Social Cognitive and Affective Neuroscience, 2016, 476–484

doi: 10.1093/scan/nsv131
Advance Access Publication Date: 1 December 2015
Original Article

When minds matter for moral judgment: intent information is neurally encoded for harmful but not impure acts

Alek Chakroff,[1] James Dungan,[1] Jorie Koster-Hale,[2] Amelia Brown,[1] Rebecca Saxe,[3] and Liane Young[1]

[1]Boston College, Department of Psychology, [2]Harvard University, Department of Psychology, and [3]Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences, Cambridge, MA, USA

Correspondence should be addressed to Alek Chakroff, Boston College Deptartment of Psychology, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA. E-mail: alekchakroff@gmail.com.

Abstract

Recent behavioral evidence indicates a key role for intent in moral judgments of harmful acts (e.g. assault) but not impure acts (e.g. incest). We tested whether the neural responses in regions for mental state reasoning, including the right temporoparietal junction (RTPJ), are greater when people evaluate harmful *vs* impure violations. In addition, using multivoxel pattern analysis, we investigated whether the voxel-wise pattern in these regions distinguishes intentional from accidental actions, for either kind of violation. The RTPJ was preferentially recruited in response to harmful *vs* impure acts. Moreover, although its response was equally high for intentional and accidental acts, the voxel-wise pattern in the RTPJ distinguished intentional from accidental acts in the harm domain but not the purity domain. Finally, we found that the degree to which the RTPJ discriminated between intentional and accidental acts predicted the impact of intent information on moral judgments but again only in the harm domain. These findings reveal intent to be a uniquely critical factor for moral evaluations of harmful *vs* impure acts and shed light on the neural computations for mental state reasoning.

Key words: moral judgment; theory of mind; purity; temporoparietal junction

Introduction

Suppose that someone is shot. In one case, the killer aims and fires. In another case, the gun the 'killer' is cleaning for his friend goes off by accident. The difference between murder and manslaughter emerges robustly in the law and in our intuitive moral judgments. Mens rea or 'guilty mind' represents a key element of criminal action (Hart, 1968). Similarly, the agent's intent represents an important factor in everyday moral and social evaluations (Mikhail, 2007). Adult observers typically judge intentional harms as worse than accidental harms (Piaget, 1965/1932; Malle and Knobe, 1997; Knobe, 2005; Borg *et al.*, 2006; Cushman, 2008; Young and Saxe, 2011; Chakroff *et al.*, 2013). Young children follow suit once they develop a theory of mind, the capacity to reason about agents' mental states (e.g. beliefs, intentions) (Killen *et al.*, 2011; Hamlin, 2013).

Recent behavioral research, however, suggests that, while intent plays a critical role in moral judgments of harmful actions (e.g. assault), intent is significantly less important for judging 'impure' acts, concerning food and sex (e.g. ingesting taboo substances or sleeping with blood relatives; Young and Saxe, 2011; Russell and Giner-Sorolla, 2011a; Chakroff *et al.*, 2013). When delivering judgments of impure acts, people focus primarily on the outcome of the act itself (Young and Saxe, 2011), rather than the agent's reasons or intentions. Convergent research has demonstrated that, during moral judgment, people place less exculpatory weight on the circumstances or

Received: 6 August 2015; **Revised:** 29 September 2015; **Accepted:** 12 October 2015

situations that led to impure (*vs* harmful) acts (Russell and Giner-Sorolla, 2011b; Piazza *et al.*, 2013). Rather than search for external reasons or circumstances that led to impure action, people are more likely to explain impure action with respect to the internal dispositions of the agent (Chakroff and Young, 2015), who may also be seen as having a corrupted or tainted character (Chakroff *et al.*, 2013; Rottman *et al.*, 2014; Uhlmann and Zhu, 2014). These findings suggest distinct cognitive signatures for distinct moral domains, as in a number of recent proposals (Graham *et al.*, 2011; Janoff-Bulman and Carnes, 2013; Rai and Fiske, 2011).

Since processing the difference between intentional and accidental harm depends on the capacity to think about another person's thoughts, we suggest that the cognitive process of theory of mind is engaged for moral evaluations of harms more so than for moral evaluations of impure acts. Note that this proposal differs subtly from alternative proposals in which theory of mind is fully deployed for evaluating both harmful and impure acts, and participants encode intent for both domains to the same extent but simply decide to assign less moral weight to intent in the case of impure acts.

Adjudicating between these proposals using purely behavioral data is challenging; however, tasks that rely on theory of mind have been shown to elicit a distinctive neural signature. One critical test then is to determine whether the neural mechanisms that support theory of mind are preferentially engaged for moral judgments of harmful *vs* impure acts. Moreover, we can ask whether there is evidence for the key cognitive computations (i.e. theory of mind) in the harm domain but not the purity domain—do these neural mechanisms support the representation of an act as intentional or accidental depending on the moral domain? The hypothesis we favor is that distinct moral violations engage theory of mind to differing degrees, and that neural evidence for the computation of intent will be present during the processing of harmful, but not impure acts.

Functional magnetic resonance imaging (fMRI) research has implicated a consistent network of brain regions for theory of mind and social cognition, including the right and left temporoparietal junction (RTPJ, LTPJ), precuneus (PC) and medial prefrontal cortex (MPFC) (Fletcher *et al.*, 1995; Gallagher *et al.*, 2000; Saxe and Kanwisher, 2003; Dodell-Feder *et al.*, 2011; Waytz *et al.*, 2012; Carter and Huettel, 2013). For example, the response in these regions is higher when participants read stories describing or requiring inferences about mental states such as false beliefs, compared with when participants read stories about physical states such as false or outdated signs, maps or photographs (Saxe and Kanwisher, 2003; Perner *et al.*, 2006; Gobbini *et al.*, 2007). Of these regions, the RTPJ appears to be especially selective (Perner *et al.*, 2006); for example, the RTPJ is recruited not only for mental states over physical states but also for mental states over other socially relevant information such as socially salient physical traits or bodily sensations (Saxe and Powell, 2006).

Moral judgment tasks also consistently elicit RTPJ activity (Borg *et al.*, 2006; Greene *et al.*, 2004; Moran *et al.*, 2011; Decety and Cacioppo, 2012; but see Parkinson *et al.*, 2011). When we encounter moral agents, we often ask ourselves: What was he thinking? Did she mean to do it? Was that an accident or on purpose? The RTPJ is robustly recruited for encoding explicit mental states for moral judgment and integrating mental states with morally relevant outcomes (Young and Saxe, 2008, 2009b), for the spontaneous inference of mental states in morally relevant contexts (Young and Saxe, 2009a) and for the extra processing of mental states for 'intuitive prosecution' (Young *et al.*, 2011). Temporarily disrupting activity in the RTPJ also disrupts mental state reasoning for moral judgment, rendering moral judgments more outcome-based (Young *et al.*, 2010a). This study probes the functional profile of the network of regions implicated in theory of mind; however, our strongest hypotheses concern the RTPJ, given prior evidence of its specificity for mental state processing in moral judgment. Specifically, we hypothesize that the RTPJ response should be higher for moral judgments of harm *vs* purity violations.

Moreover, if moral evaluation of harmful actions is sensitive to specific features of mental state content (e.g. intent), but moral evaluation of purity violations is not, then we should expect to see different computational signatures, or sensitivity to these dimensions, in the neural response elicited by each of these domains within the neural network for theory of mind. Although prior fMRI studies have shown the RTPJ to be recruited robustly for different components of mental state reasoning, across different tasks, the average RTPJ response (i.e. the mean signal change, across all voxels in the region) appears to be unaffected by a wide range of manipulations of specific features of mental states, including whether beliefs are true *vs* false (Jenkins and Mitchell, 2009), justified *vs* unjustified (Young *et al.*, 2010c), positive *vs* negative (Kliemann *et al.*, 2008), surprising *vs* unsurprising (Young *et al.*, 2010b), and whether inferences about mental states are 'constrained' *vs* 'open-ended' (Jenkins and Mitchell, 2009).

However, recent work using multivoxel pattern analysis (MVPA) has shown that it is possible to decode stimulus features or categories based on patterns of activity within regions, even in the absence of differential mean blood oxygen level dependent (BOLD) responses (Haxby *et al.*, 2001; Bedny *et al.*, 2011; Fedorenko *et al.*, 2012; Hsieh *et al.*, 2012), and to relate pattern information to differences in participant behavior (Haynes and Rees, 2006; Norman *et al.*, 2006; Raizada *et al.*, 2010). A recent study conducted using a subset of the data reported here has demonstrated that, in neurotypical adults, the pattern of activity in RTPJ distinguishes between intentional and accidental harms (Koster-Hale *et al.*, 2013). This study extends these analyses to voxel patterns associated with intentional and accidental *impure* acts. We expected the intentional-accidental dimension to be encoded in its voxel-wise pattern, but selectively for harmful acts, and not impure acts. Furthermore, we expected that the degree of pattern discrimination would predict the weight assigned to intent by participants, as in Koster-Hale *et al.* (2013), but that this relationship would be specific to harmful acts and not emerge for intentional and accidental impure acts.

This study tested these hypotheses by measuring both the average magnitude of response and the voxel-wise pattern in brain regions implicated in theory of mind. Participants read stories about harmful and impure acts and delivered judgments of moral wrongness in the scanner. We hypothesized that (i) as in prior behavioral work, intent would exert a greater influence on moral judgments of harmful *vs* impure acts, (ii) the RTPJ would be recruited more for evaluating harmful *vs* impure acts, (iii) the voxel patterns in the RTPJ would differentiate between intentional and accidental harmful acts but not impure acts and (iv) individual differences in voxel pattern discrimination in RTPJ would predict the weight placed on intent during moral judgments of harmful acts but not impure acts.

Material and methods

Participants and procedures

Twenty-three right-handed college undergraduate students (mean age = 27 years, seven women) participated in the study for

payment. All participants were native English speakers, had normal or corrected-to-normal vision and gave written informed consent in accordance with the requirements of Internal Review Board at MIT. Participants were scanned at 3T (at the MIT scanning facility in Cambridge, MA) using twenty-six 4-mm near-axial slices covering the whole brain, and 3×3-mm in-plane resolution. Standard echoplanar imaging procedures were used (TR = 2 s, TE = 40 ms, flip angle 90°). A subset of the data collected for this study were analyzed and reported previously (i.e. harmful conditions only; Koster-Hale *et al.*, 2013) alongside reanalyses of other previously published datasets; here we report results from the full dataset.

Participants were scanned while reading 60 stories: 12 intentional harmful acts, 12 accidental harmful acts, 12 intentional impure acts, 12 accidental impure acts and 12 neutral stories (see Supplementary information for full text of all stimuli). Harm stories included both physical harms (e.g. poisoning someone) and psychological harms (e.g. humiliating someone). Impure stories included both sexual violations (e.g. sex with a blood relative) and pathogen violations (e.g. eating maggots). We used three categories of mental state verbs to describe intent, applied equally across domains: [knew/thought], [realized/discovered], [saw/ noticed]. Stories were presented in cumulative segments:

1. Background (6 s)—e.g. the protagonist is at a party
2. Action (4 s)—e.g. the protagonist has sex
3. Outcome (4 s)—e.g. the sexual partner is a sibling/stranger
4. Intent (4 s)—e.g. the protagonist acted intentionally/accidentally
5. Judgment (4 s)—the question appears alone on the screen

In the scanner, for each story, participants made moral judgments of the action on a 4-point scale, 'not at all morally wrong' (1) and 'very morally wrong' (4), using a button box. Moral judgment data were not collected for three participants.

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

In the same scan session, subjects participated in four runs of a theory of mind functional localizer task (belief *vs* nonbelief). The task contrasted mental state stories requiring inferences about false beliefs with control stories requiring inferences about 'false' physical representations, i.e. a photograph or map that had become outdated. Control stories were matched for linguistic complexity and logical structure. The stimuli and story presentation for this task were exactly as described in Saxe and Kanwisher (2003), Experiment 2.

fMRI data analyses

MRI data were analyzed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm) and custom software. Each subject's high resolution T1-weighted structural scan was coregistered and normalized to Montreal Neurological Institute (MNI) brain space, and the parameters from this process were used to normalize each subject's motion-corrected EPI images into MNI space. Images were smoothed using a Gaussian filter (full width half maximum = 5 mm), and high-pass filtered during analysis. A slow event-related design was used and modeled using a boxcar regressor to estimate the hemodynamic response for each condition. An event was defined as a single story, and the event onset was defined by the onset of text on screen. The timing of story components was constant for every story, so independent parameter estimates could not be created for each component. Components were instead separated based on the time of the response corresponding with each story segment, accounting for hemodynamic lag.

Both whole-brain and tailored regions of interest (ROI) analyses were conducted. We defined ROIs for each subject individually based on a whole brain analysis of the independent theory of mind localizer. We defined ROIs as contiguous voxels significantly more active while the subject read belief stories, as compared with control stories.

Mean ROI responses

Within each ROI, we averaged across voxels to extract a single time course of BOLD response. We calculated a baseline value as the average ROI response across all inter-stimulus time points, excluding the first 4 s after each stimulus offset, to allow the hemodynamic response to settle. Then, for each condition, we calculated an average stimulus-locked BOLD time course, averaging across all blocks in the condition. These condition time courses were expressed in terms of percent signal change (PSC) relative to the baseline: PSC = 100 * (time course – baseline) / baseline. In each ROI, PSC during story segment presentation (adjusted for hemodynamic lag) was compared across conditions (Poldrack, 2006).

Multivoxel pattern analysis. Each participant's data were divided into even and odd runs and, within these partitions, the mean response (beta value) of every voxel in the individual ROI was calculated for each condition, mean centered based on the voxel's average response across conditions. Voxel patterns were correlated over even and odd runs, both *within* and *across* conditions. An index of classification was calculated for each condition pair as the z-scored within-condition correlation minus the z-scored across-condition correlation. A region was categorized as successfully classifying a category of stimuli if, across individuals, the within-condition correlation across voxels was significantly higher than the between-condition correlation (Haxby *et al.*, 2001).

Results

Behavioral results

A 2 (domain: harmful *vs* impure) × 2 (intent: intentional *vs* accidental) within-subjects ANOVA revealed that participants judged harmful violations as worse than impure violations (main effect of domain: $F(1,19) = 5.27$, $P = 0.03$, $\eta_p^2 = 0.22$), and intentional violations as worse than accidental violations (main effect of intent: $F(1,19) = 210.9$, $P < 0.001$, $\eta_p^2 = 0.92$). As in prior work (Young and Saxe, 2011; Russell and Giner-Sorolla, 2011a,b; Chakroff *et al.*, 2013), this intent effect varied in magnitude across moral domains (intent × domain interaction: $F(1,19) = 11.57$, $P = 0.003$, $\eta_p^2 = 0.38$): the effect was greater for judgments of harmful acts (intentional: 3.13, accidental: 1.41; $t(19) = 18.55$, $P < 0.001$) than impure acts (intentional: 2.62, accidental: 1.43; $t(19) = 7.74$, $P < 0.001$). The same pattern replicated across scenarios, in an item-wise

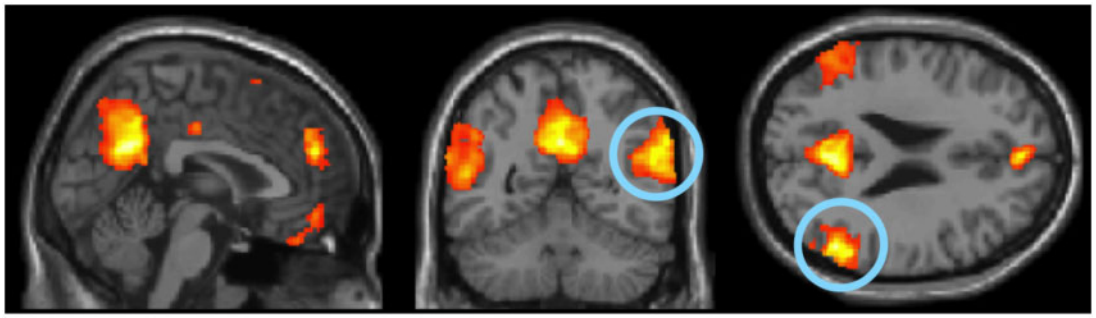

Fig. 1. Theory of mind localizer contrast of Belief > Photo, group random effects analysis, $k > 10$, $P < 0.001$, uncorrected, x = 2, y = −56, z = 24. RTPJ highlighted.

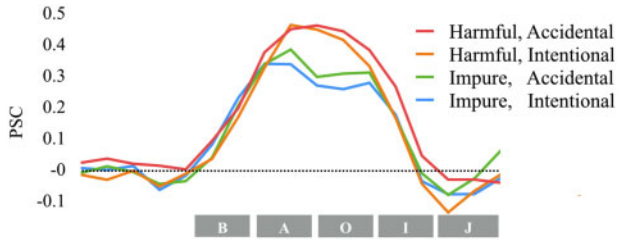

Fig. 2. Temporal BOLD response in the RTPJ for intentional and accidental harmful and impure acts, normalized as PSC from baseline. Time window labels: B = Background, A = Action, O = Outcome, I = Intent, J = Judgment.

analysis (intent × domain interaction: $F(1,11) = 6.3$, $P = 0.03$, $\eta_p^2 = 0.36$).

fMRI results

Theory of mind localizer. A whole-brain random effects analysis of BOLD response for belief *vs* non-belief replicated previous results (Saxe and Kanwisher, 2003; Dodell-Feder *et al.*, 2011), peak voxel MNI coordinates: RTPJ ($x = 58$, $y = -50$, $z = 28$), LTPJ (−56, −60, 26), PC (2, −60, 28), MPFC (2, 50, 24), Figure 1. We identified ROIs in individuals at the same threshold of $P < 0.001$, uncorrected: RTPJ (identified in 23 of 23 subjects), LTPJ (22/23), PC (23/23) and dorsal MPFC (DMPFC) (18/23). We calculated the average PSC from rest in each ROI over the entire story time course (i.e. background through judgment) and also separately for each story segment: action (e.g. the protagonist has sex), outcome (e.g. the sexual partner is a sibling/stranger) and *intent* (e.g. the protagonist knew this/didn't know this).

Contrasts of domain and intent. As displayed in Figure 2, and reported in Supplementary Table S1 in Supplementary information, a whole-brain random effects analysis of BOLD response (false-discovery rate [FDR] corrected $P < 0.05$, $k > 10$) for domain revealed activations for harmful > impure acts in the insula and TPJ bilaterally as well as the supplementary motor area. The contrast of impure > harmful acts revealed a left lateralized set of activations in frontal, temporal and parietal areas as well as in the PC. In the Supplementary information, we also report contrasts of harmful > neutral and impure > neutral, as well as their conjunction containing extensive shared activations, which may inform debates, outside the scope of this article, regarding neural systems devoted to the processing of distinct kinds of specifically immoral (*vs* morally neutral) actions (Borg *et al.*, 2008; Parkinson *et al.*, 2011). Finally, when comparing accidental and intentional conditions, no voxels reached significance at $P < 0.05$, FDR corrected, $k > 10$. A more lenient threshold of $P < 0.001$, uncorrected, $k > 10$, revealed a cluster in the left occipital lobe for accidental > intentional, but the converse contrast remained non-significant.

Mean ROI responses. We performed a 2 (domain: harmful *vs* impure) × 2 (intent: intentional *vs* accidental) within-subjects ANOVA of the RTPJ response, averaged across the entire time course. Central to our key hypothesis, the mean response was higher for harmful *vs* impure acts ($F(1,22) = 11.76$, $P = 0.002$, $\eta_p^2 = 0.35$; Figure 3); no other comparisons were significant within RTPJ. We found the same pattern when restricting our analysis to the 4 s of the stimulus when only outcome information was presented, before intent information was presented (see Supplementary information for analyses broken down by each story segment). The BOLD response was marginally higher for harmful *vs* impure acts in the LTPJ ($F(1,21) = 3.77$, $P = 0.07$, $\eta_p^2 = 0.15$) but not in the PC ($F(1,22) = 0.003$, $P = 0.96$) or the DMPFC ($F(1,17) = 0.07$., $P = 0.80$)[1]. Notably, in the RTPJ, the difference in magnitude between harmful and impure acts was not predictive of the difference in the effect of intent for moral judgments of harmful *vs* impure acts ($r(17) = -0.12$, $P = 0.63$).

Convergent with the ROI analyses, a whole-brain conjunction analysis between the ToM localizer and the harmful > impure contrasts revealed overlap in the RTPJ, LTPJ, and PC (Figure 4 and Table 1). Each voxel was considered overlap only if its value reached significance independently for each contrast at $P < 0.05$, FDR corrected, $k > 10$. Conjunction images are provided for illustrative purposes, and not to argue for a complete congruence between the ToM network and the regions revealed in the harmful > impure contrast, which we see as partially overlapping (insofar as moral judgments of harms rely on ToM) but distinct.

Multivoxel pattern analysis. As reported in Table 2, analyses revealed a separation in the pattern of response for intentional *vs* accidental acts, but only for harmful acts, and only in the RTPJ. No ROI distinguished between intentional and accidental impure acts. Across participants, the degree to which the RTPJ discriminated between intentional and accidental violations predicted the degree that participants judged intentional harms as worse than accidental harms ($r(17) = 0.6$, $P = 0.006$), as reported in Koster-Hale *et al.* (2013). This correlation was not significant for impure acts ($r(17) = 0.2$, $P = 0.41$). Notably, the degree

[1] Additional analyses on anatomically defined insula ROIs revealed higher response for harmful versus impure acts, both in left insula ($F(1,22) = 10.89$, $P = 0.003$, $\eta_p^2 = 0.33$) and right insula ($F(1,22) = 16.38$, $P = 0.001$, $\eta_p^2 = 0.43$), but no difference by intent, and no interaction.

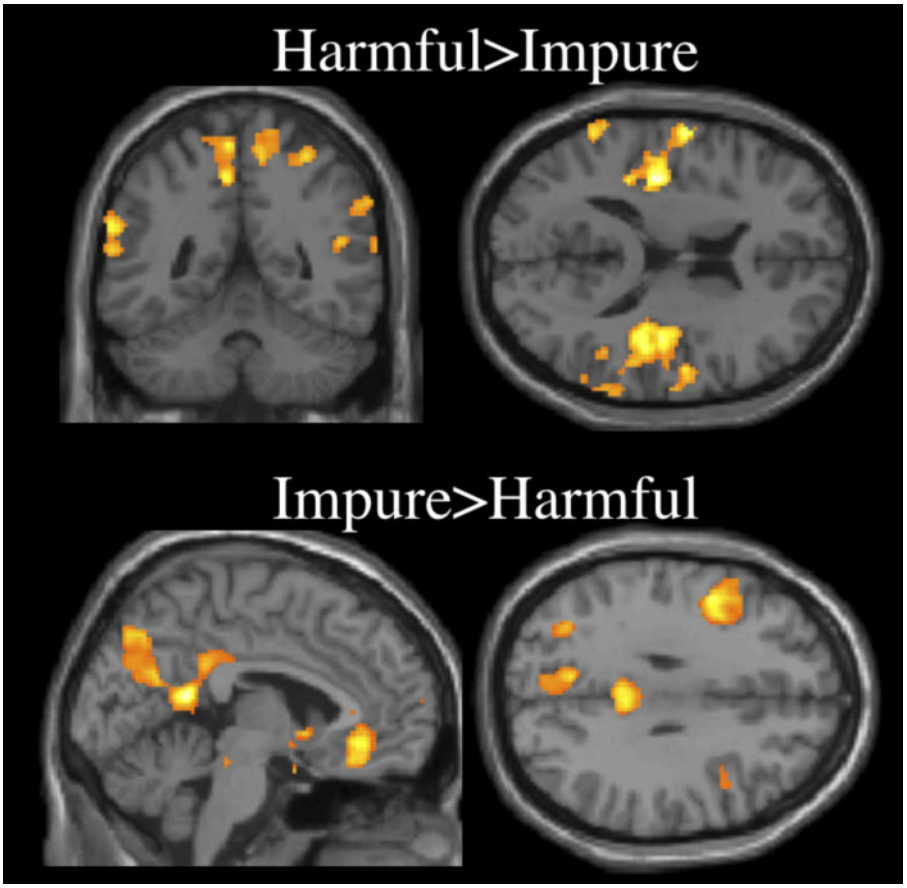

Fig. 3. Whole brain contrast of (A) Harmful > Impure, $y = -49$, $z = 16$, and (B) Impure > Harmful, $x = -4$, $z = 31$. Group random effects analysis, $P < 0.05$, FDR uncorrected.

of pattern discrimination was not significantly positively correlated with the mean PSC in RTPJ (harmful: $r(21) = -0.29$, $P = 0.18$; impure: $r(21) = 0.12$, $P = 0.60$), suggesting that domain difference in pattern discrimination was not driven by increased 'signal' in RTPJ for harmful *vs* impure acts. Finally, collapsing across intentional and accidental acts, there was a separation in the pattern of response to harmful *vs* impure acts in the RTPJ ($F(1,22) = 7.50$, $P = 0.01$, $\eta_p^2 = 0.25$), LTPJ ($F(1,21) = 11.07$, $P = 0.003$, $\eta_p^2 = 0.34$) but not DMPFC ($F(1,17) = 2.75$, $P = 0.12$, $\eta_p^2 = 0.14$). There was also a strong separation of patterns in PC ($F(1,22) = 24.60$, $P < 0.001$, $\eta_p^2 = 0.53$) despite an absence of mean BOLD differentiation in this ROI.

Discussion

The widespread role of mental state reasoning in moral judgment can be seen across disciplines, including philosophy, law, psychology and neuroscience. Yet here, using behavioral and neuroimaging methods, we show that people are not always mindful of moral minds—people encode an agent's beliefs and intentions more for some kinds of actions over others. The RTPJ, identified using an independent localizer task, was preferentially recruited when participants judged harmful *vs* impure acts, and its voxel-wise response contained information about intent in the case of harmful but not impure acts.

This study builds on a prior body of work implicating the RTPJ in mental state reasoning and moral judgment. Activity in the RTPJ correlates with the use of mental states for evaluating harmful actions: people with a high RTPJ response judge accidental harms more leniently (Young and Saxe, 2009a). Meanwhile, transiently disrupting RTPJ activity using transcranial magnetic stimulation reduces the role of mental states in moral judgment; as a consequence, moral judgments rely more on other factors, such as the action's outcome—whether or not harm was done (Young *et al.*, 2010a). Together, these findings support a critical role for the RTPJ in the neural network that supports reasoning about the minds of moral agents, especially for harmful acts.

Notably, the enhanced response in the RTPJ for harmful *vs* impure acts was observed as soon as an act was revealed to be harmful *vs* impure (e.g. the powder was poisonous/the sexual partner was a sibling; see Supplementary information). Thus, the neural difference appeared even before the presentation of explicit mental state information (e.g. whether the violation was committed intentionally or accidentally). Consistent with this finding, prior work has shown robust RTPJ recruitment for evaluating harmful actions in the absence of explicit mental state information (Kliemann *et al.*, 2008; Young and Saxe, 2009a). When reading about morally relevant actions, participants may be provoked to think about the agent's beliefs and intentions. Accordingly, the RTPJ response at this stage could reflect the process of spontaneously thinking about an agent's thoughts, without constructing a specific representation of the belief or intent.

Could the enhanced RTPJ response to harmful *vs* impure acts track with an overall increase in salience related to moral

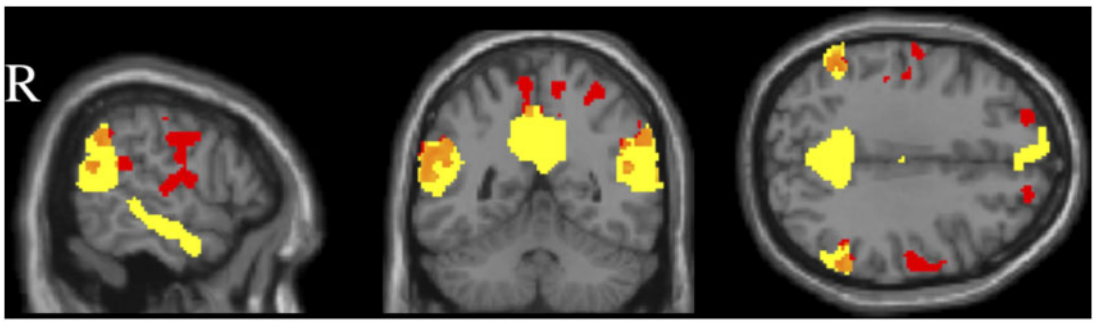

Fig. 4. Conjunction (orange) between the independent contrasts of *Belief > Photo* (Yellow), and *Harmful > Impure* (Red), each viewed at a threshold of $P < 0.05$, FDR corrected. Viewed at $x = 61$, $y = -52$, $z = 33$.

Table 1. Brain regions, X, Y, Z coordinates, and cluster sizes (k) for the conjunction between the independent contrasts of *Belief > Photo* and *Harmful > Impure*, each viewed at a threshold of $P < 0.05$, FDR corrected.

Region	X	Y	Z	k
LTPJ	−60	−52	12	512
Left middle temporal	−52	−28	−16	214
PC	−6	−46	44	112
RTPJ	48	−48	14	97
RTPJ	64	−54	30	89
Left insula/temporal	−44	20	−18	48
RTPJ	66	−50	12	41
RTPJ	48	−52	26	15

Table 2. MVPA results

	Region	*Within*	*Across*	*F*	df	*P*	η_p^2
Intentional vs accidental harmful	RTPJ	1.13(0.10)	1.05(0.11)	5.27	1,22	0.03	0.19
	LTPJ	1.28(0.06)	1.36(0.05)	2.56	1,21	0.12	0.11
	PC	0.96(0.08)	0.99(0.08)	0.36	1,22	0.55	0.02
	DMPFC	0.91(0.08)	0.87(0.11)	0.52	1,17	0.48	0.03
Intentional vs accidental impure	RTPJ	1.00(0.09)	0.97(0.09)	0.7	1,22	0.41	0.03
	LTPJ	1.26(0.07)	1.24(0.07)	0.12	1,21	0.73	0.01
	PC	0.94(0.10)	0.88(0.08)	1.44	1,22	0.24	0.06
	DMPFC	1.00(0.13)	1.05(0.09)	0.28	1,17	0.6	0.02

Within and *Across* z-scored correlation coefficients, mean (standard error), for accidental *vs* intentional acts. While RTPJ, LTPJ and PC discriminated between harm and purity violations, only RTPJ discriminated between accidental and intentional harms and no region distinguished between accidental and intentional purity violations.

wrongness? While participants judged harmful acts as more morally wrong than the impure acts, as in our prior work (Chakroff *et al.*, 2013; Chakroff and Young, 2015), impure acts may be seen as less common, more weird (e.g. bizarre, strange, not simply uncommon) and more disgusting (Chakroff and Young, 2015), all of which could lead to an overall increase in saliency for impure *vs* harmful acts. We would not necessarily predict that saliency related to moral wrongness in particular would drive activation in the ToM network, let alone the RTPJ specifically. Furthermore, we did not find that moral judgments were predictive of BOLD signal in RTPJ, a natural prediction that would follow from a saliency account of the present results.

The present work builds on recent findings demonstrating that the RTPJ encodes intent only in stable voxel-wise patterns of activity, rather than mean activity across voxels (Koster-Hale *et al.*, 2013). Here, we demonstrate that this pattern discrimination is present when participants are presented with harmful acts but not impure acts. Convergent with the observed patterns of neural activity, the behavioral data reflected a greater sensitivity to the difference between intentional and accidental harms compared with intentional and accidental purity violations, as in prior work (Young and Saxe, 2011; Russell and Giner-Sorolla, 2011a,b; Chakroff *et al.*, 2013). Here we also found that individual differences in pattern discrimination in RTPJ predicted the weight given to intent during moral judgments (as reported in Koster-Hale *et al.*, 2013), but only for judgments of harmful acts, and not impure acts. Importantly, this pattern did not emerge from analyses of the average magnitude of response in the RTPJ. Although the average response was higher for harmful *vs* impure acts, the average response magnitude did not predict behavioral sensitivity to intent. Together these results suggest that the pattern discrimination found for harmful but not impure acts was not driven by overall increased signal in the relatively high average response to harm. Indeed, across participants, pattern discriminability was not predicted by the average magnitude of response relative to baseline.

Why might encountering a harmful agent elicit more mental state reasoning than encountering an impure agent? On one account, harm norms and purity norms serve different adaptive functions. We may wish to avoid harmful agents on the basis of

what they might do to us in the future. Reliable predictions about 'friend or foe' require information about intent: an agent who causes harm intentionally (*vs* accidentally) will be more likely to do so again at the next encounter. Harm norms and their enforcement via punishment may therefore function to prevent us from hurting each other, preserving social order (Sheikh and Janoff-Bulman, 2010). Put another way, harm norms function over moral 'dyads' which consist of an 'agent' and a 'patient' (Gray and Wegner, 2009; Waytz *et al.*, 2010b). The patient, along with other observers of the agent's action, may wish to establish innocence or guilt, friend or foe, by considering the contents of the agent's mind.

In contrast, purity norms appear to lack true 'victims' in this sense—the agent and patient are often one and the same (Haidt, 2001). Purity norms may be in place to protect ourselves from a range of self-destructive behaviors—from sleeping with our siblings to consuming contaminants, for example (Chapman *et al.*, 2009; Graham *et al.*, 2011). As such, intent matters less (Appiah, 2006; Young and Tsoi, 2013). Usually our aim is simply to avoid bad outcomes for ourselves—even in the absence of possible punishment. In addition, we know our own 'friend or foe' status; usually, we are not our own enemy. Thus, it may also be the case that we do not spontaneously reflect on our own intentions (Gweon *et al.*, 2011). In this study, all moral scenarios were presented in the second person but future imaging work should directly investigate the role of intent in moral judgments of self-*vs* other across domains (Kedia *et al.*, 2008; Decety and Porges, 2011; Chakroff *et al.*, 2013).

A related account is that purity norms may help us avoid individuals who are 'impure' by our local standards, on the basis of their external actions (rather than internal mental states). Purity norms may thus reinforce group boundaries, while harm norms help us avoid enemies with ill intent, whether they are within or outside our group (Jost *et al.*, 2008; Graham *et al.*, 2011; Janoff-Bulman and Carnes, 2013). Individuals made 'impure' by their actions may even be subsequently dehumanized, and elicit less spontaneous mental state reasoning (Bandura, 1999; Harris and Fiske, 2006; Waytz *et al.*, 2010a) and harsher moral judgment (Wheatley and Haidt, 2005; Chapman *et al.*, 2009; Inbar *et al.*, 2009).

It is important to differentiate the above *adaptive* accounts from the following *processing* account of the present results. We propose that the pattern of results observed here does not merely reflect 'other-focused' *vs* 'self-focused' mental state reasoning for judgments of harmful *vs* impure acts, respectively. First, we note that in prior behavioral work the role of intent in moral judgments does not depend on the perspective taken by participants toward the act: self-focused (*you* did X) *vs* other-focused (*Sam* did X) (Young and Saxe, 2011). Second, in prior neuroimaging work, participants were asked to think about their own or another's mental states and no reliable differences in activity were found in the ToM network across the self and other conditions (Gweon *et al.*, 2011). Although the MPFC in particular has been shown to differentiate between judgments of one's own *vs* others' dispositions or preferences (Heatherton *et al.*, 2006; Mitchell *et al.*, 2006), we found no evidence of differentiation between harmful and impure acts within the DMPFC, either based on mean BOLD response or patterns of activation. Thus, we argue that, in the present work, activity in the ToM network does not reflect greater other-focused mental state reasoning for harmful acts and self-focused reasoning for impure acts.

Although mental states dominate both the law and folk morality, this study shows that the neural processes supporting mental state reasoning are engaged preferentially for considering harmful agents, compared with agents who violate purity norms concerning food and sex. Notably, strict liability, liability for which *mens rea* (guilty mind) need not be proven, is rare in criminal law; however, in many states, the key exceptions include statutory rape, the distribution of contaminated foods and pollution. Research into our punitive and moral attitudes, as well as when we are mindful of other people's minds, should shed light on the functional roles and perhaps even the evolutionary origins of different moral norms.

Acknowledgements

The authors thank Fiery Cushman, Paul Bloom, Josh Greene, Joe Henrich and the members of the *Culture and the Mind Project*, funded by the Arts and Humanities Research Council (Project Director: Stephen Laurence, University of Sheffield), for useful discussions, and two anonymous reviewers for helpful suggestions.

Funding

This project was funded by the Alfred P. Sloan Foundation, the Simons Foundation, and National Institutes of Health Grant 1R01 MH096914-01A1.

Supplementary data

Supplementary data are available at *SCAN* online.

Conflict of interest. None declared.

References

Appiah, K.A. (2006). *Cosmopolitanism: Ethics in a World of Strangers*. NY: W. W. Norton.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychological Review*, **3**, 193–209.

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E., Saxe, R. (2011). Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences*, **108**(11), 4429–34.

Borg, J.S., Hynes, C., Van Horn, J., Grafton, S., Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: an FMRI investigation. *Journal of Cognitive Neuroscience*, **18**, 803–17.

Borg, J.S., Lieberman, D., Kiehl, K.A. (2008). Infection, incest, and iniquity: investigating the neural correlates of disgust and morality. *Journal of Cognitive Neuroscience*, **20**(9), 1529–46.

Carter, R.M., Huettel, S.A. (2013). A nexus model of the temporal–parietal junction. *Trends in Cognitive Sciences*, **17**(7), 328–36.

Chakroff, A., Dungan, J., Young, L. (2013). Harming ourselves and defiling others: what determines a moral domain?. *PLoS One*, **8**(9), e74434.

Chakroff, A., Young, L. (2015). Harmful situations, impure people: An attribution asymmetry across moral domains. *Cognition*, **136**, 30–7.

Chapman, H.A., Kim, D.A., Susskind, J.M., Anderson, A.K. (2009). In bad taste: evidence for the oral origins of moral disgust. *Science*, **323**, 1222–6.

Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analysis in moral judgment. *Cognition*, **108**, 353–80.

Decety, J., Cacioppo, S. (2012). The speed of morality: a high-density electrical neuroimaging study. *Journal of Neurophysiology*, **108**(11), 3068–72.

Decety, J., Porges, E.C. (2011). Imagining being the agent of actions that carry different moral consequences: an fMRI study. *Neuropsychologia*, **49**, 2994–3001.

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

Fedorenko, E., Nieto-Castanon, A., Kanwisher, N. (2012). Lexical and syntactic representations in the brain: an fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia*, **50**(4), 499–513.

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

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

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

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

Gray, K., Wegner, D.M. (2009). Moral typecasting: divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, **96**, 505–20.

Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., Cohen, J.D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, **44**, 389–400.

Gweon, H., Young, L., Saxe, R.R. (2011). Theory of Mind for you, and for me: behavioral and neural similarities and differences in thinking about beliefs of the self and other. *In:* Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, pp. 2492–7).

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, **108**, 814–34.

Hamlin, J.K. (2013). Failed attempts to help and harm: intention versus outcome in preverbal infants' social evaluations. *Cognition*, **128**(3), 451–74.

Harris, L.T., Fiske, S.T. (2006). Dehumanizing the lowest of the low: neuroimaging responses to extreme out-groups. *Psychological Science*, **17**, 847–53.

Hart, H.L.A. (1968). *Punishment and Responsibility*. Oxford: Oxford University Press.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, **293**, 2425–30.

Haynes, J.D., Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, **7**(7), 523–34.

Heatherton, T.F., Wyland, C.L., Macrae, C.N., Demos, K.E., Denny, B.T., Kelley, W. M. (2006). Medial prefrontal activity differentiates self from close others. *Social Cognitive and Affective Neuroscience*, **1**(1), 18–25.

Hsieh, P.J., Colas, J.T., Kanwisher, N. (2012). Spatial pattern of BOLD fMRI activation reveals cross-modal information in auditory cortex. *Journal of Neurophysiology*, **107**(12), 3428–32.

Inbar, Y., Pizarro, D.A., Knobe, J., Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion*, **9**, 435–9.

Janoff-Bulman, R., Carnes, N.C. (2013). Surveying the moral landscape moral motives and group-based moralities. *Personality and Social Psychology Review*, **17**:219–36.

Jenkins, A.C., Mitchell, J.P. (2009). Mentalizing under uncertainty: dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cerebral Cortex*, **20**, 404–10.

Jost, J., Nosek, B.A., Gosling, S. (2008). Ideology: its resurgence in social, personality, and political psychology. *Perspectives on Psychological Science*, **3**, 126–36.

Kedia, G., Berthoz, S., Wessa, M., Hilton, D., Martinot, J.L. (2008). An agent harms a victim: a functional magnetic resonance imaging study on specific moral emotions. *Journal of Cognitive Neuroscience*, **20**, 1788–98.

Killen, M., Lynn Mulvey, K., Richardson, C., Jampol, N., Woodward, A. (2011). The accidental transgressor: morally-relevant theory of mind. *Cognition*, **119**, 197–215.

Kliemann, D., Young, L., Scholz, J., Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia*, **46**, 2949–57.

Knobe, J. (2005). Theory of mind and moral cognition: exploring the connections. *Trends in Cognitive Sciences*, **9**, 357–9.

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

Malle, B.F., Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, **33**, 101–21.

Mikhail, J.M. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences*, **11**, 143–52.

Mitchell, J. P., Macrae, C. N., Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, **50**(4), 655–63.

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

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, **10**(9), 424–30.

Parkinson, C., Sinnott-Armstrong, W., Koralus, P., Mendelovici, A., McGeer, V., Wheatley, T. (2011). Is morality unified? Evidence that distinct neural systems underlie moral judgments of harm, dishonesty, and disgust. *Journal of Cognitive Neuroscience*, **23**(10), 3162–80.

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**, 245–58.

Piaget, J. (1965/1932). *The Moral Judgment of the Child*. NY: Free Press.

Piazza, J., Russell, P. S., Sousa, P. (2013). Moral emotions and the envisioning of mitigating circumstances for wrongdoing. *Cognition & Emotion*, **27**, 707–22.

Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data?. *Trends in Cognitive Sciences*, **10**, 59–63.

Rai, T., Fiske, A. (2011). Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, **118**, 57–75.

Raizada, R.D., Tsao, F.M., Liu, H.M., Holloway, I.D., Ansari, D., Kuhl, P.K. (2010). Linking brain-wide multivoxel activation patterns to behaviour: examples from language and math. *NeuroImage*, **51**(1), 462–71.

Rottman, J., Kelemen, D., Young, L. (2014). Tainting the soul: purity concerns predict moral judgments of suicide. *Cognition*, **130**(2), 217–26.

Russell, P.S., Giner-Sorolla, R. (2011a). Moral anger, but not moral disgust, responds to intentionality. *Emotion*, **11**(2), 233.

Russell, P.S., Giner-Sorolla, R. (2011b). Moral anger is more flexible than moral disgust. *Social Psychological and Personality Science,* **2**, 360–4.

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

Saxe, R., Powell, L.J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological Science,* **17**, 692–9.

Sheikh, S., Janoff-Bulman, R. (2010). The "shoulds" and "should nots" of moral emotions: a self-regulatory perspective on shame and guilt. *Personality and Social Psychological Bulletin,* **36**, 213–24.

Uhlmann, E.L., Zhu, L. (2014). Acts, persons, and intuitions: Person-centered cues and gut reactions to harmless transgressions. *Social Psychological and Personality Science,* **5**, 279–85.

Waytz, A., Epley, N., Cacioppo, J.T. (2010a). Social cognition unbound: insights into anthropomorphism and dehumanization. *Current Directions in Psychological Science,* **19**, 58–62.

Waytz, A., Gray, K., Epley, N., Wegner, D.M. (2010b). Causes and consequences of mind perception. *Trends in Cognitive Sciences,* **14**, 383–8.

Waytz, A., Zaki, J., Mitchell, J.P. (2012). Response of dorsomedial prefrontal cortex predicts altruistic behavior. *The Journal of Neuroscience,* **32**(22), 7646–50.

Wheatley, T., Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, **16**(10), 780–4.

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

Young, L., Dodell-Feder, D., Saxe, R. (2010b). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia* **48**, 2658–64.

Young, L., Nichols, S., Saxe, R. (2010c). Investigating the neural and cognitive basis of moral luck: it's not what you do but what you know. *Review of Philosophy and Psychology,* **1**, 333–49.

Young, L., Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage,* **40**, 1912–20.

Young, L., Saxe, R. (2009a). An FMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience,* **21**, 1396–405.

Young, L., Saxe, R. (2009b). Innocent intentions: a correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia,* **47**, 2065–72.

Young, L., Saxe, R. (2011). When ignorance is no excuse: different roles for intent across moral domains. *Cognition,* **120**, 202–14.

Young, L., Scholz, J., Saxe, R. (2011). Neural evidence for "intuitive prosecution": the use of mental state information for negative moral verdicts. *Social Neuroscience,* **6**(3), 302–15.

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