




A functional neuroimaging investigation of Moral Foundations Theory

Ari Khoudary, Eleanor Hanna, Kevin O'Neill, Vijeth Iyengar, Scott Clifford, Roberto Cabeza, Felipe De Brigard & Walter Sinnott-Armstrong


To cite this article: Ari Khoudary, Eleanor Hanna, Kevin O'Neill, Vijeth Iyengar, Scott Clifford, Roberto Cabeza, Felipe De Brigard & Walter Sinnott-Armstrong (2022): A functional neuroimaging investigation of Moral Foundations Theory, Social Neuroscience, DOI: [10.1080/17470919.2022.2148737](https://doi.org/10.1080/17470919.2022.2148737)

To link to this article: <https://doi.org/10.1080/17470919.2022.2148737>

 View supplementary material [↗](#)

 Published online: 30 Nov 2022.

 Submit your article to this journal [↗](#)

 Article views: 287

 View related articles [↗](#)

 View Crossmark data [↗](#)



A functional neuroimaging investigation of Moral Foundations Theory

Ari Khoudary ^{a,b}, Eleanor Hanna^{b,c}, Kevin O'Neill ^{b,c}, Vijeth Iyengar^{b,c}, Scott Clifford ^d, Roberto Cabeza^{b*}, Felipe De Brigard ^{a,b,c*} and Walter Sinnott-Armstrong ^{a,b,c,e*}

^aDepartment of Philosophy, Duke University, Durham, North Carolina, USA; ^bCenter for Cognitive Neuroscience, Duke University, Durham, NC, USA; ^cDepartment of Psychology and Neuroscience, Duke University, Durham, NC, USA; ^dDepartment of Political Science, University of Houston, Houston, TX, USA; ^eKenan Institute for Ethics, Duke University, Durham, NC, USA

ABSTRACT

Moral Foundations Theory (MFT) posits that the human mind contains modules (or “foundations”) that are functionally specialized to moralize unique dimensions of the social world: Authority, Loyalty, Purity, Harm, Fairness, and Liberty. Despite this strong claim about cognitive architecture, it is unclear whether neural activity during moral reasoning exhibits this modular structure. Here, we use spatiotemporal partial least squares correlation (PLSC) analyses of fMRI data collected during judgments of foundation-specific violations to investigate whether MFT’s cognitive modularity claim extends to the neural level. A mean-centered PLSC analysis returned two latent variables that differentiated between social norm and moral foundation violations, functionally segregated Purity, Loyalty, Physical Harm, and Fairness from the other foundations, and suggested that Authority has a different neural basis than other binding foundations. Non-rotated PLSC analyses confirmed that neural activity distinguished social norm from moral foundation violations, and distinguished individualizing and binding moral foundations if Authority is dropped from the binding foundations. Purity violations were persistently associated with amygdala activity, whereas moral foundation violations more broadly tended to engage the default network. Our results constitute partial evidence for neural modularity and motivate further research on the novel groupings identified by the PLSC analyses.

ARTICLE HISTORY

Received 6 April 2022
Revised 4 November 2022
Published online 30
November 2022


KEYWORDS

Moral Foundations Theory; partial least squares; functional magnetic resonance imaging; moral cognition; modularity


Introduction

A long-standing goal of psychological, neuroscientific, and philosophical research has been to understand the principles governing human morality. One of the most influential contemporary theories of morality is Moral Foundations Theory (hereafter MFT; Haidt and Joseph 2011; Haidt & Joseph, 2007). Drawing on insights from cultural anthropology (e.g., Brown, 1991), evolutionary biology (e.g., De Waal, 1996), and cross-cultural psychology (e.g., Shweder, 1990), MFT argues that the human mind is predisposed to learn norms about, and moralize, particular dimensions of the social environment. These modules or “foundations” are claimed to be innate to all humans, and thus to constitute building blocks of morality (Graham et al., 2013). Cross-cultural and individual differences in moral judgments reflect different moral priorities, but MFT posits that all humans share the same set of foundations: Authority/Subversion, Loyalty/Betrayal, Sanctity (or Purity)/Degradation, Fairness/Cheating, and Care/Harm (Haidt and Joseph, 2011; Haidt & Joseph, 2007).

MFT has been a guiding framework for an abundance of research into human behavior, spanning domains from moral psychology to media studies (e.g., Tamborini et al., 2012) to agricultural ethics (e.g., Mäkinen et al., 2013). One particularly fruitful family of research applies MFT to questions of political ideology (Graham et al., 2009; Graham et al. 2012; Iyer et al., 2012; McAdams et al., 2008). Specifically, Graham et al. (2009) found that the foundations can be grouped into two superordinate categories: one that emphasizes the rights of individuals (*Individualizing* foundations: Harm and Fairness) vs. another that emphasizes values of group unity (*Binding* foundations: Authority, Loyalty, Purity). These superordinate categories reflect the ideological dichotomy between contemporary liberals and conservatives: liberals tend to moralize primarily violations of individualizing foundations, whereas conservatives tend to moralize all of the foundations approximately equally. The unique ideological perspective of libertarians has motivated the inclusion of a sixth category, Liberty/Oppression (Iyer et al. (2012).

CONTACT Felipe De Brigard  felipe.debrigard@duke.edu  Center for Cognitive Neuroscience, Duke University 308 Research Drive, Room C03E Durham, NC 27708-0999

*Indicates co-senior authorship.

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/17470919.2022.2148737>

© 2022 Informa UK Limited, trading as Taylor & Francis Group

Each moral foundation is supposed to correspond to a unique module that can be detected with a variety of measures (Graham et al., 2013). Although several methods have been developed to probe these modules (e.g., Clifford et al., 2015; Ferguson, 2007; Graham & Haidt, 2012; Graham et al., 2009, 2011; Payne et al., 2005; Van Berkum et al., 2009; Young & Saxe, 2011), the vast majority are purely behavioral in scope. This limitation in scope may be due to the fact that the moral foundations are postulated as cognitive modules rather than neural ones; i.e., MFT is a theory about cognitive architecture, not neural organization. The authors are explicit about this, asserting that foundations are not “spots in the brain” nor are they reducible to “one specific physiological signature” (Graham et al., 2013, p. 96). Regardless, there are a number of neuroimaging studies that use MFT to guide their inquiry into the neural basis of moral cognition. For instance, Parkinson et al. (2011) found that Physical Harm and Disgust (closely related to Purity) activated neural systems associated with action and emotion, respectively. Wasserman et al. (2017), however, found that Harm and Purity violations both activate a mentalizing network but converge onto different regions of the network (precuneus and inferior frontal gyrus, respectively). Tsoi et al. (2018) also found that Harm violations engage the precuneus, and that multi-voxel patterns in the precuneus and right temporoparietal junction (rTPJ) differentiate between Physical and Emotional Harm violations.

Taken together, these results provide convergent evidence from psychology and neuroscience for the base claim of *pluralism*: moral cognition is a multidimensional capacity composed of several systems that are distinct at both the cognitive and neural levels (Sinnott-Armstrong, 2016). We see this as a successful instance of the “method-theory coevolution” (Graham et al., 2013, p. 72) espoused by the developers of MFT. Although the foundations are not postulated at the neural level, methodological developments in functional neuroimaging revealed meaningful differences among foundations at this level. This, in turn, provided a novel set of empirical results offering support for the foundational premise of MFT: moral pluralism.

This paper aims to contribute in a similar manner, but with a different theoretical target: modularity. Specifically, our goal is to use multivariate analyses of functional neuroimaging data to investigate whether we can observe neural signatures of modularity that correspond to the cognitive modularity posited by MFT. Because we are testing the theory at a different level from which it was developed (and doing so within the null hypothesis testing framework), it is important to note that failing to observe such signatures would not

serve as evidence against cognitive modularity of MFT. Rather, it would indicate that the modular architecture does not extend to the neural level. Finding neural signatures of modularity, however, would contribute positively to the method-theory coevolution of MFT by providing converging, cross-level evidence for this core tenet of the theory. Additionally, it could provide insight into how MFT’s cognitive modularity is implemented by the brain.

This, of course, raises the question of how to operationalize modularity at the neural level (henceforward “neural modularity”). Developers of MFT posit that the cognitive modularity of the foundations consists in each foundation being a “functionally specialized mechanism” that works with other mechanisms (i.e., foundations) to solve recurrent adaptive problems (Graham et al., 2013, p. 62). One way to operationalize this functional specialization is via dimensionality reduction into statistically orthogonal latent variables – i.e., investigating whether neural activity during moral judgments about foundation violations can be decomposed into components that are uncorrelated with each other. This lack of correlation indicates that each variable accounts for a unique portion of the variance in the neural data space, which can be interpreted as “functional specialization” in the context of task-based neuroimaging. It is important to note that making an inference about the precise neural function a latent variable identifies relies critically on the choice of task and stimuli, and this “function” is not guaranteed to be preserved across tasks or stimuli. Rather, it is more akin to a “proof-of-concept” that neural activity *can be* decomposed into these dimensions when individuals perform the task at hand.

Toward this goal, we use spatiotemporal partial least squares correlation (PLSC) analyses on functional magnetic resonance (fMRI) data acquired while participants judged the wrongness of moral foundation violations. Spatiotemporal PLSC (Krishnan et al., 2011; McIntosh et al., 2004) is a multivariate technique that uses singular value decomposition to identify latent variables (LVs) that maximize covariance between the task design and neural activity matrices while remaining statistically orthogonal; i.e., each latent variable accounts for a unique dimension of the brain-task covariance. We chose PLSC instead of other multivariate techniques for four reasons. First, unlike representational similarity analyses, it guarantees orthogonality of the returned variables, which we believed was necessary for adequately probing neural modularity. Second, PLSC has the additional benefit of being a whole-brain approach, which further aligns it with MFT’s commitment to non-localist operationalizations of the foundations. Instead of

performing a contrast for each voxel (as in univariate analyses) or searching within an ROI (as is common for RSA), spatiotemporal PLSC performs the decomposition over all voxels in a single analytic step. It does identify voxels that maximally differentiate the conditions from each other, but the analysis is not conceptually restricted to inferences at the cluster level. Third, while independent component analyses meet all of the aforementioned criteria, PLSC has the additional advantage of limiting the number of returned variables to the number of task conditions, which helpfully constrains the inference space. The final advantage of PLSC comes from its capacity for both data- and hypothesis-driven analyses.

Mean-centered PLSC performs the data-driven decomposition described above, whereas non-rotated PLSC returns latent variables corresponding to experimenter-defined contrasts among task conditions (along with an estimate of their statistical significance). Given the exploratory nature of this study, we made use of both types of PLSC analyses. Our rationale was that demonstrating neural modularity via a mean-centered PLSC analysis would be strong evidence in support of MFT, indicating that the foundations can be recovered from neural data using information only about task timing. However, if the evidence from this analysis was ambiguous with respect to individual foundations, we could complement it using hypothesis-driven contrasts (Krishnan et al., 2011). We consider a foundation to exhibit neural modularity if its confidence intervals do not cross the grand mean or overlap with those of other foundations (more details provided in the Methods).

We anticipated three possible outcomes to the mean-centered PLSC analysis, listed here in decreasing order of support for neural modularity corresponding to MFT's cognitive modules: (i) it identifies one LV per foundation, (ii) it identifies one LV for each superordinate category (i.e., individualizing and binding), and (iii) it identifies LVs that differentiate some foundations from others in a way that is not predicted by the theory. Outcome (i) would provide strong evidence of MFT-like neural modularity, as it would mean that MFT's foundation-level structure can be recovered in a theory-free, data-driven manner from the brain-task covariance. Outcome (ii) would be weaker evidence for MFT-like neural modularity but strong evidence in favor of the superordinate structure of MFT because, again, it would indicate that the superordinate category structure can be recovered from the data alone. Finally, outcome (iii) would be strong evidence for the base claim of moral pluralism but only weak evidence of MFT-like neural modularity, as it would indicate some degree of functional specialization for a subset of the foundations but not a subset that is predicted by the theory. Additionally, we expected to

find (iv) just one LV that differentiates the control category (social norm violations) from all of the moral foundations. This would provide no evidence for pluralism or MFT-like neural modularity per se, but it could validate the pre-theoretical claim that neural activity does differentiate between moral and conventional/social violations in our sample.

To maximize our ability to detect neural modularity, we used the Moral Foundations Vignettes (Clifford et al., 2015), a stimulus set designed and validated to tap into each foundation individually. In addition to describing scenarios that violate the moral foundations, this stimulus set also contains vignettes describing violations of social norms to serve as a control. We believe this is an appropriate control category because it allows us to distinguish between neural activity related to generally social vs. specifically moral cognition. Thus, we included social norms as a control condition in our task to confirm that our analyses were sensitive to neural and behavioral differences between these two related forms of cognition.

Materials and methods

Participants

Thirty right-handed, native English speakers who reported no history of psychiatric or neurological disorders participated in the experiment. Data from the first three participants was excluded from analysis because of an error in the experiment script, resulting in $N = 27$ (14 male, 13 female; $M_{\text{age}} = 24.65$ (4.21) years). All participants lived in or near Durham, North Carolina at the time the study. As measured by the Social and Economic Conservatism Scale (Everett, 2013), the final sample was well-balanced with respect to political ideology ($M_{\text{Social Conservatism}} = 55.69$ (25.57), $M_{\text{Economic Conservatism}} = 52.60$ (15.23), possible scores ranging from 0 to 100). All participants provided written, informed consent in accordance with the requirements of the Duke University Health System Institutional Review Board and were compensated for their time. Sample size was decided on the basis of previous neuroimaging studies on MFT (Chakroff et al., 2016; Parkinson et al., 2011; Tsoi et al., 2018).

Stimuli

Stimuli presented in the scanner were drawn from the normed and validated Moral Foundations Vignettes stimulus set (Clifford et al., 2015). Each vignette (14–17 words) consisted of a second-person description of a social or moral violation, beginning with the words “You see” in order to encourage vivid mental simulation

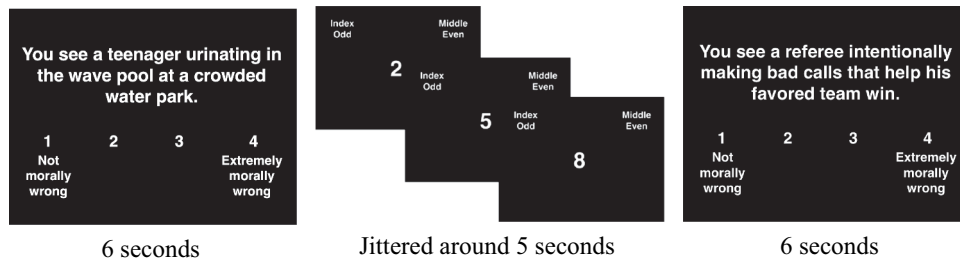


Figure 1. Experimental design. Inside the scanner, participants had 6 seconds to read and make judgments of the Moral Foundations Vignettes (15 per foundation and 15 depicting violations of social norms to serve as a control). Responses were made on an MR-compatible 4-button box using their right hand. Trials were separated by an even-odd discrimination task with a jitter drawn from a Poisson distribution (mean=5s). After the scan, participants rated how well 6 different emotions (anger, fear, sadness, contempt, amusement, disgust) described their reactions to the vignettes.

of each scene. The vignettes in the stimulus set were normed to be clear and easy to understand, and to be classified distinctively under the foundation each was meant to represent. Vignettes described violations of each of the moral foundations (including Liberty as a foundation), as well as violations of amoral social norms to serve as a control. In keeping with the factor analytic conclusions of Clifford et al. (2015) (cf. also Tsoi et al., 2018), Harm was subdivided into Emotional and Physical Harm, for a total of eight conditions. Fifteen vignettes per condition were presented to each participant in the scanner. Table S11 contains the full text of all the vignettes used in this study.

Experimental design

Participants were instructed to vividly imagine witnessing the actions depicted by the vignettes and to make a judgment of the moral wrongness of the act (Figure 1). Ratings were made on a 4-point rating scale ranging from 1 (“not at all morally wrong”) to 4 (“extremely morally wrong”). Participants had six seconds to read the vignette and make a judgment via button press, and practiced doing so prior to entering the scanner; none of the vignettes used in the practice session were viewed in the scanner. Scanning was divided into three runs of approximately 8 minutes. Each run consisted of 40 trials, pseudorandomized such that five vignettes from each foundation were presented in each run. An even-odd task served as an active baseline task between trials. Participants had 1.5 seconds to indicate whether a digit was even or odd, and did this for the duration of the jittered intertrial interval (3, 4.5, 6, 7.5, or 9 seconds drawn from a Poisson distribution with a mean of 5).

Participants then completed a number of post-scan tests. First, their recognition memory was tested by presenting them with all of the vignettes seen in the

scanner, plus 7–8 lure vignettes per condition. The vignettes had 1–3 words replaced by a blank, and participants had 3 semantically similar options or a “New” response to indicate their memory judgment. After each memory response, participants rated their confidence (1 “not confident” – 4 “extremely confident”) in that judgment. In the next task, participants viewed each vignette from the scanner once again, were provided with a list of the foundations, and were asked to choose up to 2 foundations they thought the vignette belonged to. Next, participants rated how much six different emotions (anger, amusement, sadness, fear, contempt, and disgust) described their experience of the vignette on a 7-point rating scale ranging from 1 (“not at all”) to 7 (“perfectly”). Then, participants completed the Moral Foundations Questionnaire (Graham, Haidt & Nosek, 2008), the Disgust Scale-Revised (Haidt et al., 1994; Olatunji et al., 2007), the Interpersonal Reactivity Index (Davis, 1980, 1983), and the Social and Economic Conservatism Scale (Everett, 2013). Following this post-scan session, participants were debriefed and compensated for their time. Apart from the emotional reaction task, none of the post-scan data were used for subsequent analyses, as they were collected to be used for a separate paper. Summary statistics for all post-scan surveys can be found in the Supplementary Materials (Table S1B).

Scanning parameters

Scanning was conducted on a research-dedicated 3T GE MR750 scanner with an 8-channel head coil. The scanning session began with a high-resolution T1-weighted structural scan followed by a five-minute resting state scan and three runs of the moral judgment task. Functional scans were collected using a whole-brain spiral-in sequence (TR = 2s, TE = 30 ms, flip angle = 70°). Slices were acquired in an interleaved fashion, and

participants' heads were kept in place with cushions to limit head motion. The task was projected into the scanner and viewed by participants with a mirror placed above the head coil. Stimuli were presented in white letters on a black background, using Psychtoolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997), which was also used to collect behavioral responses. Each run began and ended with 10 seconds (5 TRs) of fixation that were dropped from analysis. The scanning session concluded with five minutes of a localizer task for emotional faces. Data from the resting state and localizer task are not reported here.

fMRI data preprocessing

Functional neural data were preprocessed using FSL 5.0.1 (Jenkinson et al., 2012). Images were reoriented, slice-time corrected using an interleaved pattern, motion-corrected using MCFLIRT, realigned and unwrapped, normalized to the Montreal Neurological Institute template (resampled at $2 \times 2 \times 2$ mm voxels), temporally filtered using a highpass filter, and spatially smoothed using a 6 mm full-width half-maximum Gaussian kernel. An independent components analysis for each run was conducted using FSL MELODIC in order to identify noise components to be filtered out of the functional data prior to the analysis. Denoising was conducted according to guidelines set out by Kelly et al. (2010). Additionally, in-house scripts implemented in MATLAB flagged components related to fat suppression artifacts by identifying maximum frequency peaks at 0.14 Hz or 0.22 Hz that were at least 5 standard deviations above mean frequency. The FSL regfilt function was then used to filter out noise components from the preprocessed data.

Statistical analyses

The preprocessed neural data were analyzed using a spatiotemporal partial least squares (PLSC) toolbox in MATLAB developed at the Rotman Research Institute (<https://www.baycrest.org/>). Trial-related activity was binned across runs of the task, and was defined as all activity occurring within 7 TRs (14s) of vignette onset. Since the PLSC toolbox does not fit the data to an estimate of the hemodynamic response function (HRF), we investigated temporal windows spanning from 3–7 TRs. We found that the pattern of brain scores remained stable across temporal window definitions only after the window was extended to 6 TRs, consistent with the canonical peak of the HRF occurring roughly 6 seconds after stimulus onset. We report results from the 7 TR analysis to remain consistent with the common practice

of extending the temporal window by one additional TR when using this toolbox (De Brigard et al., 2015; Faul et al., 2020; Hassabis et al., 2014; McIntosh et al., 2004).

We performed two types of PLSC analyses – mean-centered and non-rotated – to conduct theory-free and theory-guided analyses, respectively. Both analyses return orthogonal latent variables that maximize the covariance between whole-brain activity and task timing (in mean-centered analyses) or experimenter-defined contrasts (in non-rotated analyses). The statistical significance of an LV is determined using permutation testing. Because PLSC compares all voxels in a single step, there is no need to adjust for multiple comparisons (i.e., $\alpha = 0.05$). Each LV is composed of “brain score” values for each task condition. The brain score of a condition is the sum, across all voxels in all participants, of the product of the BOLD signal and the singular value weight for that voxel; in this way, brain scores can be interpreted analogously to component scores in principal component analyses. Within an LV, confidence intervals (CIs) for the brain scores are used to determine whether a condition contributed significantly to the variance accounted for by LV. Brain score CIs are computed via bootstrap resampling performed in an independent computational step from permutation testing for LV significance. We interpret an experimental condition as significantly contributing to an LV if its confidence intervals do not cross the grand mean for that LV. Additionally, we consider conditions to be statistically different from each other (i.e., are “functionally specialized”) if their CIs do not overlap (and are also different from the grand mean).

We report results from one mean-centered PLSC analysis and three non-rotated analyses that investigated the following contrasts: Social Norms vs. Moral Foundations, Binding vs. Individualizing foundations, and Binding (minus Authority) vs. Individualizing foundations. All PLSC analyses were performed with 100 rounds of bootstrapping and 500 permutations. In line with previous work (e.g., De Brigard et al., 2015), clusters larger than 10 voxels (40 mm^3) with a bootstrap ratio (BSR; ratio of brain score to standard error for each condition) greater than 3.2 ($p < .002$) for each significant LV are reported. Again, in line with previous work (De Brigard et al., 2015; Faul et al., 2020; McIntosh et al., 2004), we report clusters from the TR where maximal differentiation among conditions was observed. This occurred at TR6 (12s) for all PLSC analyses.

Behavioral data were analyzed with linear mixed effects models using the *lme4* package (Bates et al., 2015; version 1.1–23) in R version 3.6.3 (R Core Team, 2020). Effect sizes (Cohen's *d*) were estimated using the *eff_size()* function in the *emmeans* package

(Lenth, 2020; version 1.4.5). Confidence intervals were computed by the `tab_model()` function in the R package `sjPlot` (Lüdtke, 2020; version 2.8.3). We fit two models: one testing differences in moral judgments across foundations for our sample of participants and another testing how those judgments compared to the normed values of the stimulus set. The first model had random intercepts for participants and vignettes, and the second had

random intercepts for vignettes only (since participant information was not available for the normed dataset). Both models were fit with restricted maximum likelihood estimation (REML). p values and 95% confidence intervals were estimated using the Satterthwaite method. All significance testing was performed with $\alpha = 0.05$, and degrees of freedom were computed using the Kenward-Roger method.

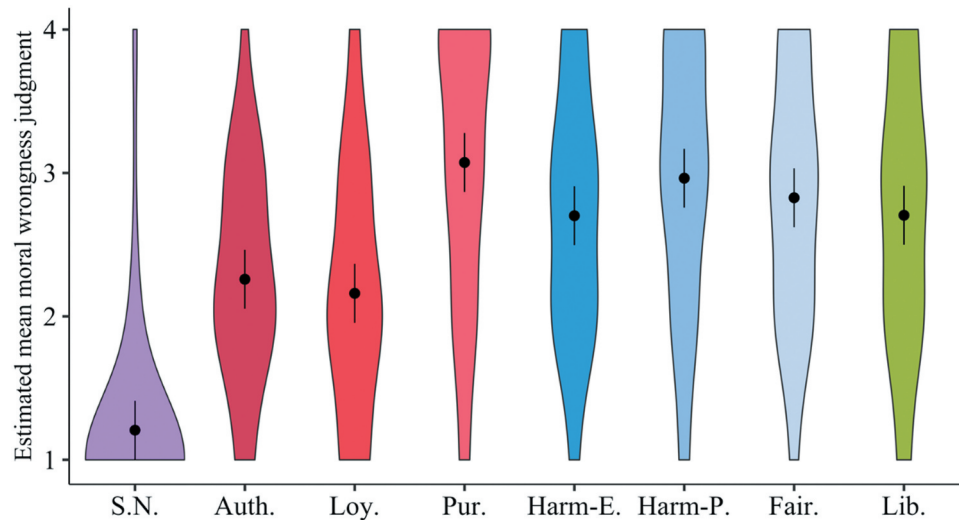


Figure 2. Average moral wrongness ratings per foundation. Responses were made on a 4-point scale (1=not morally wrong, 4=extremely morally wrong). Colors reflect superordinate categories: binding foundations are shown in shades of red, individualizing foundations in shades of blue. Violin width reflects response densities. Error bars represent 95% confidence intervals shown in Table 1. S.N.=Social Norms, Auth.=Authority, Loy.=Loyalty, Pur.=Purity, Harm-E.=Emotional Harm, Harm-P.=Physical Harm, Fair.=Fairness, Lib.=Liberty.

Table 1. Summary of moral judgments model. Linear mixed effects model fit with restricted maximum likelihood (REML). Maximum number of iterations for fitting = 5000. REML criterion at convergence = 7352.3.

Model Formula	moral judgment foundation + (1 Participant) + (1 Vignette)					
Factors	Estimates	CI (95%)	df	Cohen's d	t value	p value
Intercept (Social Norms)	1.21	1.00–1.41	123.09		11.636	<0.001
Authority	1.05	0.81–1.29	111.64	1.03	8.517	<0.001
Loyalty	1.00	0.71–1.20	112.06	0.93	7.714	<0.001
Purity	1.87	1.62–2.11	112.19	1.83	15.089	<0.001
Harm-Emo	1.50	1.25–1.74	111.69	1.46	12.101	<0.001
Harm-Phys	1.76	1.51–2.00	111.53	1.72	14.223	<0.001
Fairness	1.62	1.38–1.86	111.74	1.59	13.114	<0.001
Liberty	1.50	1.26–1.74	111.86	1.47	12.122	<0.001
Random Effects						
Residual variance	0.55					
Vignette variance (Intercept)	0.09					
Participant variance (Intercept)	0.08					
N _{Participant}	27					
N _{Vignette}	120					
Observations	3154					
Model Performance						
Marginal R ²	0.31					
Conditional R ²	0.48					
ICC	0.25					
RMSE	0.73					

Results

Behavioral results: Moral judgments

As illustrated by Figure 2, vignettes depicting violations of the foundations were judged as more morally wrong than vignettes depicting violations of Social Norms (all β values between 0.95–1.87, d values between 0.93–1.83, t values between 7.71–15.09, p values <.001; Figure 2 ; Table 1). To confirm that our sample did not exhibit abnormal response patterns, we ran a sum-coded linear mixed model comparing our sample's moral judgments to the normed data reported by Clifford et al. (2015). Because the normed data were collected on a 5-point scale and ours were collected with a 4-point scale, we z-scored the moral judgments prior to analysis. However, the additional variance introduced by a 5-point scale still resulted in artificially larger statistical differences for a number of interaction terms. The biggest difference was observed for the interaction of Social Norms*dataset, with our sample appearing to rate Social Norm violations as less wrong than the normed sample. However, our sample had a mean moral wrongness judgment of 1.2 (on a scale of 1–4) for Social Norms, indicating that participants appropriately judged these violations as

not morally wrong. In addition to this misleading interaction term, the inclusion of Social Norms in this model considerably decreased the intercept (grand mean for a sum-coded model), resulting in artificially significant interaction terms for other foundations as well. Fitting a model that excluded Social Norms eliminated all of the significant interaction terms, indicating that our sample did not exhibit abnormal responses to the Moral Foundations Vignettes. A summary of the full model is presented in Table 2, and a summary of the model excluding Social Norms is presented in Table S3 and Figure S2.

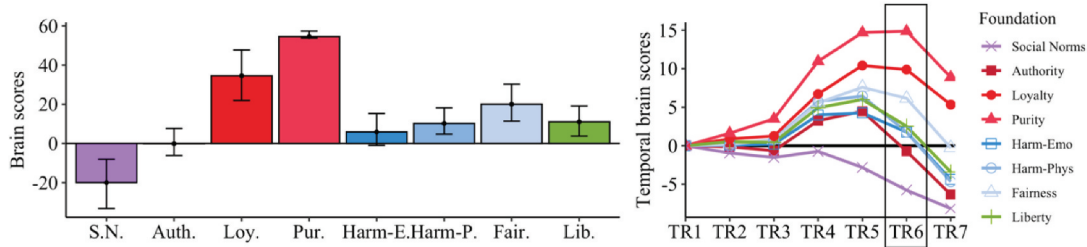
fMRI results: Mean-centered PLSC

As shown in Figure 3, the mean-centered PLSC analysis returned two significant LVs (hereafter MC-LVs) that together accounted for 47.50% of the crossblock variance. A third MC LV, accounting for 15.96% of the crossblock variance, was exactly at the significance threshold of $p < 0.050$. To remain statistically conservative, we report this MC-LV in Figure S6 and Table S7 but are not interpreting it along with the other mean-centered and non-rotated LVs. MC-LV1 (26.99% of crossblock variance, $p < .001$) differentiated Social Norms from Loyalty, Purity,

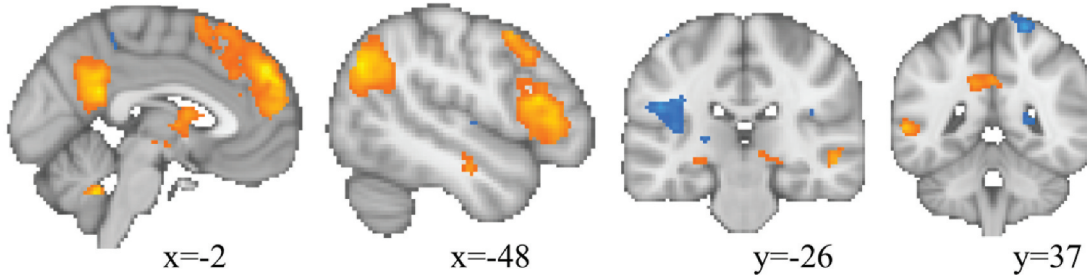
Table 2. Summary of norm data comparison model. Linear mixed effects model comparing the average moral judgment of each vignette used in this study to the normed data collected by Clifford et al. (2015). Moral judgment scores were z-scored to account for differences in the scales used between the two experiments. The model was fit with restricted maximum likelihood (REML); REML criterion at convergence = 7352.3.

Model formula	moral judgment (z-scored) dataset*foundation + (1 Vignette)				
Factors	Estimates	CI (95%)	df	t value	p value
Intercept (Grand mean)	0.01	−0.06–0.09	175.62	0.38	0.702
Our sample	−0.01	−0.09–0.06	112.00	−0.39	0.701
Social Norms	−2.23	−2.43 – −2.04	175.62	−22.62	<0.001
Authority	0.05	−0.14–0.25	175.62	0.55	0.584
Loyalty	−0.32	−0.52 – −0.13	175.62	−3.28	0.001
Purity	0.66	0.47–0.86	175.62	6.71	<0.001
Harm-Emo	0.37	0.18–0.56	175.62	3.74	<0.001
Harm-Phys	0.66	0.47–0.86	175.62	6.71	<0.001
Fairness	0.59	0.39–0.78	175.62	5.93	<0.001
Our sample*Social Norms	0.98	0.79–1.17	112.00	10.17	<0.001
Our sample*Authority	−0.28	−0.47 – −0.09	112.00	−2.93	0.004
Our sample*Loyalty	−0.00	−0.19–0.19	112.00	−0.00	0.999
Our sample*Purity	−0.08	−0.27–0.11	112.00	−0.81	0.418
Our sample*Harm-Emo	−0.16	−0.35–0.03	112.00	−1.65	0.102
Our sample*Harm-Phys	−0.20	−0.39 – −0.01	112.00	−2.05	0.043
Our sample*Fairness	−0.25	−0.44 – −0.06	112.00	−2.61	0.010
Random Effects					
Residual variance	0.08				
Vignette variance	0.09				
ICC	0.52				
N vignettes	120				
Observations	240				
Model Performance					
Marginal R ²	0.771				
Conditional R ²	0.891				
ICC	0.525				
RMSE	0.221				

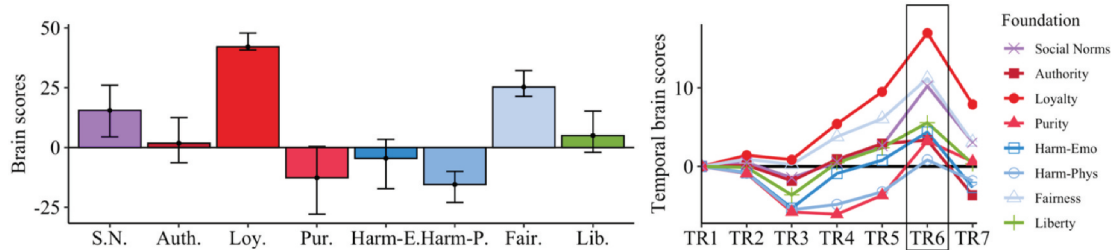
a. MC-LV1 ($p < 0.001$, 26.99% crossblock variance)



b. MC-LV1 singular value maps at TR6



c. MC-LV2 ($p < 0.010$, 20.51% crossblock variance)



d. MC-LV2 singular value maps at TR6

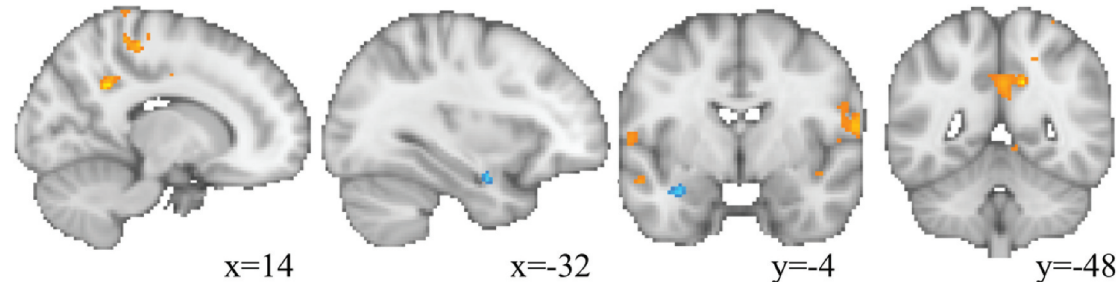


Figure 3. Mean-centered (MC) PLSC results. On the graphs, binding foundations are plotted in shades of red, individualizing foundations are plotted in shades of blue. The solid line at 0 represents the grand mean across all voxels in all conditions, and is the value against which the significance of a brain score is tested. On the brain images, positive singular values are plotted in warm colors and negative singular values are plotted in cool colors. Darker color=greater BSR. (A) Brain scores for MC-LV1, which dissociated between moral foundation and social norm violations. The black box overlaid on the temporal brain scores indicates the TR (TR 6; 12–14s) at which cluster reports and singular value maps were extracted. (B) Singular value map for MC-LV1. Activity associated with Loyalty, Purity, Physical Harm, Fairness, and Liberty is shown in orange (superior frontal gyrus, posterior cingulate, angular gyrus, inferior frontal gyrus); activity associated with Social Norms is shown in blue (insula, postcentral gyrus). (C) Brain scores for MC-LV2. (D) Singular value maps for MC-LV2. Activity associated with Social Norms, Loyalty, and Fairness is shown in orange (precuneus, precentral gyrus); activity associated with Emotional Harm is shown in blue (amygdala). S.N.=Social Norms, Auth.=Authority, Loy.=Loyalty, Pur.=Purity, Harm-E.=Emotional Harm, Harm-P.=Physical Harm, Fair.=Fairness, Lib.=Liberty.

Physical Harm, Fairness, and Liberty, though this seems largely driven by Loyalty and Purity (see the non-rotated analysis, below). Indeed, the CIs for Purity did not overlap with the CIs for any of the other foundations, indicating a unique contribution of Purity to MC-LV1. During the window of maximal differentiation of temporal brain scores (TR 6; Figure 3A), there were large clusters in the bilateral insulae, right postcentral gyrus, and right precuneus associated with positive singular values (Social Norms). Most clusters associated with negative singular values (the foundations) were left-lateralized, with the largest occurring in the superior frontal gyrus, posterior cingulate, angular gyrus, and inferior frontal gyrus (Figure 3C; see Table 3 for cluster report).

MC-LV2 (20.51% of crossblock variance, $p < .010$) differentiated Social Norms, Loyalty, and Fairness from Physical Harm. CIs for Loyalty were completely non-overlapping with Fairness and Social Norms, indicating that it contributed a unique dimension to the positive singular values of MC-LV2. Social Norms and Fairness did not exhibit such functional differentiation (i.e., their CIs were overlapping), and thus seem to be contributing

similar dimensions to the positive singular values. Additionally, the upper bound of Purity's CIs was minimally above 0 (exact value = 0.41), suggesting that it likely contributed to activity associated with negative singular values for MC-LV2. During the window of maximal differentiation (TR 6; Figure 3D), there were large, right-lateralized clusters in the precuneus and precentral gyrus associated with positive singular values (Social Norms, Loyalty, and Fairness). There was only one significant cluster associated with negative singular values (Physical Harm) in the left amygdala (Figure 3E; see Table 4 for cluster report). We then extracted the BOLD timecourse from this left amygdala cluster and found that it was greatest for Purity violations (Figure S6).

fMRI results: Non-rotated PLSC

The non-rotated PLSC analysis allowed us to specify contrasts of interest among experimental conditions. The first non-rotated analysis served as a manipulation check and test of prediction (iv): that neural activity

Table 3. Cluster report for MC-LV1 at TR6. Clusters are reported if they pass the BSR threshold of $\neq 3.2$, have a minimum of 10 voxels, and are separated by at least 10 mm.

Region(s)	BA	Hemi.	Peak MNI Coordinates			# of Voxels	BSR
			X	Y	Z		
Positive Singular Values (Loyalty, Purity, Physical Harm, Fairness, Liberty > Social Norms)							
Superior Frontal Gyrus	9	L	-4	50	36	6157	7.4896
Posterior Cingulate	23	L	-2	-50	24	1026	7.0375
Angular Gyrus	39	L	-48	-68	34	1225	6.9582
Middle Temporal Gyrus	21	L	-56	-18	-10	624	6.9327
	21	R	54	-12	-16	36	4.4532
Inferior Frontal Gyrus	47	L	-40	26	-6	1226	6.8166
Caudate	-	L	-8	2	10	576	5.3726
Inferior Parietal Lobule	40	R	52	-62	36	310	5.593
Amygdala	-	L	-18	-6	-12	116	5.2652
Middle Frontal Gyrus	9	R	48	22	36	41	3.8977
	6	R	36	12	48	65	3.6968
Thalamus	-	L	-10	-28	-2	45	3.7848
		L	-10	-8	6	11	3.7034
Superior Temporal Gyrus	38	L	-52	12	-10	21	3.766
Negative Singular Values (Social Norms > Loyalty, Purity, Physical Harm, Fairness, Liberty)							
Insula	13	R	40	0	12	147	-4.882
	13	R	48	-26	24	373	-4.8022
	13	L	-34	-14	22	30	-4.0129
	13	R	40	-16	-4	10	-3.7996
	13	L	-54	-32	20	16	-3.7974
Caudate	-	R	22	-42	12	50	-4.0664
Superior Temporal Gyrus	13	L	-42	-22	4	10	-3.8051
	42	R	64	-36	20	10	-3.6658
Precuneus	7	R	10	-84	46	87	-5.5069
Precentral Gyrus	6	R	62	-2	12	35	-3.9605
Postcentral Gyrus	5	R	22	-46	70	107	-5.1884
Cingulate Gyrus	24	R	12	-8	42	14	-4.8546
Paracentral Lobule	3	L	-18	-40	60	13	-4.284
	5	R	14	-32	48	81	-4.038

Table 4. Cluster report for MC-LV2 at TR6. Clusters are reported if they pass the BSR threshold of ≥ 3.2 , have a minimum of 10 voxels, and are separated by at least 10 mm.

Region(s)	BA	Hemi.	Peak MNI Coordinates			# of Voxels	BSR
			X	Y	Z		
Positive Singular Values (Social Norms, Loyalty, Fairness > Physical Harm)							
Precentral Gyrus	6	R	62	0	14	123	6.2705
	4	R	36	-22	42	69	3.9409
	4	R	24	-24	64	31	3.6995
Precuneus	6	R	44	-14	34	99	4.5711
	31	R	14	-48	34	466	5.7874
	19	R	22	-86	42	27	4.4006
Medial Frontal Gyrus	7	R	20	-50	48	19	3.8979
	6	R	14	-30	58	36	4.8601
Culmen	—	L	0	-62	-8	77	4.6977
Clastrum	—	R	38	-10	6	16	4.496
Postcentral Gyrus	5	R	26	-46	66	47	4.0553
Middle Temporal Gyrus	21	L	-54	-4	-16	17	3.897
	39	L	-42	-64	30	54	3.6647
Insula	13	R	34	-34	20	53	3.8357
	13	R	34	-8	16	13	3.6466
Superior Temporal Gyrus	22	L	-60	-6	6	49	3.8108
Lingual Gyrus	19	L	-22	-72	-6	22	3.7902
Thalamus	—	L	0	-18	6	12	3.6116
Negative Singular Values (Physical Harm > Social Norms, Loyalty, Fairness)							
Amygdala	—	L	-32	-4	-22	25	-5.564

differentiates between moral and conventional/social norm violations. As shown in [Figure 4A](#), this contrast (Social Norms vs. Foundations; hereafter NR-LV1) was statistically significant ($p < 0.010$). Similar to MC-LV1, at the window of maximal differentiation of brain scores (TR 6), clusters related to Social Norms were identified in the precuneus bilaterally, inferior parietal lobule, and bilateral insulae. And similar to MC-LV2, activity related to the moral foundations was found in the left middle temporal gyrus, bilateral angular gyrus, and left amygdala (see [Table 5](#) and [Figure 4B](#)).

The second non-rotated analysis contrasted the foundations along traditional superordinate lines (Emotional Harm, Physical Harm, Fairness vs. Authority, Loyalty, Purity) and failed to reach significance ($p < 0.138$), suggesting the brain does not differentiate between canonical superordinate groupings of the moral foundations. Then, we performed a third contrast to investigate a novel superordinate grouping inspired by the pattern of MC-LV1: individualizing foundations (Emotional Harm, Physical Harm, and Fairness) vs. Loyalty and Purity (without Authority). As shown in [Figure 4C](#), this contrast (hereafter NR-LV2) was statistically significant ($p < 0.019$), suggesting that Authority does not resemble the other binding foundations at the neural level. That is, removing Authority from the binding foundations returned significant neural differences along superordinate lines whereas including it did not, suggesting that Authority violations invoke different neural

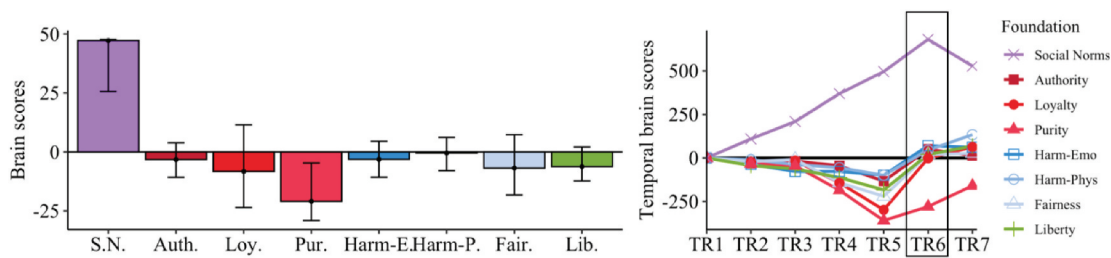
processes than the other binding foundations (Loyalty and Purity). Clusters associated with Loyalty and Purity were found in bilateral inferior frontal gyrus, left precuneus, left middle temporal gyrus, and right amygdala (see [Table 6](#) and [Figure 4D](#)). No clusters associated with the individualizing foundations were large enough to pass our reporting thresholds.

Discussion

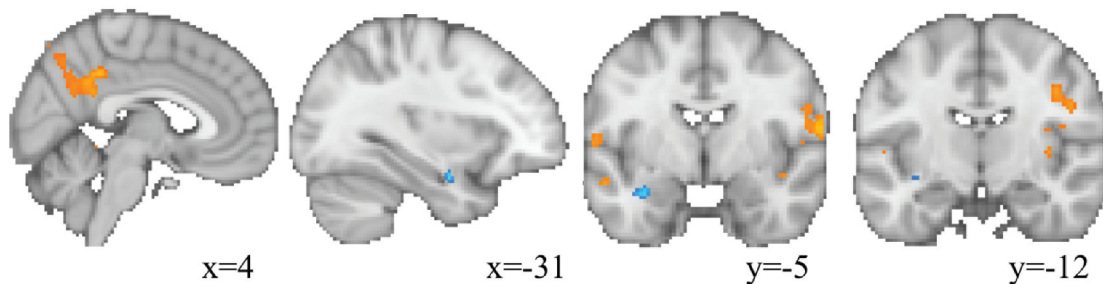
The present study asked whether the modular cognitive architecture posited by Moral Foundations Theory extends to a level measurable by functional neuroimaging. In line with MFT's claim that each moral foundation is a "functionally specialized mechanism" (Graham et al., 2013, p. 62), we used both mean-centered and non-rotated spatiotemporal PLSC analyses to investigate whether making judgments of moral foundation violations elicited statistically orthogonal patterns of neural activity. The mean-centered analysis allowed us to ask this question in a data-driven manner, whereas the non-rotated analyses allowed us to test theory-specific predictions as well as further investigate trends returned by the mean-centered analysis.

We anticipated three potential outcomes to the mean-centered analysis, each of which would lend decreasing degree of support for modular organization at the neural level corresponding to MFT's cognitive modules: (i) the identification of one latent

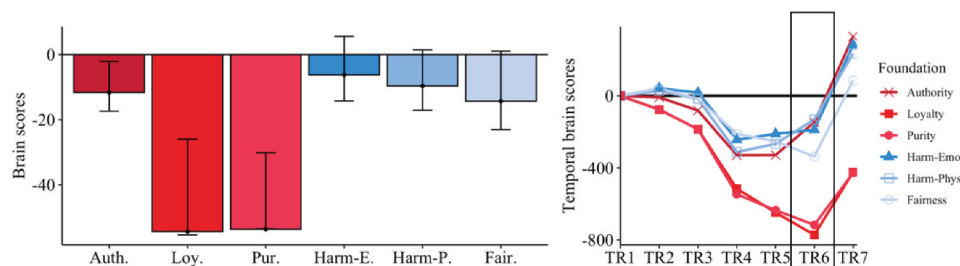
a. NR-LV1: Social Norms vs. Foundations ($p < 0.010$)



b. NR-LV1 singular value map at TR6 ■ Positive singular values ■ Negative singular values



c. NR-LV2: Loyalty & Purity vs. Individualizing ($p < 0.019$)



d. NR-LV2 singular value map at TR6 ■ Positive singular values ■ Negative singular values

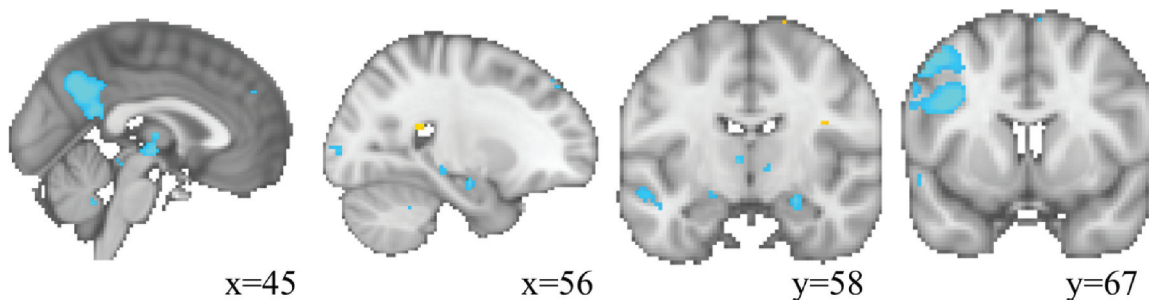


Figure 4. Non-rotated (NR) PLSC results. (A) Brain scores for the non-rotated analysis contrasting Social Norms against the moral foundations (NR-LV1). The black box overlaid on the temporal brain scores indicates the TR (TR 6; 12–14s) at which cluster reports and singular value maps were extracted. (B) Singular value maps for NR-LV1. Activity associated with Social Norms is shown in orange (precuneus, inferior parietal lobule, insula, middle temporal gyrus); activity associated with the foundations is shown in blue (amygdala). (C) Brain scores for the non-rotated analysis contrasting Loyalty and Purity against the individualizing foundations (Emotional Harm, Physical Harm, and Fairness; NR-LV2). (D) Singular value maps for NR-LV2. Activity associated with Loyalty and Purity is shown in blue (inferior frontal gyrus, precuneus, amygdala). S.N.=Social Norms, Auth.=Authority, Loy.=Loyalty, Pur.=Purity, Harm-E.=Emotional Harm, Harm-P.=Physical Harm, Fair.=Fairness, Lib.=Liberty.

Table 5. Cluster report for NR-LV1. Clusters are reported if they pass the BSR threshold of ∓ 3.2 , have a minimum of 10 voxels, and are separated by at least 10 mm.

Region(s)	BA	Hemi.	Peak MNI Coordinates			# of Voxels	BSR
			X	Y	Z		
Positive Singular Values (Social Norms > Foundations)							
Precuneus	19	L	-8	-86	44	267	5.4386
	7	R	10	-42	46	483	4.7452
	7	R	22	-60	42	149	4.4587
	19	R	10	-84	46	14	4.1413
	7	L	-24	-70	38	17	3.5226
Inferior Parietal Lobule	40	R	64	-36	38	994	5.3907
Insula	13	R	42	0	10	425	4.7077
	13	R	40	-16	-6	36	4.4624
	13	L	-38	0	12	18	3.6062
Superior Temporal Gyrus	42	L	-62	-32	20	59	4.6441
Fusiform Gyrus	37	L	-40	-44	-8	43	4.6058
Precentral Gyrus	6	R	38	-10	42	54	4.5265
	6	L	-56	-4	34	19	3.6156
Precentral Gyrus	6	R	14	-18	68	41	3.7503
Middle Occipital Gyrus	18	R	10	-94	18	13	4.49
Postcentral Gyrus	43	L	-60	-6	16	63	4.3455
	7	R	20	-44	72	24	4.1149
	43	R	62	-16	22	93	3.7501
	3	R	16	-38	64	22	3.783
Middle Temporal Gyrus	19	R	42	-64	14	18	4.0566
	21	R	58	-50	0	17	3.5378
Cuneus	19	R	18	-90	24	15	4.0196
Medial Frontal Gyrus	6	L	0	-18	70	29	4.0095
Cingulate Gyrus	24	R	12	-8	40	15	3.8711
Inferior Frontal Gyrus	46	R	50	38	8	16	3.7776
Putamen	—	R	28	-18	4	19	3.7499
Anterior Cingulate	24	R	6	36	2	15	3.7295
Clastrum	—	L	-30	-18	16	26	3.667
Negative Singular Values (Foundations > Social Norms)							
Middle Temporal Gyrus	21	L	-54	-16	-12	106	-5.4398
Angular Gyrus	39	L	-50	-68	34	474	-5.238
	39	R	56	-64	40	75	-4.2155
Precuneus	31	L	-4	-52	32	423	-5.1482
Superior Frontal Gyrus	6	L	-14	36	56	743	-4.7554
Amygdala	—	L	-32	-2	-22	17	-4.7353
Thalamus	—	L	-4	-2	6	21	-3.954

Table 6. Cluster report for NR-LV2. Clusters are reported if they pass the BSR threshold of ∓ 3.2 , have a minimum of 10 voxels, and are separated by at least 10 mm.

Region(s)	BA	Hemi.	Peak MNI Coordinates			# of Voxels	BSR
			X	Y	Z		
Negative Singular Values (Purity and Loyalty > Physical Harm, Emotional Harm, Fairness)							
Inferior Frontal Gyrus	47	L	-42	18	-2	11593	-7.0571
	47	R	42	22	-2	1676	-4.7661
Precuneus	7	L	-2	-60	36	5063	-5.564
Middle Temporal Gyrus	21	L	-58	-22	-8	8302	-5.3465
Amygdala	—	R	24	-10	-16	139	-4.1369
Middle Occipital Gyrus	18	R	28	-98	6	1115	-3.9065
Angular Gyrus	39	R	48	-64	32	1763	-3.8824
Medial Frontal Gyrus	—	L	-14	50	4	447	-3.3966
Culmen	—	R	24	-48	-32	41	-3.3856

variable per foundation, (ii) the identification of one latent variable per superordinate category, or (iii) the identification of any number of latent variables differentiating foundations along categories not predicted by MFT. Additionally, to corroborate that the behaviorally validated stimuli we employed in fact tapped

into moral as opposed to social violations, we expected our analysis to (iv) identify a latent variable differentiating social norms versus MFT's moral violations.

Our results are most in line with prediction (iii): we found some evidence for pluralism and neural

modularity, but the neural modules we found do not correspond to the cognitive modules postulated by MFT. Specifically, we found that only Purity, Loyalty, Physical Harm, and Fairness elicit statistically unique patterns of neural activity and that Authority does not resemble the other binding foundations at the neural level. We take these results to be paradigmatic of Graham et al.'s (Graham et al., 2013, p. 72) call for *method theory co-evolution*: “theoretical constructs inspiring the creation of new ways to measure them, and data from the measurements guiding development of the theory.” We applied a novel analytic technique to a comprehensive sampling of neural activity during judgments of foundation violations, and found that only a subset of the foundations exhibit neural modularity and that Authority violations do not engage the same neural processes as violations of the other two binding foundations (Purity and Loyalty). We interpret these results both as providing evidence in support of a conceptual claim of MFT – namely, a partial extension of the modularity claim to the neural level – and as motivating further developments of the theory, for instance, considering why this subset of foundations might exhibit stronger evidence of neural modularity and how Authority might differ from the other binding foundations.

Evidence for the functional specialization of Purity, Loyalty, Physical Harm, and Fairness comes from the mean-centered PLSC analyses. Our operationalization of “functional specialization” (or neural modularity) is whether a particular foundation’s BSR in the mean-centered analyses has confidence intervals (CIs) that overlap with another foundation and/or the grand mean. If the CIs do not overlap in these ways, then that foundation explains a unique dimension of the brain-task covariance (and thus is “functionally specialized” in this task). Purity met these criteria in MC-LV1, and Loyalty, Physical Harm, and Fairness met these criteria in MC-LV2. Interestingly, Fairness’s BSR CIs overlapped with those of Social Norms in MC-LV2, suggesting that similar processes might be underlying assessments of these two classes of violations (we consider what these might be in later paragraphs). Also interesting was the lack of a unique signature for Emotional Harm, given that Tsoi et al. (2018) found that multivoxel patterns in the precuneus differentiated between Physical and Emotional Harm violations. A likely explanation for this discrepancy is that, by choosing to include all of the foundation in our stimulus set, we may have been underpowered for detecting small differences between similar foundations. Indeed, it is possible that we failed to detect strong evidence of foundation-level modularity because our sample was only large enough to detect

differences at the superordinate level. However, because our goal was a comprehensive investigation into the potential neural modularity of MFT, we had to forego high-powered between-foundation comparisons to focus on big-picture questions of functional specialization.

Evidence for the lack of neural similarity between Authority and the other binding foundations (Purity and Loyalty) comes from both the mean-centered and non-rotated analyses. In addition to differentiating between Social Norms and most of the foundations, MC-LV1 also appeared to differentiate between the individualizing foundations and a subset of the binding foundations (Purity and Loyalty). We used a non-rotated analysis (NR-LV2) to probe the significance of this difference and found that dropping Authority from the binding foundations did in fact result in a significant binding vs. individualizing contrast. Combining this with the fact that the contrast was not significant when Authority was included with the binding foundations strongly suggests that Authority does not resemble the other binding foundations at the neural level. This does not seem to be an artifact of sampling error – our sample did not display atypical responses to the Moral Foundations Vignettes, and the moral wrongness judgments of Authority violations were equivalent with those of Loyalty violations. Additionally, our sample was well-balanced with respect to political ideology. Conservatism scores were just around the midpoint of the SECS, with standard deviations large enough to indicate a representative sample. And although our sample does come from a predominantly Western, educated, industrialized, rich, and democratic (WEIRD) society that places less emphasis on Authority relative to other foundations (Graham et al., 2013), the theory behind the superordinate groupings was developed precisely with respect to this context (Graham et al., 2009). However, it is also important to keep in mind that we sampled only a subset of individuals from this WEIRD society, and it is likely that their geographical proximity resulted in some amount of idiosyncrasy in their cultural values, which may place a limit on the extent to which this particular neural result generalizes to other samples from different regions. An important question for future work is characterizing how political ideology modulates MFT-related neural activity, in a larger and more politically heterogeneous sample.

Our results also aligned with prediction (iv), that we would find a latent variable dissociating between Social Norm and Moral Foundation violations. MC-LV1 demonstrated this in a totally data-driven manner, and an explicit contrast between these categories (NR-LV1)

confirmed the significance of the distinction. Activity preferentially associated with MFT violations in MC-LV1 included superior and inferior frontal, posterior cingulate, angular, middle temporal and inferior parietal gyri, all of which are core nodes of the brain's default network and have been consistently associated with moral reasoning (Buckner, Andrews-Hanna, and Schacter, 2008; Andrews-Hanna et al., 2014; Buckner & Carroll, 2007). The engagement of the default network in MFT violations was corroborated by the cluster report for NR-LV1. It likewise found that precuneus, middle temporal, superior frontal, and angular gyri were associated with MFT violations,

By contrast, social norm violations were more likely to engage regions not traditionally associated with the brain's default network, including insula, superior temporal, precentral and postcentral gyri. The largest clusters associated with Social Norms across both LVs were identified, bilaterally, in the insula, which is typically associated with the brain's salience network, rather than the default network (Menon & Uddin, 2010; Yeo et al., 2011). Insula activation has been commonly reported in response to a large variety of stimuli spanning many tasks, from interoception and autonomic control to somatic processing to chemosensory functions (see Uddin et al., 2017, for a review). Related to the current project, insula activation has been reported in relation to socio-emotional processing, both with emotionally arousing stimuli such as disgust, fear, and sadness (Chakroff et al., 2016; Decety et al., 2012; Hutcherson et al., 2015; Uddin et al., 2017; Wicker et al., 2003) as well as empathy and social cognition (Boucher et al., 2015; Fan et al., 2011), including the so-called "social pain", which is associated with social exclusion and rejection (Bolling et al., 2011; Eisenberger et al., 2003, 2011; Masten et al., 2011). The fact that MC-LV1 revealed such preferential engagement of bilateral insula during social norm violations, as opposed to MFT violations, suggest that the role played by the insula in processing social stimuli exceeds its role in processing exclusively moral information. We believe that these findings should motivate further research on the precise role insula plays in the processing of moral and non-moral social information.

We also observed a number of different clusters in the precuneus, and these were associated both with Social Norm and Moral Foundation violations. Specifically, precuneus activity was associated with Social Norms in MC-LV1, with Social Norms, Loyalty, and Fairness in MC-LV2, with both Social Norms and all the foundations in NR-LV1, and with Loyalty and Purity in NR-LV2. The precuneus has been associated with a large array of tasks,

from attentional orientation to motor imagery to music perception (see Cavanna & Trimble, 2006, for a review), and it also forms part of the brain's default network. Relevant to the current study are findings suggesting that the precuneus plays a critical role in mentalizing (Chakroff et al., 2016; Koster-Hale et al., 2013; Wasserman et al., 2017; Young & Saxe, 2011; Young et al., 2010) and specifically in processing Emotional Harm violations (Tsoi et al., 2018). Coupling these previous findings with our present observations suggests that mentalizing is a core feature of moral cognition, including assessing the moral wrongness of non-moral, social norm violations.

There was, however, one brain structure that was associated only with moral foundation violations: the amygdala. In MC-LV1, it was associated with all Loyalty, Purity, Physical Harm, Fairness, and Liberty, and Purity had a significantly higher BSR than all of the other foundations on this LV. In MC-LV2, it was the only cluster that survived correction and was associated with Physical Harm and Purity. In NR-LV1, it was associated with all of the foundations, and in NR-LV2 it was associated with Purity and Loyalty. The amygdala has long been implicated in emotional arousal (Inman et al., 2020; LeDoux, 2003; Phelps, 2006), and a meta-analysis of moral reasoning and moral emotions found amygdala involvement for processing passive and emotionally salient, as opposed to active and less emotionally salient, moral stimuli (Sevinc et al., 2014). Indeed, when we extracted the BOLD response of the left amygdala cluster in MC-LV2 (Figure S6), Purity and Harm violations activated it most strongly. We interpret this as indicative of the emotional content of Purity and Harm violations, which is further corroborated by the fact that Purity and both Harm violations elicited stronger emotional reactions on average than all of the other foundations (See Supplemental Materials, Table S11).

There are several factors limiting the scope of this study. First, given our sample size, we might have been underpowered to detect more subtle differences among foundations. As a result, we cannot rule out that the patterns observed in our study are due to sample idiosyncrasies than actual functional differences among the foundations. We attempted to mitigate this as much as possible by recruiting a politically balanced sample and ensuring that the PLSC results reported were stable across specifications of temporal windows. Additionally, we utilized a conservative reporting threshold for reporting clusters from all of the PLSC analyses to reduce the probability of false positive findings. Another concern might be that conducting multiple exploratory statistical tests inflates the false positive rate across analyses. However, all p values for the latent variables remain significant after

Bonferroni correcting the alpha value across both PLSC analyses ($\alpha = 0.025$).

Along with these statistical limitations, it is important to highlight a conceptual limitation of these findings: they cannot, in principle, support inferences about genetic predispositions or other evolutionary claims made by MFT. As pointed out by a helpful reviewer, these analyses could return differences in neural processing for categories that cannot have intrinsically differentiated neural substrates (e.g., models of cars). Thus, we echo our points made in the introduction about the task-contingency of modularity claims tested with functional neuroimaging, and emphasize that these results are more aligned with a “proof of concept” that neural activity can be decomposed along dimensions that loosely correspond to MFT posits.

Taken together, the results of the current study suggest that some but not all of the modular architecture posited by MFT extends to the neural level. We observed unique neural signatures for the individual foundations of Purity, Loyalty, Physical Harm, and Fairness, but not for any other foundations. We also found that a totally data-driven analysis of whole brain activity did differentiate between binding and individualizing foundations, but only if Authority is not included with the binding foundations. This was corroborated by two non-rotated PLSC analyses: one that explicitly contrasted traditional binding vs. individualizing foundations and found no effect and another that dropped Authority from the binding foundation and did reach statistical significance. Additionally, we found robust evidence (from both data- and theory-driven analyses) to the effect that neural activity differentiates between moral foundation violations and social norm violations.

Acknowledgements

The authors thank members of the Moral Attitudes and Decision-Making (MAD) lab and Imagination and Modal Cognition (IMC) lab for helpful feedback on previous drafts.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partially supported by the Duke Institute for Brain Sciences incubator grant awarded to E.H. and W.S.A. and additional support from the Duke Institute for Brain Sciences for FDB.

ORCID

Ari Khoudary  <http://orcid.org/0000-0002-1339-2600>
 Kevin O’Neill  <http://orcid.org/0000-0001-7401-9802>
 Scott Clifford  <http://orcid.org/0000-0002-9401-7481>
 Felipe De Brigard  <http://orcid.org/0000-0003-0169-1360>
 Walter Sinnott-Armstrong  <http://orcid.org/0000-0003-2579-9966>

References

- Andrews-Hanna, J. R., Smallwood, J., & Spreng, R. N. (2014). The default network and self-generated thought: Component processes, dynamic control, and clinical relevance. *Annals of the New York Academy of Sciences*, 1316(1), 29. <https://doi.org/10.1111/nyas.12360>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bolling, D. Z., Pitskel, N. B., Deen, B., Crowley, M. J., McPartland, J. C., Mayes, L. C., & Pelphrey, K. A. (2011). Dissociable brain mechanisms for processing social exclusion and rule violation. *NeuroImage*, 54(3), 2462–2471. <https://doi.org/10.1016/j.neuroimage.2010.10.049>
- Boucher, O., Rouleau, I., Lassonde, M., Lepore, F., Bouthillier, A., & Nguyen, D. K. (2015). Social information processing following resection of the insular cortex. *Neuropsychologia*, 71, 1–10. <https://doi.org/10.1016/j.neuropsychologia.2015.03.008>
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10(4), 433–436. <https://doi.org/10.1163/156856897X00357>
- Brown, D. E. (1991). *Human universals*. New York: McGraw-Hill.
- Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain’s default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, 1124(1), 1–38. <https://doi.org/10.1196/annals.1440.011>
- Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, 11(2), 49–57. <https://doi.org/10.1016/j.tics.2006.11.004>
- Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: A review of its functional anatomy and behavioural correlates. *Brain: A Journal of Neurology*, 129(3), 564–583. <https://doi.org/10.1093/brain/awl004>
- Chakroff, A., Dungan, J., Koster-Hale, J., Brown, A., Saxe, R., & Young, L. (2016). When minds matter for moral judgment: Intent information is neurally encoded for harmful but not impure acts. *Social Cognitive and Affective Neuroscience*, 11(3), 476–484. <https://doi.org/10.1093/scan/nsv131>
- Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods*, 47(4), 1178–1198. <https://doi.org/10.3758/s13428-014-0551-2>
- Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. <https://puso.pw/sat.pdf>
- Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1), 113–126. <https://doi.org/10.1037/0022-3514.44.1.113>

- De Brigard, F., Spreng, R. N., Mitchell, J. P., & Schacter, D. L. (2015). Neural activity associated with self, other, and object-based counterfactual thinking. *NeuroImage*, *109*, 12–26. <https://doi.org/10.1016/j.neuroimage.2014.12.075>
- Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. *Cerebral Cortex*, *22*(1), 209–220. <https://doi.org/10.1093/cercor/bhr111>
- De Waal, F. (1996). Good natured. The origins of right and wrong in humans and other animals. <http://www.diacronia.ro/en/indexing/details/B440>
- Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, *302*(5643), 290–292. <https://doi.org/10.1126/science.1089134>
- Eisenberger, N. I., Master, S. L., Inagaki, T. K., Taylor, S. E., Shirinyan, D., Lieberman, M. D., & Naliboff, B. D. (2011). Attachment figures activate a safety signal-related neural region and reduce pain experience. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(28), 11721–11726. <https://doi.org/10.1073/pnas.1108239108>
- Everett, J. A. C. (2013). The 12 item Social and Economic Conservatism Scale (SECS). *PloS One*, *8*(12), e82131. <https://doi.org/10.1371/journal.pone.0082131>
- Fan, Y., Duncan, N. W., de Greck, M., & Northoff, G. (2011). Is there a core neural network in empathy? An fMRI based quantitative meta-analysis. *Neuroscience and Biobehavioral Reviews*, *35*(3), 903–911. <https://doi.org/10.1016/j.neubiorev.2010.10.009>
- Faul, L., St Jacques, P. L., DeRosa, J. T., Parikh, N., & De Brigard, F. (2020). Differential contribution of anterior and posterior midline regions during mental simulation of counterfactual and perspective shifts in autobiographical memories. *NeuroImage*, *215*, 116843. <https://doi.org/10.1016/j.neuroimage.2020.116843>
- Ferguson, M. J. (2007). On the automatic evaluation of end-states. *Journal of Personality and Social Psychology*, *92*(4), 596–611. <https://doi.org/10.1037/0022-3514.92.4.596>
- Graham, J., & Haidt, J. (2012). Sacred values and evil adversaries: A moral foundations approach. <https://psycnet.apa.org/record/2011-09275-001>
- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In P. Devine & A. Plant (Eds.), *Advances in Experimental Social Psychology* (Vol. 47, pp. 55–130). Academic Press.
- Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. <https://doi.org/10.1037/a0015141>
- 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), 366–385. <https://doi.org/10.1037/a0021847>
- Haidt, J., & Joseph, C. (2007). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In: Carruthers, P., Laurence, S., Stich, S. (Eds.), *The Innate Mind*, (Vol. 3, pp. 367–391). New York: Oxford.
- Haidt, J., & Joseph, C. (2011). How Moral Foundations Theory succeeded in building on sand: A response to Suhler and Churchland. *Journal of Cognitive Neuroscience*, *23*(9), 2117–2122. <https://doi.org/10.1162/jocn.2011.21638>
- Haidt, J., McCauley, C., & Rozin, P. (1994). Individual differences in sensitivity to disgust: A scale sampling seven domains of disgust elicitors. *Personality and Individual Differences*, *16*(5), 701–713. [https://doi.org/10.1016/0191-8869\(94\)90212-7](https://doi.org/10.1016/0191-8869(94)90212-7)
- Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2014). Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex*, *24*(8), 1979–1987. <https://doi.org/10.1093/cercor/bht042>
- Hutcherson, C. A., Montaser-Kouhsari, L., Woodward, J., & Rangel, A. (2015). Emotional and utilitarian appraisals of moral dilemmas are encoded in separate areas and integrated in Ventromedial Prefrontal Cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(36), 12593–12605. <https://doi.org/10.1523/JNEUROSCI.3402-14.2015>
- Inman, C. S., Bijanki, K. R., Bass, D. I., Gross, R. E., Hamann, S., & Willie, J. T. (2020). Human amygdala stimulation effects on emotion physiology and emotional experience. *Neuropsychologia*, *145*, 106722. <https://doi.org/10.1016/j.neuropsychologia.2018.03.019>
- Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PloS One*, *7*(8), e42366. <https://doi.org/10.1371/journal.pone.0042366>
- Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, *62*(2), 782–790. <https://doi.org/10.1016/j.neuroimage.2011.09.015>
- Kelly, R. E., Jr., Alexopoulos, G. S., Wang, Z., Gunning, F. M., Murphy, C. F., Morimoto, S. S., Kanellopoulos, D., Jia, Z., Lim, K. O., & Hoptman, M. J. (2010). Visual inspection of independent components: Defining a procedure for artifact removal from fMRI data. *Journal of Neuroscience Methods*, *189*(2), 233–245. <https://doi.org/10.1016/j.jneumeth.2010.03.028>
- Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? https://pure.mpg.de/rest/items/item_1790332/component/file_3136265/content
- 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), 5648–5653. <https://doi.org/10.1073/pnas.1207992110>
- Krishnan, A., Williams, L. J., McIntosh, A. R., & Abdi, H. (2011). Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage*, *56*(2), 455–475. <https://doi.org/10.1016/j.neuroimage.2010.07.034>
- LeDoux, J. (2003). The emotional brain, fear, and the amygdala. *Cellular and Molecular Neurobiology*, *23*(4–5), 727–738. <https://doi.org/10.1023/A:1025048802629>
- Lenth, R. (2020). Emmeans: Estimated marginal means, aka least-squares means. R package version 1.4.6. <https://CRAN.R-project.org/package=emmeans>
- Mäkineniemi, J.-P., Pirttilä-Backman, A.-M., & Pieri, M. (2013). The endorsement of the moral foundations in food-related moral thinking in three European countries. *Journal of*

- Agricultural & Environmental Ethics*, 26(4), 771–786. <https://doi.org/10.1007/s10806-012-9401-3>
- Masten, C. L., Morelli, S. A., & Eisenberger, N. I. (2011). An fMRI investigation of empathy for “social pain” and subsequent prosocial behavior. *NeuroImage*, 55(1), 381–388. <https://doi.org/10.1016/j.neuroimage.2010.11.060>
- McAdams, D. P., Albaugh, M., Farber, E., Daniels, J., Logan, R. L., & Olson, B. (2008). Family metaphors and moral intuitions: How conservatives and liberals narrate their lives. *Journal of Personality and Social Psychology*, 95(4), 978. <https://doi.org/10.1037/a0012650>
- McIntosh, A. R., Chau, W. K., & Protzner, A. B. (2004). Spatiotemporal analysis of event-related fMRI data using partial least squares. *NeuroImage*, 23(2), 764–775. <https://doi.org/10.1016/j.neuroimage.2004.05.018>
- Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: A network model of insula function. *Brain Structure & Function*, 214(5–6), 655–667. <https://doi.org/10.1007/s00429-010-0262-0>
- Olatunji, B. O., Williams, N. L., Tolin, D. F., Abramowitz, J. S., Sawchuk, C. N., Lohr, J. M., & Elwood, L. S. (2007). The disgust scale: Item analysis, factor structure, and suggestions for refinement. *Psychological Assessment*, 19(3), 281–297. <https://doi.org/10.1037/1040-3590.19.3.281>
- Parkinson, C., Sinnott-Armstrong, W., Koralus, P. E., 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), 3162–3180. https://doi.org/10.1162/jocn_a_00017
- Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89(3), 277–293. <https://doi.org/10.1037/0022-3514.89.3.277>
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10(4), 437–442. <https://doi.org/10.1163/156856897X00366>
- Phelps, E. A. (2006). Emotion and cognition: Insights from studies of the human amygdala. *Annual Review of Psychology*, 57(1), 27–53. <https://doi.org/10.1146/annurev.psych.56.091103.070234>
- R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>
- Sevinc, G., Spreng, R. N., & Soriano-Mas, C. (2014). Contextual and perceptual brain processes underlying moral cognition: A quantitative meta-analysis of moral reasoning and moral emotions. *PLoS One*, 9(2), e87427. <https://doi.org/10.1371/journal.pone.0087427>
- Shweder, R. A. (1990). In defense of moral realism: Reply to Gabennesch. *Child Development*, 61(6), 2060. <https://doi.org/10.2307/1130859>
- Sinnott-Armstrong, W. (2016). The disunity of morality. In S. Matthew Liao (Ed.), *Moral brains: The Neuroscience of Morality* (pp. 331–354). Oxford University Press.
- Tamborini, R., Eden, A., Bowman, N. D., Grizzard, M., & Lachlan, K. A. (2012). The influence of morality subcultures on the acceptance and appeal of violence. *The Journal of Communication*, 62(1), 136–157. <https://doi.org/10.1111/j.1460-2466.2011.01620.x>
- Tsoi, L., Dungan, J. A., Chakroff, A., & Young, L. L. (2018). Neural substrates for moral judgments of psychological versus physical harm. *Social Cognitive and Affective Neuroscience*, 13(5), 460–470. <https://doi.org/10.1093/scan/nsy029>
- Uddin, L. Q., Nomi, J. S., Hébert-Seropian, B., Ghaziri, J., & Boucher, O. (2017). Structure and function of the human insula. *Journal of Clinical Neurophysiology: Official Publication of the American Electroencephalographic Society*, 34(4), 300–306. <https://doi.org/10.1097/WNP.0000000000000377>
- Van Berkum, J. J. A., Holleman, B., Nieuwland, M., Otten, M., & Murre, J. (2009). Right or wrong? The brain’s fast response to morally objectionable statements. *Psychological Science*, 20(9), 1092–1099. <https://doi.org/10.1111/j.1467-9280.2009.02411.x>
- Wasserman, E. A., Chakroff, A., Saxe, R., & Young, L. (2017). Illuminating the conceptual structure of the space of moral violations with searchlight representational similarity analysis. *NeuroImage*, 159, 371–387. <https://doi.org/10.1016/j.neuroimage.2017.07.043>
- Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in my insula: The common neural basis of seeing and feeling disgust. *Neuron*, 40(3), 655–664. [https://doi.org/10.1016/S0896-6273\(03\)00679-2](https://doi.org/10.1016/S0896-6273(03)00679-2)
- Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., & Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3), 1125–1165. <https://doi.org/10.1152/jn.00338.2011>
- Young, L., Camprodon, J. A., Hauser, M., Pascual Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of the United States of America*, 107(15), 6753–6758. <https://doi.org/10.1073/pnas.0914826107>
- Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, 120(2), 202–214. <https://doi.org/10.1016/j.cognition.2011.04.005>