

## Discussion

Transiently disrupting RTPJ activity with offline and online repetitive TMS reduced the influence of beliefs on moral judgments. Normal moral judgment often represents a response to a constellation of features, including not only the agent's beliefs but the agent's desires (15), the magnitude of the consequences (16, 17), the agent's prior record (18), the means used by the agent to cause the harm (17, 19), the external constraints on the agent (e.g., coercion, self-defense) (20), and so on (21). In the current experiments, we manipulated two of these factors, the agent's belief and the outcome of the action, and found that the effect of TMS to the RTPJ was specific to the agent's belief. We found an interaction between TMS site (RTPJ vs. control) and belief (i.e., whether the agent believed he or she would cause harm) in participants' moral judgments and no interaction involving TMS site and outcome (i.e., whether the harm actually occurred).

TMS did not disrupt participants' ability to make any moral judgment. On the contrary, moral judgments of intentional harms and nonharms were unaffected by TMS to either the RTPJ or the control site; presumably, however, people typically make moral judgments of intentional harms by considering not only the action's harmful outcome but the agent's intentions and beliefs. So why were moral judgments of intentional harms not affected by TMS to the RTPJ? One possibility is that moral judgments typically reflect a weighted function of any morally relevant information that is available at the time. On the basis of this view, when information concerning the agent's belief is unavailable or degraded, the resulting moral judgment simply reflects a higher weighting of other morally relevant factors (e.g., outcome). Alternatively, following TMS to the RTPJ, moral judgments might be made via an abnormal processing route that does not take belief into account. On either account, when belief information is degraded or unavailable, moral judgments are shifted toward other morally relevant factors (e.g., outcome). For intentional harms and nonharms, however, the outcome suggests the same moral judgment as the intention. Thus, we suggest that TMS to the RTPJ disrupted the processing of negative beliefs for both intentional harms and attempted harms, but the current design allowed us to detect this effect only in the case of attempted harms, in which the neutral outcomes did not afford harsh moral judgments on their own.

Our hypothesis therefore is that TMS to the RTPJ affects an input to moral judgment (i.e., belief information) but not the process of moral judgment per se. An alternative hypothesis might be that TMS to the RTPJ impaired participants' ability to make moral judgments per se, especially when participants must consider multiple competing factors. On the basis of this alternative account, participants would be worse, following TMS to RTPJ, at integrating information about any two morally relevant factors (e.g., agent's prior record, means used, external constraints on the agent). We do not favor this hypothesis, however, given that it does not predict the direction of our observed effects. If TMS to the RTPJ rendered participants generally worse at combining any two factors in their moral judgments, participants' judgments might have been slower or more variable, which they were not (see below), but not systematically biased, which they were. Nevertheless, this alternative hypothesis deserves further empirical investigation using scenarios featuring other morally relevant features.

TMS to the RTPJ significantly reduced but did not eliminate the role of beliefs in moral judgment. Participants continued to judge accidental harms (neutral belief, negative outcome) as more permissible than intentional harms (negative belief, negative outcome) and attempted harms (negative belief, neutral outcome) as more forbidden than nonharms (neutral belief, neutral outcome) and even accidental harms. This pattern reflects the persistent role of beliefs in their judgments. Previous animal and human studies show that trains of TMS at 1 Hz reduce metabolic activity in the target region by 5–30% (11). Similarly, in previous experiments with human participants, TMS slows or impairs task performance but does not block cognitive task performance completely (22, 23). Consistent with prior estimates, we found that TMS to the RTPJ reduced participants' use of beliefs (by ≈15%) but did not block the use of beliefs completely. In the current scenarios, however, the agents' beliefs dominated participants' moral judgments in the absence of TMS. Other moral scenarios or moral dilemmas exist, for which moral judgments are dominated by other morally relevant factors like the means of the action or the external constraints on the agent. For these scenarios, the initial contribution of beliefs to moral judgment is much smaller; TMS to the RTPJ might therefore eliminate the influence of beliefs altogether. We are testing this hypothesis in ongoing research.

One unpredicted aspect of the current results was the more pronounced effect of TMS on judgments of attempted harms

(negative belief, neutral outcome) than on judgments of accidental harms (neutral belief, negative outcome). Specifically, in analyses of individual conditions, only attempted harms showed an independently significant effect of TMS. Notably, however, there was no significant difference between the effects of TMS on attempted harms and accidental harms (no interaction of belief by outcome by TMS) in any analysis, and for accidental harms, the change in the mean judgment following TMS was in the predicted direction (more forbidden/less permissible) in both experiments. Nevertheless, the hint of asymmetry between attempted and accidental harms is interesting, partly because of its convergence with recent fMRI results. Activity in the RTPJ while participants make moral judgments about attempted and accidental harms shows the same asymmetry: greater activity for attempted than accidental harms (9, 10, 24, 25). The enhanced RTPJ response for attempted harms at the time of judgment appears to reflect enhanced mental state processing for negative moral judgments that rely exclusively on mental state information (9, 10); that is, moral judgment and mental state reasoning appear to interact: Mental states are weighed more heavily when (*i*) they form the predominant basis of moral judgment (e.g., when belief and outcome conflict) and (*ii*) they support negative (as opposed to neutral or positive) moral judgments.

Also of interest is that the hint of asymmetry between attempted and accidental harms appeared to be more pronounced in experiment 1 than in experiment 2, although statistical analyses across both experiments did not reveal any effect of experiment; that is, there was no significant difference between the pattern of results for experiments 1 and 2. Nevertheless, prior fMRI evidence suggests an interpretation of the qualitatively more symmetrical results in experiment 1 (i.e., effects on both attempted and accidental harms) than in experiment 2 (i.e., more pronounced effect on attempted harms). When participants perform the current task in the scanner (i.e., read moral scenarios and then make moral judgments), the RTPJ shows two distinct phases of response: (*i*) a high response to both attempted and accidental harms while participants are first reading the scenarios and (*ii*) as described above, an enhanced response to attempted harms while participants are making moral judgments (10). In the offline paradigm used in experiment 1, TMS effects are expected to be extended in time, including both while participants are reading the scenarios and while participants are making judgments. Thus, we might predict effects of TMS for both conditions. By contrast, in the online paradigm used in experiment 2, TMS was applied only at the moment when participants made their moral judgments, so we might predict relatively greater effects of TMS on judgments of attempted harm. The overall pattern of the current results is thus consistent with the hypothesis that TMS disrupted function in the RTPJ and that the behavioral consequences of this disruption were proportional to the amount of activity previously observed in the RTPJ for each condition and time period. Again, however, these interpretations must be taken lightly because neither the belief by outcome by TMS interaction nor the belief by outcome by experiment interaction was significant in the current data.

The RTPJ was targeted here because of prior neuroimaging evidence that activity in the RTPJ is relatively selective in processing mental states (e.g., beliefs) as opposed to other socially relevant information (8). However, the RTPJ is not the only brain region involved in processing mental states in the context of moral judgment or in other nonmoral contexts. An important consideration for any TMS study, especially if using offline stimulation, is the degree to which the observed effects are specific to the targeted region or reflect the combined result of suppressing function in that region and other regions to which it is differentially connected. Evidence from animal models (26, 27) suggests that the neuromodulatory effects of TMS are maximal on the directly targeted region (11). Nevertheless, the effects of TMS spread to other regions via connections from the target region (12). Our results are thus consistent with the hypothesis that either the (*i*) RTPJ is specifically necessary for belief attribution or (*ii*) the RTPJ and regions to which it is connected are jointly necessary. The RTPJ appears to be strongly connected to other regions implicated in mental state attribution (4, 13, 28–30) and moral cognition (16, 19, 31–34), such as the left TPJ, precuneus, and medial prefrontal cortex. These regions as well as other regions, including the dorsolateral prefrontal cortex, recently implicated in studies on social and moral cognition (35, 36) deserve attention in future research. Importantly, however, TMS generally did not affect moral judgments. Both experiments included an active TMS control site to determine anatomical specificity. The control site was chosen to be close to (5 cm) and in the same (right) hemisphere as the experimental site to control for any secondary nonspecific effects of TMS (e.g., auditory sensations, somatic and tactile stimulation, potential startle effects). Moral judgments in the control TMS condition were no different from moral judgments made in the absence of TMS. Thus, the effects observed here were selective to a specific cortical site — the RTPJ.

Although we consider our results to support the hypothesis that TMS to the RTPJ caused disruption of belief attribution, an alternative hypothesis is that TMS to the RTPJ actually caused disruption of other cognitive functions. Near the right TPJ is a lateral inferior parietal region involved in attentional shifting, one component of the "ventral attention network" (37–39). However, there is significant anatomical separation between our target region and the region involved in attentional processing. Two recent studies have found that the regions associated with belief attribution and attentional reorienting are separated by 10 mm (39, 40). Modeling and experimental work suggest that the spatial resolution of direct TMS stimulation is 5–10 mm (11, 41). Using a functional localizer and image-guided TMS, we targeted the specific region of the RTPJ implicated in mental state attribution. In addition, the behavioral evidence in the current study was not consistent with an effect of attention or any other general effect on task performance (e.g., making participants overall slower, more variable, or generally less able to combine multiple factors when making judgments). TMS to the RTPJ (or the control site) did not render participants more conflicted (i.e., slower moral judgments) or less reliable (i.e., more variable judgments) on any condition. On the other hand, we noted two potentially problematic features of the stimuli of experiment 1: (*i*) participants were presented with outcome information twice and belief information once and (*ii*) participants saw outcome information immediately before making their moral judgments. In experiment 2, we modified the stimuli to remove these concerns and replicated the results of experiment 1. These results support the hypothesis that the effect of TMS to the RTPJ was specific to the content of the stimulus (i.e., belief information).

We therefore interpret the current results as evidence that RTPJ activity is causally implicated in belief attribution and, at least for the scenarios tested, that belief attribution is necessary for typical moral judgment. These results may relate to a pattern commonly observed in moral development in which children up to the age of 6 years rely primarily on the observable outcomes of an action when making moral judgments of the action (3). In fact, young children judge someone who accidentally hurts another person as worse (e.g., "more naughty") than someone who maliciously attempts to hurt another person but fails to do so. Our results suggest that one source of this developmental change may be the maturation of specific brain regions, including the RTPJ (42, 43). Consistent with this idea, recent research suggests that the RTPJ is late maturing (44). In addition, the functional selectivity of the RTPJ for beliefs increases in children from 6 to 11 years old. The link between the maturation of this brain region and moral judgment is an interesting topic for future studies.

The current results may also relate to recent work on neurodevelopmental disorders, such as autism spectrum disorders

(ASDs). Prior studies have found no difference between participants with ASDs and neurotypical controls on various measures of moral judgment. These studies, however, have focused on participants' ability to evaluate intentional violations of moral norms (e.g., harming others) as "bad" and the ability to distinguish moral norms from social norms (e.g., wearing pajamas to school) (45, 46). We suggest that ASDs would lead to impairments in moral judgments, specifically when moral judgment depends on reasoning about an agent's (false) belief, and thus on intact RTPJ function, as in the current experiment. Children with ASDs often show pronounced deficits on nonmoral tasks that depend on considering an agent's belief (47), compared with closely matched control tasks. Even high-functioning adults with ASDs show impaired representations of false beliefs, when measured by their spontaneous-looking behavior (48). Furthermore, reduced capacity for processing mental states in ASDs is associated with reduced RTPJ activity (49). We therefore predict that even high-functioning adults with ASDs would show atypical moral judgments on the kinds of scenarios used in the current study. We are testing this hypothesis in ongoing research.

In sum, both folk moral judgments and legal decisions depend on our ability to look beyond the consequences of an individual's actions to the beliefs and intentions that underlie those actions. In some cases, even if no harm is done, we can "call foul," especially if the individual believed he or she would cause harm by acting and intended to do so. Our experiments show that belief attribution in the service of deciding right and wrong, especially in the case of failed attempts to harm, depends critically on normal neural activity in the RTPJ. When activity in the RTPJ is disrupted, participants' moral judgments shift toward a "no harm, no foul" mentality. Future experiments should explore the relevance of these findings for real-life judgments made by judges and juries, who routinely make very detailed distinctions based on mental state information, such as that between negligence and recklessness (50). Research in this area is likely to inform neural models of moral judgment and moral development in typically developing people and in individuals with neurodevelopmental disorders such as autism (45, 46).

## Materials and Methods

**Experiment 1.** ***fMRI.*** Eight right-handed subjects (aged 18–30 years, five women; *SI Text*) were scanned at 3 T (Athinoula A. Martinos Imaging Center, Michigan Institute of Technology) using twenty-six 4-mm-thick near-axial slices covering the whole brain. Standard echoplanar imaging procedures were used [repetition time = 2 s, echo time = 40 ms, flip angle = 90°]. Subjects participated in four runs of the mental state attribution functional localizer, contrasting stories requiring inferences about a character's beliefs with stories requiring inferences about a physical representation (i.e., an outdated photograph) (7). Fixation blocks of 12 s were interleaved between each story.

fMRI data were analyzed using SPM2 and custom software. Each subject's data were motion-corrected and then smoothed using a Gaussian filter (full-width half-maximum = 5 mm), and the data were high-pass filtered during analysis. A slow event-related design was used and modeled using a boxcar regressor. An event was defined as a single story, with the event onset defined by the onset of text on the screen.

The RTPJ was defined for each subject individually based on a whole-brain analysis of the localizer contrast and as contiguous voxels that were significantly more active ($P < 0.001$, uncorrected) for belief stories vs. physical stories. The peak voxel coordinates in the whole-brain random effects group analysis after normalization onto the Montreal Neurological Institute template were [60, −54, 34]; the average size was 509 $mm^3$.

***TMS.*** Offline TMS sessions took place at the Berenson–Allen Center for Noninvasive Brain Stimulation and the Harvard–Thorndike General Clinical Research Center at Beth Israel Deaconess Medical Center. We used a Magstim SuperRapid biphasic stimulator and a commercially available, air-cooled, eight-shaped, 70-mm coil (MagStim Corporation). The intensity of stimulation was 70% of the stimulator's maximum output for all subjects; the frequency was 1 Hz, and the duration was 25 min. The coil was oriented in the anteroposterior axis with the handle pointing posteriorly.

In most subjects, low-frequency TMS leads to a disruption of activity in the stimulated brain region outlasting the duration of TMS (51, 52). The duration of TMS effects is variable and dependent on experimental conditions, task, stimulation parameters, and characteristics of the subjects. However, the behavioral impact is broadly defined as ranging between 50% and 200% of the time of stimulation (14), and studies in animal models have revealed that 30 min of low-frequency TMS leads to suppression of activity in the target brain region for ≈15 min, with a complete return to baseline after 30 min (11, 12). Offline TMS has the advantage that any general effects of stimulation would not be concurrent with the behavioral experiment. We applied TMS for 25 min. A conservative estimate of the duration of the TMS effects was 12.5 min (50% of the duration). The behavioral task, computer-paced, took 11.2 min. Analyses were performed to determine whether the effects wore off during the session; no within-session effects were found. TMS was applied to the fMRI-defined subject-specific RTPJ in one session and to a control region 5 cm posterior to the RTPJ on the axial plane in the other session, counterbalancing for order of stimulation site (Fig. S2), to control for any nonspecific secondary effects of rTMS. Using Brainsight software (Rogue Industries), we created a 3D reconstruction of the fMRI localizer scan for every subject and graphically represented both the RTPJ and the control region. These individual brain images were used to plan, guide, and monitor the stimulation in real time using a stereotaxic infrared system, ensuring that every TMS pulse was delivered to the predetermined cortical location (53).

***Moral judgment.*** Immediately after stimulation, subjects completed the moral judgment task in each TMS session. Stimuli consisted of four variations of 48 scenarios, for a total of 192 stories with an average of 86 words per story; word count and average reading time were matched across conditions (Fig. 1 and *SI Text*). A 2 × 2 design was used for each scenario, such that protagonists (*i*) produced either a negative or neutral outcome and (*ii*) believed that they were causing either the negative outcome ("negative" belief) or the neutral outcome ("neutral" belief). Moral judgments of these scenarios are determined primarily by the belief (9, 10). Stories were presented in cumulative segments: (*i*) background information (6 s), (*ii*) foreshadow (6 s), (*iii*) belief (6 s), and (*iv*) action (6 s). Stories were then removed from the screen and replaced with a question about the moral permissibility of the action (4 s). Participants made judgments on a scale of 1 (forbidden) to 7 (permissible), using a computer keyboard.

Subjects saw 24 scenarios during each of two 11.2-min sessions (six stories per condition). Stories were presented in a pseudorandom order; the order of conditions was counterbalanced across runs and across subjects. Across subjects, every scenario occurred in each of the four conditions. Individual subjects saw each scenario only once: half after TMS to RTPJ and half after TMS to the control TMS.

**Experiment 2.** Twelve different subjects (aged 18–30 years, seven women) were scanned exactly as in experiment 1. The peak voxel coordinates in the whole-brain random effects group analysis after normalization onto the Montreal Neurological Institute template were [52, −52, 28]; the average size was 151 $mm^3$. TMS sessions took place at the Michigan Institute of Technology and were conducted as in experiment 1, with the following changes: Intensity was 60% of the stimulator's maximum output, the frequency was 10 Hz, and the duration was 500 ms, with the onset of TMS time-locked to be concurrent with the onset of the moral judgment question for each story. The following changes were made to the content and presentation of the moral stimuli: (*a*) the removal of outcome information from the action segment and (*b*) shorter timing (3 s) for the action and judgment segments. Subjects participated in one TMS session, in which they completed six 3.2-min-long runs of the moral judgment task. In each run, subjects were presented with eight stories (two stories per condition). Subjects were allowed minute-long breaks between runs. TMS was applied to the fMRI-defined subject-specific RTPJ in three runs and to a control region in the other three runs, which were interleaved during the session, counterbalancing for order of stimulation site.

**ACKNOWLEDGMENTS.** We thank David Pitcher, Elizabeth Spelke, Alfonso Caramazza, Maurizio Corbetta, Fiery Cushman, Marina Bedny, and Nancy Kanwisher for their helpful comments and David Dodell-Feder and Jonathan Scholz for technical support. This project was supported by the National Center for Research Resources (Grant P41RR14075), Medical Investigations of Neurodevelopmental Disorders and the Athinoula A. Martinos Center for Biomedical Imaging. J.A.C. was supported by the Fundación Rafael del Pino and the Harvard Center for Neurodegeneration and Repair. M.H. was supported by the National Science Foundation–Human Social Dynamics, J. Epstein, and S. Shuman. A.P.L. was supported in part by Grant Number UL1 RR025758 - Harvard Clinical and Translational Science Center, from the National Center for Research Resources and National Institutes of Health grant K 24 RR018875. R.S. was supported by Massachusetts Institute of Technology, the John Merck Scholars program, and the David and Lucile Packard Foundation. R.S. and L.Y. were supported by the Simons Foundation.

1. 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* 103: 37–49.
2. Karniol R (1978) Children's use of intention cues in evaluating behavior. *Psychol Bull* 85:76–85.
3. Piaget J (1932) *The Moral Judgment of the Child* (Free Press, New York).
4. Gallagher HL, Frith CD (2003) Functional imaging of 'theory of mind'. *Trends Cogn Sci* 7:77–83.
5. Gobbini MI, Koralek AC, Bryan RE, Montgomery KJ, Haxby JV (2007) Two takes on the social brain: A comparison of theory of mind tasks. *J Cognit Neurosci* 19:1803–1814.
6. Perner J, Aichhorn M, Kronbichler M, Staffen W, Ladurner G (2006) Thinking of mental and other representations: The roles of left and right temporo-parietal junction. *Social Neuroscience* 1:245–258.
7. Saxe R, Kanwisher N (2003) People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind." *NeuroImage* 19:1835–1842.
8. Saxe R, Powell L (2006) It's the thought that counts: Specific brain regions for one component of Theory of Mind. *Psychol Sci* 17:692–699.
9. Young L, Cushman F, Hauser M, Saxe R (2007) The neural basis of the interaction between theory of mind and moral judgment. *Proc Natl Acad Sci USA* 104:8235–8240.
10. Young L, Saxe R (2008) The neural basis of belief encoding and integration in moral judgment. *NeuroImage* 40:1912–1920.
11. Valero-Cabre A, Payne BR, Pascual-Leone A (2007) Opposite impact on 14C-2-deoxyglucose brain metabolism following patterns of high and low frequency repetitive transcranial magnetic stimulation in the posterior parietal cortex. *Exp Brain Res* 176:603–615.
12. Valero-Cabre A, Payne B, Rushmore J, Lomber S, Pascual-Leone A (2005) Impact of repetitive transcranial magnetic stimulation of the parietal cortex on metabolic brain activity: A 14C-2DG tracing study in the cat. *Exp Brain Res* 163:1–12.
13. Greicius MD, Krasnow B, Reiss AL, Menon V (2003) Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proc Natl Acad Sci USA* 100:253–258.
14. Walsh V, Cowey A (2000) Transcranial magnetic stimulation and cognitive neuroscience. *Nat Rev Neurosci* 1:73–79.
15. Cushman F (2008) Crime and punishment: Distinguishing the roles of causal and intentional analysis in moral judgment. *Cognition* 108:353–380.
16. Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD (2001) An fMRI investigation of emotional engagement in moral judgment. *Science* 293:2105–2108.
17. Cushman F, Young L, Hauser MD (2006) The role of conscious reasoning and intuitions in moral judgment: Testing three principles of harm. *Psychol Sci* 17:1082–1089.
18. Kliemann D, Young L, Scholz J, Saxe R (2008) The influence of prior record on moral judgment. *Neuropsychologia* 46:2949–2957.
19. Greene JD, Nystrom LE, Engell AD, Darley JM, Cohen JD (2004) The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44:389–400.
20. Woolfolk RL, Doris JM, Darley JM (2006) Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition* 100: 283–301.
21. Young L, Nichols S, Saxe R Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology*, in press.
22. Robertson EM, Theoret H, Pascual-Leone A (2003) Studies in cognition: The problems solved and created by transcranial magnetic stimulation. *J Cognit Neurosci* 15: 948–960.
23. Wagner T, Valero-Cabre A, Pascual-Leone A (2007) Noninvasive human brain stimulation. *Annu Rev Biomed Eng* 9:527–565.
24. Young L, Saxe R (2009) Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia* 47:2065–2072.
25. Young L, Saxe R (2009) An FMRI investigation of spontaneous mental state inference for moral judgment. *J Cognit Neurosci* 21:1396–1405.
26. Valero-Cabre A, Rushmore RJ, Payne BR (2006) Low frequency transcranial magnetic stimulation on the posterior parietal cortex induces visuotopically specific neglect-like syndrome. *Exp Brain Res* 172:14–21.
27. Wagner T, et al. (2007) Transcranial direct current stimulation: A computer-based human model study. *NeuroImage* 35:1113–1124.
28. Apperly IA, Samson D, Chiavarino C, Humphreys GW (2004) Frontal and temporo-parietal lobe contributions to theory of mind: Neuropsychological evidence from a false-belief task with reduced language and executive demands. *J Cognit Neurosci* 16: 1773–1784.
29. Adolphs R (2009) The social brain: Neural basis of social knowledge. *Annu Rev Psychol* 60:693–716.
30. Bechara A (2002) The neurology of social cognition. *Brain* 125:1673–1675.
31. Ciaramelli E, Muccioli M, Ladavas E, di Pellegrino G (2007) Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience* 2:84–92.
32. Koenigs M, et al. (2007) Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature* 446:908–911.
33. Mendez MF, Anderson E, Shapira JS (2005) An investigation of moral judgment in frontotemporal dementia. *Cognitive and Behavioral Neurology* 18:193–197.
34. Young L, Cushman FA, Adolphs R, Tranel D, Hauser MD (2006) Does emotion mediate the relationship between an action's moral status and its intentional status? Neuropsychological evidence. *Journal of Cognition and Culture* 6:291–304.
35. Knoch D, Pascual-Leone A, Meyer K, Treyer V, Fehr E (2006) Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science* 314:829–832.
36. Priori A, et al. (2008) Lie-specific involvement of dorsolateral prefrontal cortex in deception. *Cereb Cortex* 18:451–455.
37. Corbetta M, Kincade JM, Ollinger JM, McAvoy MP, Shulman GL (2000) Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nat Neurosci* 3:292–297.
38. Mitchell JP (2007) Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cereb Cortex* 18:262–271.
39. Scholz J, Triantafyllou C, Whitfield-Gabrieli S, Brown EN, Saxe R (2009) Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLoS One* 4:1–7.
40. Decety J, Lamm C (2007) The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *Neuroscientist* 13:580–593.
41. Kammer T (1999) Phosphenes and transient scotomas induced by magnetic stimulation of the occipital lobe: Their topographic relationship. *Neuropsychologia* 37:191–198.
42. Gogtay N, et al. (2008) Three-dimensional brain growth abnormalities in childhood-onset schizophrenia visualized by using tensor-based morphometry. *Proc Natl Acad Sci USA* 105:15979–15984.
43. Blakemore SJ (2008) The social brain in adolescence. *Nat Rev Neurosci* 9:267–277.
44. Saxe RR, Whitfield-Gabrieli S, Scholz J, Pelphrey KA (2009) Brain regions for perceiving and reasoning about other people in school-aged children. *Child Dev* 80: 1197–1209.
45. Leslie AM, Mallon R, DiCorcia JA (2006) Transgressors, victims, and crybabies: Is basic moral judgment spared in autism? *Social Neuroscience* 1:270–283.
46. Blair RJ (1996) Brief report: Morality in the autistic child. *J Autism Dev Disord* 26: 571–579.
47. Baron-Cohen S, Leslie AM, Frith U (1985) Does the autistic child have a "theory of mind"? *Cognition* 21:37–46.
48. Senju A, Southgate V, White S, Frith U (2009) Mindblind eyes: An absence of spontaneous theory of mind in Asperger syndrome. *Science* 325:883–885.
49. Kana RK, Keller TA, Cherkassky VL, Minshew NJ, Just MA (2009) Atypical frontal-posterior synchronization of Theory of Mind regions in autism during mental state attribution. *Soc Neurosci* 4:135–152.
50. Mikhail JM (2007) Universal moral grammar: Theory, evidence and the future. *Trends Cogn Sci* 11:143–152.
51. Chen R, et al. (1997) Safety of different inter-train intervals for repetitive transcranial magnetic stimulation and recommendations for safe ranges of stimulation parameters. *Electroencephalogr Clin Neurophysiol* 105:415–421.
52. Maeda F, Keenan JP, Tormos JM, Topka H, Pascual-Leone A (2000) Interindividual variability of the modulatory effects of repetitive transcranial magnetic stimulation on cortical excitability. *Exp Brain Res* 133:425–430.
53. Gugino LD, et al. (2001) Transcranial magnetic stimulation coregistered with MRI: A comparison of a guided versus blind stimulation technique and its effect on evoked compound muscle action potentials. *Clin Neurophysiol* 112:1781–1792.