Moral Foundations Elicit Shared and Dissociable Cortical Activation Modulated by Political Ideology

Frederic Hopp
University of Amsterdam
Ori Amir
Pomona College
Jacob Fisher
University of Illinois Urbana-Champaign https://orcid.org/0000-0002-2968-2557
Scott Grafton
UCSB
Walter Sinnott-Armstrong
Kenan Institute for Ethics
René Weber (renew@comm.ucsb.edu)
University of California Santa Barbara https://orcid.org/0000-0002-8247-7341

Article

Keywords:

Posted Date: October 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2133317/v1

Version of Record: A version of this preprint was published at Nature Human Behaviour on September 7th, 2023. See the published version at https://doi.org/10.1038/s41562-023-01693-8.

Abstract

Moral Foundations Theory (MFT) holds that moral judgments are driven by modular and ideologically variable moral foundations, but where and how they are represented in the brain and shaped by political beliefs remains an open question. Using a moral judgment task of moral foundation vignettes, we probed the neural (dis)unity of moral foundations. Univariate analyses revealed that moral judgment of moral foundations, versus conventional norms, reliably recruits core areas implied in emotional processing and theory of mind. Yet, multivariate pattern analysis demonstrated that each moral foundation has dissociable neural representations distributed throughout the cortex. As predicted by MFT, political ideology modulated neural responses to moral foundations. Our results confirm that each moral foundation recruits domain-general mechanisms of social cognition, but has a dissociable neural signature malleable by sociomoral experience. We discuss these findings in view of unified versus dissociable accounts of morality and their neurological support for MFT.

Introduction

Human morality comprises a set of diverse norms that prescribe what is considered morally right or wrong (Buckholtz & Marois, 2012). Despite ongoing debates about the specific nature and function of morality (Curry, 2016; Haidt, 2001; Shweder et al., 1997; Suhler & Churchland, 2011), diversity and pluralism in moral norms is increasingly recognized (Curry et al., 2021; Graham et al., 2013; Haidt & Joseph, 2007; Sinnott-Armstrong & Wheatley, 2012). Moral Foundations Theory (MFT, Graham et al., 2013; Haidt, 2007) posits that there are five universal, but contextually variable groups of moral intuitions: *Care/harm*, *Fairness/cheating*, *Loyalty/betrayal*, *Authority/subversion*, and *Sanctity/degradation*. A sixth *Liberty/oppression* foundation has been identified (Iyer et al., 2012). Despite the wide-ranging relevance of moral foundations (Amin et al., 2017; Brady et al., 2017; Hoover et al., 2021; Mooijman et al., 2018; Morgan et al., 2010), where and how they are represented in the brain remains an open question.

A central proposition of MFT is that moral foundations are sufficiently distinct to be treated as separate cognitive modules (Haidt & Joseph, 2007), yet behavioral evidence for this prediction remains mixed (Doğruyol et al., 2019; Dungan & Young, 2012). Although mounting literature suggests that different forms of moral judgment may rely on distinguishable neural systems (FeldmanHall et al., 2014; Greene & Haidt, 2002; Parkinson et al., 2011; Tsoi et al., 2018; Wasserman et al., 2017; Young, & Dungan, 2012), no neuroimaging study has examined the full spectrum of morality as proposed by MFT.

Furthermore, research on MFT consistently reports that liberals (progressives) endorse individualizing intuitions of Care and Fairness more than the binding intuitions of Loyalty, Authority, and Sanctity, whereas conservatives endorse all five foundations more or less equally (Graham et al., 2009; Kivikangas et al., 2020). While individualizing and binding moral concerns have been related to individual differences in regional brain volume (Lewis et al., 2012; Nash et al., 2017), it is unknown whether political ideology moderates *functional* variation in neural responses to moral foundation vignettes.

We tested MFT's predictions in an fMRI study (n = 64; Fig. 1a.), using a standardized and validated stimulus database of moral foundation vignettes (MFVs; Clifford et al., 2015). The MFVs were designed for use in neuroimaging studies, describe behaviors that violate a particular moral foundation and not others, and also feature transgressions of conventional, social norms as a non-moral control category. Using functional neuroimaging allowed us to examine whether moral intuitions are encoded in unified or distributed brain systems. Furthermore, by treating moral foundations as multivariate neural representations, we moved beyond asking *where* moral foundations are located in the brain and examined *how* moral foundations are neurally organized. Specifically, we tested whether moral foundations have shared or dissociable neural signatures, and where substructures of moral foundations can be recovered. Finally, we combined neural responses to the MFV with behavioral data to examine whether political ideology moderates individual differences in neural processing of moral foundations.

Results

Behavioral results.

During the fMRI experiment, judgments of moral wrongness (1 = *not morally wrong* to 4 = *extremely morally wrong*) were collected from each participant for each of the 120 vignettes (Fig. 1a). Each vignette described a violation of either: (1) one of seven moral foundations (Physical Care, Emotional Care, Fairness, Liberty, Authority, Loyalty, Sanctity), or (2) a non-moral social norm transgression. 15 vignettes were presented in each condition. Vignettes were randomly distributed across three runs, with 5 vignettes per condition per run (40 vignettes/run).

To assess the impact of vignette condition on moral wrongness ratings and response times, we used linear mixed-effect modeling and performed likelihood ratio tests (LRTs) to test whether the model including vignette condition (moral vs. non-moral) would provide a better fit to the data than a model without (Tsoi et al., 2018). There was a significant main effect of vignette condition ($\chi2$ (7) = 202.29, p < 0.001). Pairwise contrasts revealed that each moral violation was judged as more morally wrong than were social norm transgressions (Fig. 1b): Physical Care (z = 18.315, p < 0.001), Emotional Care (z = 16.119, p < 0.001), Fairness (z = 17.186, p < 0.001), Liberty (z = 16.869, p < 0.001), Loyalty (z = 12.197, p < 0.001), Authority (z = 12.975, p < 0.001), and Sanctity (z = 16.895, p < 0.001). This observation assured that ensuing contrasts between moral versus non-moral vignettes reflect neural activity relevant to participants' judgments of *moral* wrongness. There was also a main effect of vignette condition on response times (RT; $\chi2(7)$ = 74.088, p < 0.001). Ratings of moral transgressions were slower than were ratings of non-moral transgressions (Fig. 1c): Physical Care (z = 4.923, p < 0.001), Emotional Care (z = 4.351, p < 0.001), Fairness (z = 4.092, p = 0.001), Liberty (z = 5.624, p < 0.001), Loyalty (z = 8.869, p < 0.001), Authority (z = 7.086, p < 0.001), and Sanctity (z = 6.632, p < 0.001). These findings suggest that judging moral transgressions, as compared with non-moral social norm violations, may involve a deeper evaluation of individuals' actions and how they relate to one's own values (Parkinson et al., 2011).

Identifying brain regions that encode moral foundations.

We examined the neural activation underlying the moral judgment across different moral foundations with a univariate General Linear Model (GLM), contrasting each type of moral violation with the social norm violations (Fig. 2a–g). We also computed contrasts for the binding (Loyalty, Authority, Sanctity) versus the individualizing (Physical Care, Emotional Care, Fairness) foundations (Fig. 2h), and for all moral foundations versus social norm violations (Fig. 2i). For each contrast, participants' statistical parametric maps were used to perform group-level analyses. To correct for multiple comparisons, images from the group-level analyses were subjected to a voxel-wise threshold of $p < 0.001$ (uncorrected) and a cluster extent threshold ensuring $q < 0.01$ (false discovery rate (FDR)-corrected).

Moral judgment of each moral foundation versus social norms (Fig. 2a–g) recruited multiple regions (cluster peaks are summarized in Supplementary Table 1), demonstrating that no foundation is encoded in a single area but distributed throughout the brain. To identify whether any neural system was *uniquely* activated across each of the seven moral foundation versus social contrasts (Fig. 2a–g), we combined the corresponding thresholded statistical maps and determined which voxel preferentially responded to only one of the seven moral foundations. We found that each moral foundation evoked dissociable neural activation (see Supplementary Fig. 1), confirming that each moral foundation relies partially on separable brain systems. Notably, the Sanctity foundation recruited most areas not evoked by any other moral foundation, including the Thalamus and bilateral caudate nucleus, highlighting its neurological distinctiveness from other moral domains (Parkinson et al., 2011; Wasserman et al., 2017; Young & Saxe, 2009).

Next, we examined which neural systems were *commonly* activated across each of the seven moral foundation versus social contrasts (Fig. 2a–g) by performing a conjunction analysis of the corresponding thresholded statistical maps (Nichols et al., 2005). We find common activation in dorsomedial prefrontal cortex (dmPFC), posterior cingulate cortex (PCC) and precuneus (PC), bilateral temporoparietal junction (TPJ), supplementary motor area (SMA), and primary visual cortex (V1). Previous studies have only shown independent activation in dmPFC when reasoning through harmful, disgusting, and dishonest versus neutral scenarios (Parkinson et al., 2011). Here we find that evaluating violations of moral foundations versus social norms not only jointly recruits the dmPFC, but multiple additional areas commonly observed in the moral neurosciences (Eres et al., 2018; FeldmanHall & Mobbs, 2015). Furthermore, these areas largely overlap with the theory of mind (ToM) network, which has been attributed a central function in moral judgment studies (FeldmanHall & Mobbs, 2015; Greene & Haidt, 2002; Tsoi et al., 2018; Wasserman et al., 2017; Young, & Dungan, 2012).

Contrasting binding versus individualizing foundations also indicated that binding foundations correspond to greater activity in regions relating to ToM (Fig. 2h). While every vignette only described an action without explicitly stating an agent's intention or ensuing outcomes, group-based moral foundations may have triggered spontaneous mental state inference (Young & Saxe, 2009) to assess hidden goals and determine whether an action will have harmful consequences. In contrast, violations of individualizing foundations may be quickly categorized as blatant moral wrongs without eliciting deeper mentalizing processes, for instance, because they will directly result in physical harm. This notion is

supported by both faster RTs and higher moral wrongness ratings for individualizing violations compared to slower RTs and lower moral wrongness ratings for binding violations (Fig. 2b–c). Analogously, Amit and Greene (2012) showed that high-level action construals (abstract, goal-focused) favor utilitarian moral judgment, whereas low-level action construals (concrete, means-focused) facilitate deontological moral judgment. Although speculative, violations of binding foundations may have elicited more abstract construals, whereas transgressions of individualizing foundations may lend themselves more to low-level, concrete construals.

Moral foundations elicit dissociable cortical activation patterns.

Univariate analyses are useful to examine which brain regions are more engaged during moral (versus social) judgment. However, these approaches cannot determine whether the brain, individual networks, or specific regions show convergence of multivoxel patterns for moral foundations. Hence, we employed a multivariate pattern analysis approach, training a support vector machine (SVM) to examine whether moral foundations elicit shared or dissociable neural activation patterns. To this end, we first averaged the runwise, whole-brain statistical parametric maps (beta estimates) for each vignette condition and participant from the GLM reported earlier, creating one average beta map per condition and participant. We then used a leave-one-subject-out (LOSO) cross-validation procedure to evaluate the performance of our SVM model in classifying which of two vignette conditions a participant was judging using data from the rest of the participants (a “forced-choice” test). Forced-choice tests compare the relative pattern expression of the model between brain maps within the same participant and are particularly well suited for fMRI because they do not require signals to be on the same scale across individuals (Wager et al., 2013). Moreover, pairwise classification–as opposed to classifying between *all* moral foundations–allowed us to determine whether every pair of foundations can be reliably decoded.

The SVM was able to accurately distinguish between all moral versus social vignette conditions within each participant (all forced-choice accuracies ≥ 98%, $p < 0.001$; Fig. 3a). Compellingly, within the moral space, the SVM also accurately distinguished between every one of the moral foundations (all forced-choice accuracies ≥ 89%, $p < 0.001$), suggesting that moral foundations are indeed distinctly represented in the brain. Speaking to the validity of the model, performance was relatively lower for moral foundations that have been argued to be similar in content and function: Physical Care versus Sanctity (96.87%) and Fairness versus Liberty (89.06%).

Next, we examined whether morally-relevant activation patterns are distributed throughout the brain, or localized to specific networks or regions identified in the literature. To this end, we created a mask from a meta-analytic map associated with the term “moral” from Neurosynth (uniformity test map, thresholded at $P_{FDR} < 0.01$, Fig. 3c; Yarkoni et al. 2011). This mask was chosen to select voxels that are presumably activated in moral judgment studies. Furthermore, we selected regions of interest (ROI) commonly reported in moral neuroscience (Eres et al., 2018) from an independent parcellation based on meta-analytic functional coactivation of the Neurosynth database (Yarkoni et al. 2011; see Supplementary Fig. 2). We then retrained the SVM using only the averaged beta estimates from voxels of each of these

masks. Notably, the SVM trained solely on the Neurosynth mask (N_{voxels} = 1,595) approached the accuracy of the whole-brain (N_{voxels} = 238,955) model (mean difference in forced-choice accuracy: 1.1%, $t(28) = 1.52$, $p = .134$; Fig. 2b). This finding validates that voxels commonly recruited during moral judgment may not only be consistently *activated* during moral judgment, but do contain relevant information to *distinguish* moral foundations. As expected, individual ROIs comprising the moral brain also contained information to reliably decode each pair of moral foundations (Fig. 3b). Yet, none of the functionally-parcelated networks reached the accuracy of the moral mask or whole-brain model. This provides compelling evidence that moral foundations are not encoded within isolated “moral hotspots”, but distributed across the human cortex.

To corroborate these findings, we trained a multiclass SVM to distinguish between *all* MFV conditions, applying both a leave-one-run-out (LORO) within-subject classifier (WSC) and LOSO between-subject classifier (BSC) using only voxels from Neurosynth’s moral mask. The WSC was able to discriminate between all vignette conditions with high accuracy (average classification accuracy: 83.14%; chance level: 12.5%; $t(64) = 89.82$, $p < .000$), demonstrating that cortical activation was dissociable across MFV conditions within each participant (Fig. 3d). As expected, social norms were most accurately classified. The BSC also accurately discriminated across all MFV vignette conditions, but with even higher accuracy than the WSC classifier (average classification accuracy: 92.94%; chance level: 12.5%; $t(64) = 139.43$, $p < .000$), likely due to the greater number of training samples available. These results undergird that neural activation patterns for each moral foundation can reliably be decoded in morally relevant voxels, both *within* and *across* participants.

Neural representational mapping of moral foundations.

Having established that moral foundations elicit dissociable neural activation patterns raises the question how these patterns are organized in the high-dimensional activity space. We used representational similarity analysis (RSA; Kriegeskorte et al., 2008) to examine the cortical structure and hierarchical division of moral foundations. We first created conceptual representational dissimilarity matrices (RDMs) denoting different theoretical predictions about the representational geometry of moral foundations (Fig. 4a). Specifically, we computed four conceptual models that monotonically increase the nested hierarchy across moral foundations: (1) independent (no similarity); (2) ind/bind/social (similarity within individualizing foundations, binding foundations, and social norms, but dissimilarity between these clusters); (3) ind:bind/social (individualizing foundations similar to binding foundations but dissimilar to social norms); (4) moral/social (moral foundations similar to each other but dissimilar to social norms. Additionally, we created behavior-based RDMs based on participants’ moral wrongness ratings and response times. Next, we obtained runwise statistical maps for each vignette condition and participant from the GLM reported earlier. These statistical maps were then used to compute neural reference RDMs (Fig. 4b). We constrained the neural RDMs to independent parcels implicated in moral judgment as well as Neurosynth’s moral mask, and also examined visual cortex (V1) to study representations of moral intuitions in a lower-order processing system.

In a first exploratory step, we computed the whitened cosine similarity (Diedrichsen et al., 2020) across conceptual, behavior-based, and neural RDMs and subjected the resulting similarity matrix to a hierarchical clustering algorithm (Fig. 4c, left panel). This revealed that the representational geometry of moral foundations is more similar within ROIs of the same functional network as indexed by clusters spanning ToM (vmPFC, TPJ, PCC/Precuneus, PCC/Superior LOC, and STS) and executive control/conflict monitoring networks (dlPFC, dACC). Compellingly, the cluster solution of dmPFC and Neurosynth's moral mask suggests that the neural population codes of moral foundations is highly similar for both cortical structures, undergirding the central role of dmPFC for representing moral information in larger, more distributed moral judgment networks (Parkinson et al., 2011). Visual inspection of a t-distributed stochastic neighbor embedding (t-SNE; Fig. 4c, right panel) of RDM similarities revealed that the visual cortex represents moral foundations distinctly from other brain networks. Notably, the close proximity of V1 to conceptual models predicting distinct representations of individualizing, binding, and social norms indicates that the vignettes may have elicited qualitatively different mental imaginations. This aligns with research demonstrating that visual cortex is activated not only during visual perception but also during visual imagery (O'Craven & Kanwisher, 2000), and that this activation correlates with the vividness of the imagery (Cui et al., 2007). Considering that the MFV were extensively pretested to ensure "the ease with which a vivid image could be formed from each vignette" and only differed in the described moral action (Clifford et al., 2015), it is likely that differences in low-level action representation are most salient along the visual stream. These results also corroborate the importance of morally relevant information in directing perceptual awareness (Gantman & Van Bavel, 2015; Gantman et al., 2020) and suggest that distinctions between moral versus non-moral stimuli are formed early in the visual system.

Importantly, the close proximity of the moral judgment RDM to V1 indicates that visual representations of moral foundations informed moral wrongness ratings, undergirding the role of visual cortex and imagery in moral judgment (Amit & Greene, 2012; Caruso & Gino, 2011; Kahane et al., 2012). In contrast, (dis)similarities in response times were rather reflected in ROIs involved in problem solving (dACC, dlPFC) and mentalizing (dmPFC, PCC). Given that individualizing foundations and social norms were judged faster than binding intuitions (Fig. 1c), moral judgment of group-oriented foundations may have recruited deeper mentalizing and intent inference processes. In line with this reasoning, the Loyalty/betrayal foundation–which is strongly concerned with tribalism and "us versus them" thinking–showed the most distinct neural representations across ToM networks (PCC/Superior LOC, STS, and Neurosynth's moral mask).

To confirm these descriptive results, we conducted statistical inference on the RDMs (Fig. 5). We first compared the performance of every candidate model in each ROI against chance to determine whether there is a statistically significant association between model and neural RDM. We found that conceptual models denoting hypothetical structures across moral foundations were significantly related to neural representations of moral foundations in the majority of our a priori ROIs, confirming that these networks represent theoretically-specified patterns of moral foundations. Yet, in parcels of the insula and in TPJ/Parietal Operculum, most conceptual models did not reach statistical significance.

Next, we compared the performance of models in each ROI to adjudicate whether some models explain the neural representations of moral foundations better than others. Across conceptual models, we found that the nested hierarchy structure across moral foundations (ind:bind/social) as well as a rigid split between moral versus social norms (moral/social) better predicts neural patterns in V1 than models denoting either complete independence or a rigid division into individualizing, binding, and social norms (ind/bind/social). In vmPFC and Neurosynth's moral mask, the independent model performed better than the ind/bind/social model, but other conceptual model comparisons were not significant. In contrasting behavior-based models, we find that response times were a significantly better predictor for neural representations in ToM networks including vmPFC, dmPFC, TPJ/Angular Gyrus, PCC/Precuneus, PCC/Superior LOC, STS, and Neurosynth's moral mask than were participants' self-reported moral wrongness ratings. In contrast, we find that the moral judgment RDM was a significantly better predictor in V1 than the response time RDM. These findings undergird that variability in response times for reaching a moral judgment is more closely reflected in brain networks associated with spontaneous mental state inference, whereas the visual representation of moral foundations may better predict the assigned moral wrongness rating.

Political ideology modulates neural responses to moral foundations.

Research on MFT consistently reports that liberals (progressives) endorse individualizing intuitions more than binding, whereas conservatives endorse all foundations more or less equally (Graham et al., 2009; Kivikangas et al., 2020). We examined this 'moral foundations hypothesis' (Graham et al., 2009) combining neural responses to the vignettes with self-report and behavioral data. Participants indicated their political ideology using a slider from "very liberal" (0) to "very conservative" (100; Fig. 6a) and also reported their sensitivity towards each moral foundation via the Moral Foundations Questionnaire (MFQ; Graham et al., 2009). Participants' moral wrongness ratings for the MFV were used as behavioral markers for moral judgment. Starting with the self-report data, we largely replicated Graham and colleagues' (2009) initial result, finding that conservatives rated the individualizing foundations of the MFQ as less relevant to their moral judgment than did liberals, whereas conservatives rated the binding foundations as more relevant to their moral judgment than did liberals (Fig. 6b). Controlling for age and gender revealed that political ideology remained the only significant predictor for the endorsement of each moral foundation as indexed by the MFQ with the exception of Care, for which gender was a stronger predictor. We observed similar patterns for responses to the MFV (Fig. 6c): The positive slopes for the binding foundations mean that conservatives rated these vignettes as more morally wrong than did liberals, whereas the negative slopes for Physical Care and Liberty mean that conservatives rated these as less morally wrong than did liberals. Interestingly, the positive slopes for Fairness and Social Norms suggests that conservatives also rated these vignettes as more morally wrong than did liberals, whereas no differences emerged for vignettes pertaining to Emotional Care. We also find that political ideology remains a significant predictor for moral wrongness ratings for each vignette condition–with the exception of Emotional Care–when controlling for age and gender.

Next, we examined whether liberals and conservatives show differential neural activity while processing moral foundations. For each participant, we first retrieved the seven *t*-contrast images from the previous GLM that modeled each type of moral violation versus social norm violations (Fig. 2a–g). In a second step, we divided participants into either “liberal” or “conservative” groups by performing a median split on the political orientation scale (median = 41). We then entered the contrast estimates into a 2 x 7 repeated-measures ANOVA and performed an *F*-test for the group-type × condition-type interaction. While conservative correction for multiple comparisons (voxelwise $p < .001$, (uncorrected); FDR $q < .05$) revealed no significant clusters, using a more lenient threshold (voxelwise $p < .005$ (uncorrected), $k > 10$)

showed that liberals and conservatives differentially processed moral foundations in several networks related to semantic and affective processing (Fig. 6d; for cluster peaks, see Supplementary Table 3), including the lingual gyrus, PC, left orbitofrontal cortex (OFC), and anterior prefrontal cortex (aPFC), as well as the temporal pole. Particularly PC and OFC have been involved in idiosyncratic processing of political debates between liberals and conservatives (van Baar et al., 2021). Analogously, the aPFC exerts robust meta-analytic functional coactivation with the medial prefrontal cortex ($r = 0.48$; Yarkoni et al., 2011), a region that has been linked to “neural polarization” in liberal versus conservative leaning individuals (Leong et al., 2020). Furthermore, the likely function of the temporal pole for binding complex, highly processed perceptual inputs to visceral emotional responses (Olsen et al. 2007) underlines that liberals and conservatives may have differential affect-laden imagery of moral transgressions that influence ensuing moral judgments.

Discussion

We reported an fMRI experiment that offers a neurobiological account of the distinguishable nature of moral judgment across moral foundations as outlined in Moral Foundations Theory (MFT). Replicating previous studies (Clifford et al., 2015), we showed that people deemed vignettes displaying transgressions of moral foundations–as opposed to conventional social norms–as more morally wrong. Each evaluation of a moral foundation versus social norms was associated with separable and overlapping brain systems, suggesting that moral foundations have specialized and shared neural bases. The common involvement of dmPFC, PCC/PC, and TPJ across all moral foundations and in past neuroimaging studies of moral judgment confirm that moral foundations are encoded in mental faculties that are characteristic of human moral cognition. Corroborating the group-based nature of binding moral intuitions, we demonstrated that transgressions of Loyalty, Authority, and Sanctity, compared to individualizing intuitions of Care and Fairness, resulted in greater activity in regions associated with processing of others, as opposed to self (Van Overwalle, 2009; Northoff et al., 2006).

The conjoined activation of dmPFC, PCC/PC, bilateral TPJ, SMA, and V1 across the spectrum of moral foundations may suggest that moral foundations are unified in these brain systems, and thus not modular enough to be treated as separate cognitive modules. However, simultaneous activation for a set of stimuli does not imply that the stimuli are phenomenologically unified in the brain. For instance, viewing images of houses and cats reliably activates the ventral temporal cortex, yet the multivariate

activation patterns of these stimuli are highly dissociable (Haxby et al., 2001). Analogously, we showed that a cross-validated multivariate pattern classifier can accurately distinguish the neural signatures between all and every pair of moral foundations. This demonstrates that moral foundations indeed elicit dissociable cortical activation patterns. In line with MFT's "massive modularity hypothesis" (Sperber, 2005), we showed that morally-informative voxels are distributed throughout the brain and not limited to any single region. In fact, morally relevant, a priori regions discriminated slightly less accurately between moral foundations compared to a distributed, whole-brain model. Going forward, illuminating the specificity and sensitivity of these moral signatures beyond text-based moral vignettes has relevant implications for the study of morality and for identifying biomarkers for pathological moral judgment (Woo et al., 2017).

By studying the representational geometry of moral foundations, we showed that the neural organization of moral foundations aligns with MFT's predicted organizational structure in core regions of the moral brain. The observation that moral foundations are more similar in functionally-connected areas spanning ToM and executive control/conflict monitoring networks corroborates that these networks encapsulate distinct aspects of moral judgment (e.g., spontaneous mental state attribution versus blame computation). Rigid hierarchical divisions of moral foundations into binding and individualizing categories emerged in early visual cortex and moral wrongness ratings, undergirding that visual imagery facilitates low-level construals of moral actions which subsequently inform moral judgment (Amit & Greene, 2012; Caruso & Gino, 2011; Kahane et al., 2012). These findings provide a fertile ground for richer delineations of the moral space by moving beyond taxonomies limited to harm versus purity (Tsoi et al., 2018; Wassermann et al., 2018).

Moreover, our results shed light on the neural systems that modulate liberals' and conservatives' idiosyncratic responses to individualizing and binding moral intuitions. Previous studies have shown that individualizing is linked to increased volume in dmPFC and reduced volume in bilateral PC, whereas increased adherence to binding foundations was positively related to bilateral subcallosal gyrus volume and reduced volume in ACC and LPFC (Lewis et al., 2012; Nash et al., 2017). Here we showed that liberals and conservatives exert differential neural responses when evaluating moral foundations in PC, OFC, and aPFC, as well as lingual gyrus and right TP, suggesting that political ideology moderates the social-affective experience of moral violations. These findings confirm that moral foundations modulate neural polarization processes (Leong et al., 2020) and may offer solutions for message interventions that re-align representations of core human values across political divides (van Baar et al., 2021).

Taken together, we showed that MFT's moral foundations robustly recruit mental systems commonly observed in the moral neurosciences, albeit the multivariate patterns associated with moral foundations are highly dissociable and distributed throughout the cortex. We encourage the use of the MFV as a controlled localizer for tracing neural signatures of moral foundations and for illuminating the space of moral violations that separate political camps.

Online Methods

Participants.

Healthy volunteers were recruited from the University of California Santa Barbara (UCSB) Department of Communication participant pools and from the local Santa Barbara community. Exclusion criteria included a history of systemic or neurological disorders, psychiatric disorders, psychoactive medication or drug use, and pregnancy. For the fMRI study, we recruited 64 participants (33 males; mean age 20.78 years; 63 right-handed; 63 native English speakers), who completed a moral judgment task divided across three runs in the fMRI scanner. No statistical methods were used to predetermine sample size, but our sample size is well above those reported in previous fMRI studies using similar moral judgment tasks (Parkinson et al., 2011; Tsoi et al., 2018; van Baar et al., 2019). No participants were excluded from analyses, but one run for a single participant could not be completed due to issues with stimulus presentation.

Procedure.

The study took place at UCSB's Brain Imaging Center and was approved by the institutional review board of the University of California at Santa Barbara (protocol number: 21-17-0123). Participants completed a battery of online trait questionnaires approximately 1 week before the MRI session. After providing informed consent upon arrival to the laboratory, participants underwent a first fMRI scan, divided across three runs, while completing the moral foundation vignettes paradigm (Clifford et al., 2015). Thereafter, participants completed two additional fMRI scans not reported here. Before departing the laboratory all participants completed debriefing questionnaires.

Moral judgment task.

Participants were presented with the moral foundation vignettes (MFVs; Clifford et al., 2015) while undergoing fMRI. The MFVs span 120, one sentence descriptions (14–17 words) detailing the violation of one (and only one) of seven moral foundations: Physical Care, Emotional Care, Fairness, Liberty, Loyalty, Authority, and Sanctity. The vignettes also contain a non-moral, social norm transgression category. Each of the eight conditions features 15 vignettes. Vignettes were organized in an event-related design, randomly distributed over three approximately 8-minute functional runs, with five vignettes of each category in each of the three runs. Participants viewed one vignette at a time and were instructed to vividly imagine the described scene. While the vignette was on screen, participants were asked to make a judgment of how morally wrong the action described in the vignette was using an MRI-safe button box (1 = *not morally wrong* to 4 = *extremely morally wrong*). After 8 seconds, the vignette disappeared but the scale remained on screen and participants could still respond during the inter-trial interval (ITI). ITIs were on average 4 seconds long, with a jitter of +/- 2.16 seconds (jitter length was calculated so that each trial would begin at exactly the beginning of the scanner's collection of the next volume).

Behavioral analyses.

Behavioral analyses were conducted in R (version 3.6.3) and Python (version 3.7.10). Moral judgments were analyzed using cumulative link mixed models with an ordinal response term (from a scale of 1–4). Mixed models were run using the package “ordinal” (Christensen, 2015). We were primarily interested in understanding whether ratings differed across vignette conditions (physical care, emotional care, fairness, liberty, loyalty, authority, and sanctity, and social norm transgressions). Our full model included only vignette condition as a predictor. Participant and item were included as random effects, and we fit an intercept for each participant and for each item, allowing the intercept to vary across individuals and items. To assess the importance of our predictor, we performed likelihood ratio tests (LRTs) to test whether the model including the vignette condition would provide a better fit to the data than a model without that term. Response times were analyzed using linear mixed effect models using “lme4” (Bates et al., 2015) in R, with the same predictor and random effects as the analyses for ratings above. Participants’ responses to the Moral Foundations Questionnaire (MFQ; Graham et al., 2011) and the MFV were regressed onto their political orientation (0 = extremely liberal; 100 = extremely conservative) using the statsmodel (https://www.statsmodels.org/) package.

fMRI acquisition and preprocessing.

fMRI scanning was performed on a 3-Tesla Siemens Magnetom Prisma with a Siemens head coil, at the Brain Imaging Center of the University of California, Santa Barbara. Functional images were taken using a multiband echo-planar gradient sequence (repetition time = 720 ms, echo time = 37 ms, flip angle = 52°, field of view = 208 mm, acceleration factor = 8). Volumes consisted of 72 interleaved slices (2 mm isotropic) acquired with an angle of ~ 20º relative to the AC-PC plane, so that the slices are acquired more dorsally near the eyes relative to the back of the brain (in that fashion we were able to acquire the entire brain volume including the cerebellum for every participant). High-resolution T1-weighted whole brain acquisitions were collected prior to functional image acquisition (repetition time = 2500 ms, echo time 2.22 ms, flip angle = 7°, field of view = 241 mm, .9 mm isotropic resolution).

Results included in this manuscript come from preprocessing performed using *fMRIPrep* 20.2.1 (Esteban, Markiewicz, et al. (2019); RRID:SCR_016216), which is based on *Nipype* 1.5.1 (Gorgolewski et al., 2011; Gorgolewski et al. 2018); RRID:SCR_002502). The pipeline description below was copied from the fMRIprep boilerplate text, leaving out unused components.

The T1w image was corrected for intensity nonuniformity with N4BiasFieldCorrection (Tustison et al. 2010), distributed with ANTs 2.3.3 (Avants et al. 2008, RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a *Nipype* implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, RRID:SCR_002823, Zhang, Brady, and Smith 2001). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization:

ICBM 152 Nonlinear Asymmetrical template version 2009c [Fonov et al. (2009), RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym].

For each of the three BOLD runs per subject, the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. The BOLD reference was then co-registered to the T1w reference using flirt (FSL 5.0.9, Jenkinson and Smith 2001) with the boundary-based registration (Greve and Fischl 2009) cost function. Coregistration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson et al. 2002). BOLD runs were slice-time corrected using 3dTshift from AFNI 20160207 (Cox and Hyde 1997, RRID:SCR_005927). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying the transforms to correct for head-motion. These resampled BOLD time-series will be referred to as preprocessed BOLD in original space, or just preprocessed BOLD. The BOLD time-series were resampled into standard space, generating a preprocessed BOLD run in MNI152NLin2009cAsym space. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD was computed using two formulations following Power (absolute sum of relative motions, Power et al. (2014)) and Jenkinson (relative root mean square displacement between affines, Jenkinson et al. (2002)). FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by Power et al. 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al. 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos 1964).

fMRI GLM analysis.

Whole-brain univariate GLM analyses were conducted in SPM12 www.fil.ion.ucl.ac.uk/spm) using custom scripts written in Nipype (Gorgolewski et al., 2011). Each run started with a tail of 11 TRs which were discarded. Thereafter, preprocessed images were spatially smoothed using a Gaussian filter (full-width half-maximum = 6 mm kernel) and analyzed using a general linear model (GLM). A GLM was constructed for each participant using boxcar regressors for each of the eight vignette conditions. For each trial, a 7.92-second window (11 TRs) of volumes of interest was included as an event, capturing the time the vignette text was displayed on screen. The three runs were modeled by separate regressors in the same GLM. To account for residual variance, the temporal derivative of each condition regressor was added to the model as well as a constant regressor for each entire run. The resulting GLM was convolved

with SPM's canonical hemodynamic response function. The model was corrected for temporal autocorrelations using a first-order autoregressive model. In addition, we included DVARS (D, temporal derivative of time courses; VARS, RMS variance over voxels), framewise displacement, six anatomical component-based noise correction (CompCor) components and four cosine drift terms as well as a standard high-pass filter (90s cutoff) to exclude low-frequency drifts. Planned contrasts then modeled activation unique to each moral violation (e.g., Fairness versus Social). These contrasts were then analyzed at the second level using a mixed-effect model. To correct for multiple comparisons, images from the second-level analyses were subjected to a voxelwise threshold of $p < 0.001$ (uncorrected) and a cluster extent threshold ensuring $q < 0.01$ (false discovery rate (FDR)-corrected). The conjunction analysis tested the minimum-statistic/conjunction null hypothesis (MS/CN; Nichols et al., 2005). Accordingly, we determined common activations across all seven moral versus social contrasts by creating the intersection of the thresholded statistical maps. Coordinate tables and region labels from statistical parametric maps were retrieved using *AtlasReader* (Notter et al., 2019). Significant activations were projected onto a cortical surface via *Surfplot* (Gale et al., 2021).

The interaction analysis between political orientation (liberals vs. conservative) and moral vignette condition (seven moral categories) proceeded as follows. For each participant, seven differential *t*-contrast images modeling each moral foundation versus social norms were collected from the previous univariate GLM analyses. Participants were then assigned to a liberal or conservative group depending on whether their self-reported political orientation fell below or above the median of the political orientation scale. We then entered the *t*-maps into a 2 x 7 repeated-measures ANOVA including participant, group, and condition as main effect and group × condition as interaction effect. The resulting statistical parametric map for the *F*-test of the interaction effect was then thresholded at $p < .005$, $k > 10$.

Multivariate pattern analysis.

Multivariate pattern classification was performed using the *NLTools* package version 0.4.5 (http://github.com/ljchang/nltools) and *nilearn* package version 0.7.1 (https://github.com/nilearn/nilearn). We first obtained each participant's mean vignette condition activity map by averaging over the corresponding GLM beta maps for the three runs. We then mean-centered these maps across all voxels within each beta map.

Next, we trained a linear Support Vector Machine (SVM) to discriminate between each of the 28 pairwise vignette conditions and used a leave-one-subject-out (LOSO) cross-validation procedure, ensuring that every subject served as both training and testing data (Chang et al., 2015). This allowed us to evaluate how a model trained on 63 participants could classify between two vignette conditions from the left-out participant. To evaluate the accuracy of the SVM, we used forced-choice methods from receiver operating characteristic curves (ROC). Forced-choice accuracy examines the relative expressions of the model between two brain images collected from the same participant and is well suited for fMRI as the input images are unlikely to be on the same scale across individuals. We performed hypothesis tests using a two-tailed independent binomial test for forced-choice classification accuracy.

To evaluate the multivariate response patterns in voxels commonly activated in moral judgment studies, we created a mask from a meta-analytic map associated with the term “moral” from Neurosynth (Yarkoni et al. 2011) and thresholded the mask at $P_{FDR} < 0.01$. Furthermore, we selected a priori regions of interest (ROI) commonly reported in moral neuroscience (Eres et al., 2018) from a parcellation created using a whole-brain parcellation based on meta-analytic functional coactivation of the Neurosynth database (Yarkoni et al., 2011) (parcellation available at https://neurovault.org/images/395092/ and displayed in Supplementary Fig. 2). We then retrained and evaluated the SVM using only features (voxels) from these masks and ROIs.

Representational similarity analysis (RSA).

Representational similarity analysis was performed using the *Python Representational Similarity Analysis toolbox* (https://github.com/rsagroup/rsatoolbox). We computed candidate representational dissimilarity matrices (RDMs) denoting MFT’s predictions about the hierarchical structure of moral foundations as well as behavioral responses for all vignette conditions. We then obtained runwise beta maps for each vignette condition and participant from the previously reported GLM. These beta maps were then used to compute the neural reference RDMs. Neural RDMs were estimated using cross-validated squared Mahalanobis (crossnobis) distances by multiplying pattern estimates for stimuli of the same condition across runs (Walther et al., 2016). We constrained the neural RDMs to the same ROIs described for MVPA as well as Neurosynth’s moral mask. To conduct statistical inference on the RDMs, we computed the whitened cosine similarity between each subject’s neural reference RDM and candidate RDM and averaged the resulting cosine similarities across subjects (Schütt et al., 2021). To perform statistical comparisons and estimate the uncertainty of a model’s performance, we randomly sampled the subjects 2,000 times. In addition to comparing models to each other, we compared models to chance performance and to a noise ceiling (Nili et al., 2014). The noise ceiling provides an estimate of the performance the true (data-generating) model would achieve.

Data availability.

The behavioral data that support the findings of this study are available at The Open Science Framework platform (https://osf.io/dfmu6/). The fMRI data are available from the corresponding author upon reasonable request.

Code availability.

All custom code required to reproduce the results in this paper can be found at https://github.com/medianeuroscience/mft_vignettes.

References

1. Amin, A. B., Bednarczyk, R. A., Ray, C. E., Melchiori, K. J., Graham, J., Huntsinger, J. R., & Omer, S. B. (2017). Association of moral values with vaccine hesitancy. Nature Human Behaviour, *1*(12), 873–880.
2. Amit, E., & Greene, J. D. (2012). You see, the ends don't justify the means: Visual imagery and moral judgment. Psychological Science, *23*(8), 861–868.
3. Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318.
4. Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. Nature Neuroscience, *15*(5), 655–661.
5. Caruso, E. M., & Gino, F. (2011). Blind ethics: Closing one's eyes polarizes moral judgments and discourages dishonest behavior. Cognition, *118*(2), 280–285.
6. Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A., & Wager, T. D. (2015). A sensitive and specific neural signature for picture-induced negative affect. PLoS biology, *13*(6), e1002180.
7. 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.
8. Cui, X., Jeter, C. B., Yang, D., Montague, P. R., & Eagleman, D. M. (2007). Vividness of mental imagery: individual variability can be measured objectively. Vision Research, *47*(4), 474–478.
9. Curry, O. S. (2016). Morality as cooperation: A problem-centred approach. In *The evolution of morality* (pp. 27–51). Springer, Cham.
10. Curry, O. S., Alfano, M., Brandt, M. J., & Pelican, C. (2021). Moral molecules: Morality as a combinatorial system. Review of Philosophy and Psychology, 1–20.
11. Diedrichsen, J., Berlot, E., Mur, M., Schütt, H. H., Shahbazi, M., & Kriegeskorte, N. (2020). Comparing representational geometries using whitened unbiased-distance-matrix similarity. *arXiv preprint arXiv:2007.02789*.
12. Doğruyol, B., Alper, S., & Yilmaz, O. (2019). The five-factor model of the moral foundations theory is stable across WEIRD and non-WEIRD cultures. Personality and Individual Differences, *151*, 109547.
13. Dungan, J., Young, L. (2012). The two-type model of morality. In D. Fassin (ed.) *Companion to Moral Anthropology*. Wiley-Blackwell
14. Eres, R., Louis, W. R., & Molenberghs, P. (2018). Common and distinct neural networks involved in fMRI studies investigating morality: an ALE meta-analysis. Social Neuroscience, *13*(4), 384–398.
15. Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe,A., … & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116.
16. FeldmanHall, O., & Mobbs, D. (2015). A neural network for moral decision making. In AW Toga, & MD Lieberman (eds.*) Brain Mapping: An Encyclopedic Reference. Elsevier: Oxford*.

17. FeldmanHall, O., Mobbs, D., & Dalgleish, T. (2014). Deconstructing the brain's moral network: dissociable functionality between the temporoparietal junction and ventro-medial prefrontal cortex. Social Cognitive and Affective Neuroscience, *9*(3), 297–306.
18. Gale, Daniel J., Vos de Wael., Reinder, Benkarim, Oualid, & Bernhardt, Boris. (2021). *Surfplot: Publication-ready brain surface figures* (v0.1.0). Zenodo.
19. Gantman, A., Devraj-Kizuk, S., Mende-Siedlecki, P., Van Bavel, J. J., & Mathewson, K. E. (2020). The time course of moral perception: an ERP investigation of the moral pop-out effect. Social Cognitive and Affective Neuroscience, *15*(2), 235–246.
20. Gantman, A. P., & Van Bavel, J. J. (2015). Moral perception. Trends in Cognitive Sciences, *19*(11), 631–633.
21. 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.
22. Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (Vol. 47, pp. 55–130). Academic Press.
23. 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.
24. Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? Trends in Cognitive Sciences, *6*(12), 517–523.
25. Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological Review, *108*(4), 814.
26. Haidt, J. (2007). The new synthesis in moral psychology. Science, *316*(5827), 998–1002.
27. 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. The innate mind, *3*, 367–391.
28. 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. doi:10.1371/journal.pone.0042366
29. Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2012). The neural basis of intuitive and counterintuitive moral judgment. Social Cognitive and Affective Neuroscience, *7*(4), 393–402.
30. Kivikangas, J. M., Fernández-Castilla, B., Järvelä, S., Ravaja, N. & Lönnqvist, J.-E. Moral foundations and political orientation: systematic review and meta-analysis. Psychol. Bull. 147, 55–94 (2020).
31. Hoover, J., Atari, M., Mostafazadeh Davani, A., Kennedy, B., Portillo-Wightman, G., Yeh, L., & Dehghani, M. (2021). Investigating the role of group-based morality in extreme behavioral expressions of prejudice. Nature Communications, *12*(1), 1–13.

32. Leong, Y. C., Chen, J., Willer, R., & Zaki, J. (2020). Conservative and liberal attitudes drive polarized neural responses to political content. *Proceedings of the National Academy of Sciences*, *117*(44), 27731–27739.
33. Lewis, G. J., Kanai, R., Bates, T. C., & Rees, G. (2012). Moral values are associated with individual differences in regional brain volume. Journal of Cognitive Neuroscience, *24*(8), 1657–1663.
34. Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. Nature Human Behaviour, 389–396.
35. Morgan, G. S., Skitka, L. J., & Wisneski, D. C. (2010). Moral and religious convictions and intentions to vote in the 2008 presidential election. Analyses of Social Issues and Public Policy, 10(1), 307320.
36. Nash, K., Baumgartner, T., & Knoch, D. (2017). Group-focused morality is associated with limited conflict detection and resolution capacity: Neuroanatomical evidence. Biological Psychology, *123*, 235–240.
37. Nichols, T., Brett, M., Andersson, J., Wager, T., & Poline, J. B. (2005). Valid conjunction inference with the minimum statistic. Neuroimage, *25*(3), 653–660.
38. Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. PLoS computational biology, *10*(4), e1003553.
39. Northoff, G., Heinzel, A., De Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain—a meta-analysis of imaging studies on the self. Neuroimage, *31*(1), 440–457.
40. Notter M. P., Gale D., Herholz P., Markello R. D., Notter-Bielser M.-L., & Whitaker K. (2019). AtlasReader: A Python package to generate coordinate tables, region labels, and informative figures from statistical MRI images. Journal of Open Source Software, *4*(*34*), 1257.
41. O'Craven, K. M., & Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. Journal of cognitive neuroscience, *12*(6), 1013–1023.
42. Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The enigmatic temporal pole: a review of findings on social and emotional processing. Brain, *130*(7), 1718–1731.
43. 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.
44. Schütt, H. H., Kipnis, A. D., Diedrichsen, J., & Kriegeskorte, N. (2021). Statistical inference on representational geometries. *arXiv preprint arXiv:2112.09200*.
45. Shweder, R.A., Much, N. C., Mahapatra, M., & Park, L. (1997). The “big three” of morality (autonomy, community, and divinity), and the “big three” explanations of suffering. In A. Brandt, P. Rozin (Eds.), *Morality and health*, Routledge, New York, NY, 119–169
46. Sinnott-Armstrong, W., & Wheatley, T. (2012). The disunity of morality and why it matters to philosophy. The Monist, *95*(3), 355–377.

47. Sperber, D. (2005). Modularity and relevance: How can a massively modular mind be flexible and context-sensitive? In P. Carruthers, S. Laurence, & S. P. Stich (Eds.), *The innate mind: Structure and contents*: Vol. 1. (pp. 53–68). New York: Oxford University Press
48. Suhler, C. L., & Churchland, P. (2011). Can innate, modular “foundations” explain morality? Challenges for Haidt's moral foundations theory. Journal of Cognitive Neuroscience, *23*(9), 2103–2116.
49. 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.
50. van Baar, J. M., Chang, L. J., & Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. Nature Communications, *10*(1), 1–14.
51. van Baar, J. M., Halpern, D. J., & FeldmanHall, O. (2021). Intolerance of uncertainty modulates brain-to-brain synchrony during politically polarized perception. *Proceedings of the National Academy of Sciences*, *118*(20), e2022491118.
52. Van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. Human Brain Mapping,*30*(3), 829–858.
53. Wager TD, Atlas LY, Lindquist MA, Roy M, Woo CW, Kross E. An fMRI-based neurologic signature of physical pain. The New England journal of medicine. 2013; 368(15):1388–97. doi: 10.1056/NEJMoa1204471 PMID: 23574118
54. Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage, *137*, 188–200.
55. 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.
56. Woo, C. W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: brain models in translational neuroimaging. Nature Neuroscience, *20*(3), 365–377.
57. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. Nature Methods, *8*(8), 665–670.
58. Young, L., & Dungan, J. (2012). Where in the brain is morality? Everywhere and maybe nowhere. Social Neuroscience, *7*(1), 1–10.
59. Young, L., & Saxe, R. (2009). An FMRI investigation of spontaneous mental state inference for moral judgment. Journal of Cognitive Neuroscience, *21*(7), 1396–1405.

Figures

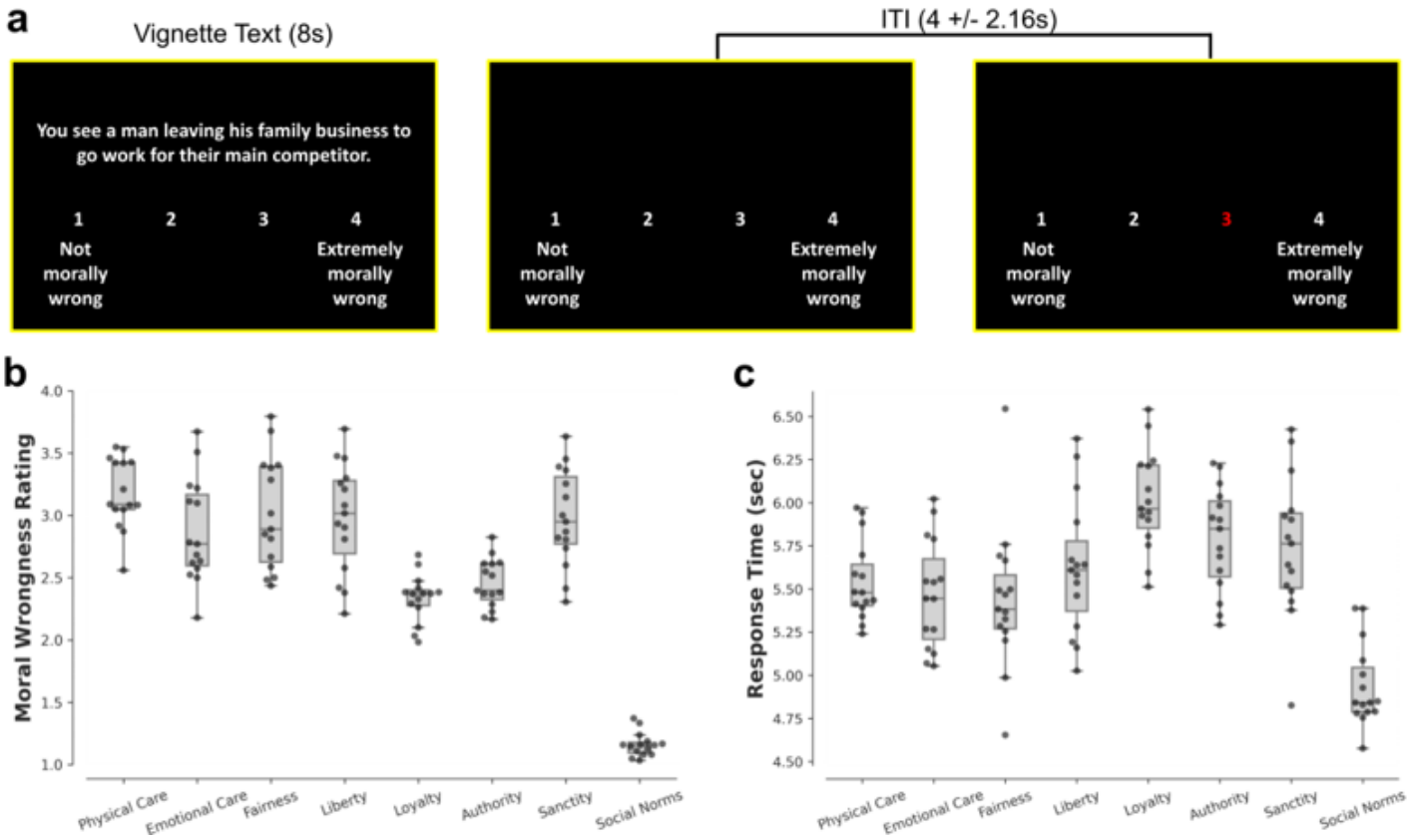

Figure 1

Moral judgment task and behavioral results. (a) In the fMRI study, participants (n = 64) rated the moral wrongness of 120 vignettes (Clifford et al., 2015). Each vignette described a violation of one of seven moral foundations or a social norm transgression. Vignettes were randomly distributed across 3 runs, with 5 vignettes per condition per run. Participants viewed one vignette at a time and were instructed to vividly imagine the described scene. While the vignette was on screen, participants were asked to make a judgment of how morally wrong the action described in the vignette was. The selected rating was highlighted in red. After 8 seconds, the vignette disappeared but the scale remained on screen and participants could still respond during the inter-trial interval (ITI). **(b)** Mean moral wrongness ratings were significantly higher for each moral foundation compared to social norms ($\chi2$ (7) = 202.29, p< 0.001). **(c)** Mean response times (in seconds) until a moral judgment was made were significantly slower for each moral foundation compared to social norms ($\chi2$ (7) = 74.088, p < 0.001). Each dot in (b) and (c) represents the mean response for each vignette.

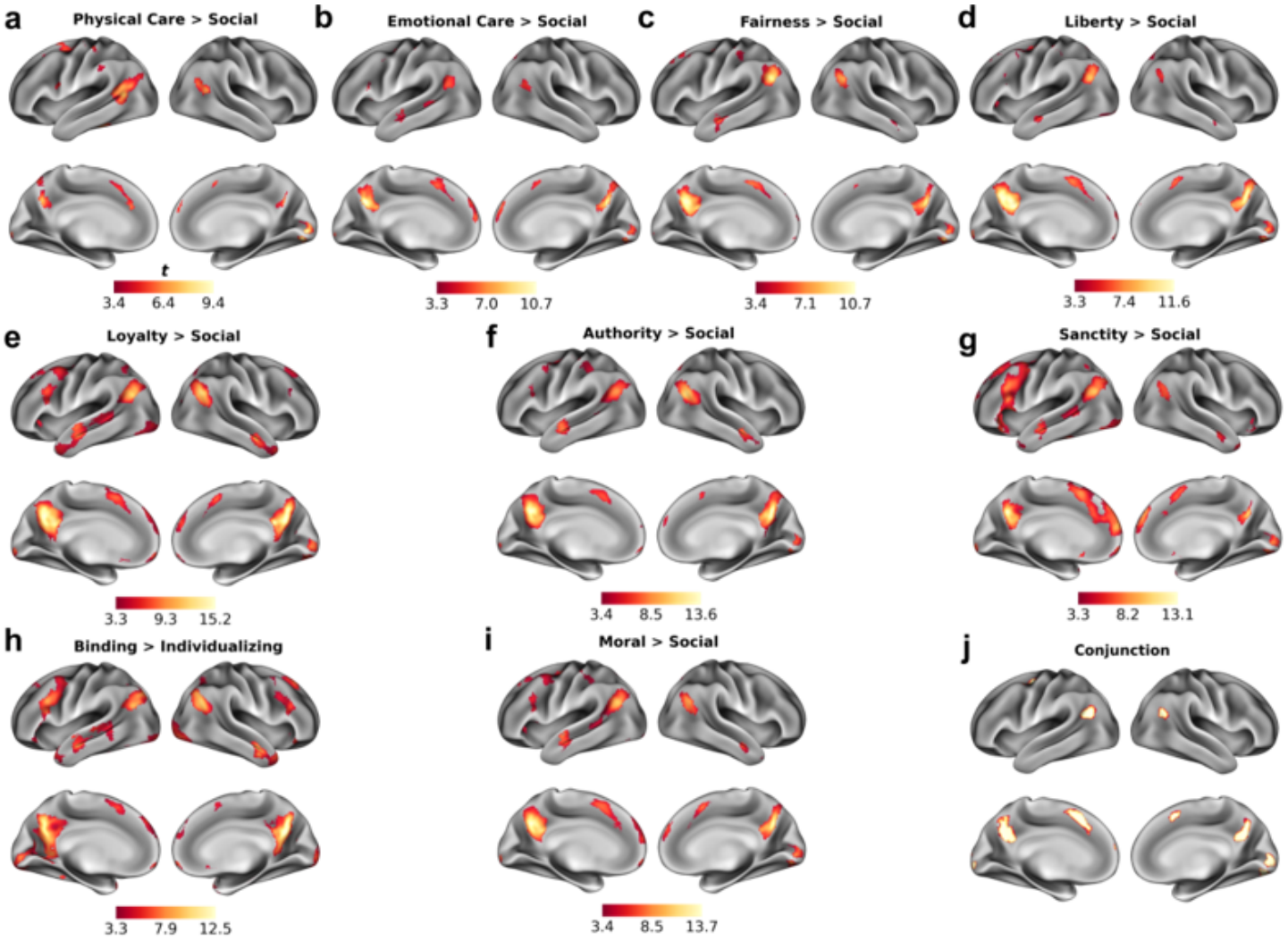

Figure 2

Results of whole-brain univariate analyses. **(a)–(i)** Statistical parametric *t*-maps for group analysis (N = 64) are projected onto a cortical surface for visualization. Contrasts were computed for every moral foundation versus social norms (a–g), for binding versus individualizing foundations (h), and for all averaged moral foundations versus social norms (i). For all images, cluster-level correction (FDR, q < 0.01) was applied. **(j)** A conjunction analysis identified voxels that survived FDR in each of the seven moral foundation versus social norm contrasts (a–g).

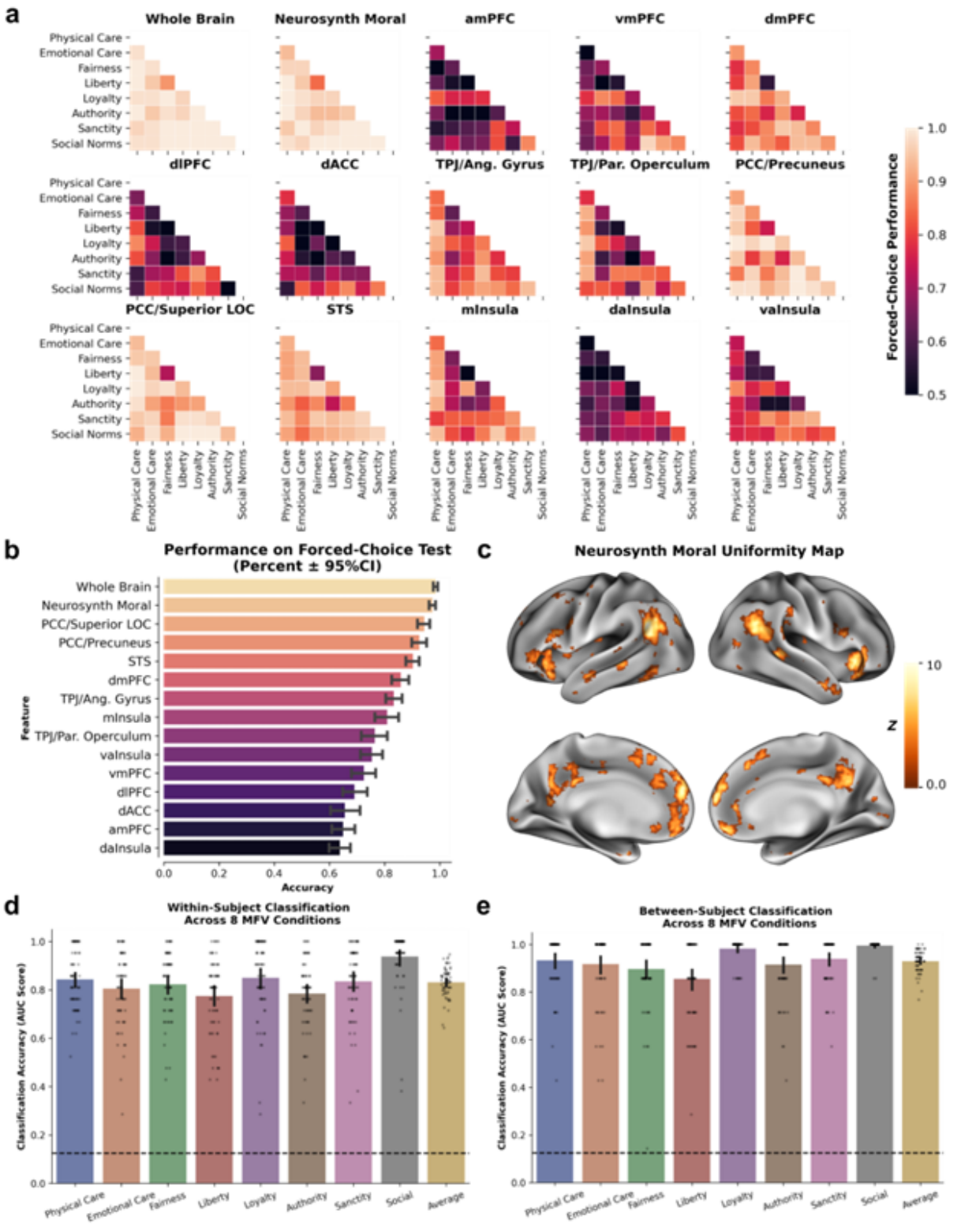

Figure 3

Multivoxel pattern classification of moral foundation vignettes. A linear support vector machine (SVM) was trained via leave-one-subject-out (LOSO) cross-validation to classify which of two vignette conditions an individual was judging. **(a)** Forced-choice accuracy is reported in heatmaps, with bright (dark) colors denoting higher (lower) performance. **(b)** Average performance for forced-choice tests across different feature (voxel) sets. **(c)** Cortical projection of the meta-analytic map associated with the term “moral”

from Neurosynth (uniformity test map, thresholded at $P_{FDR} < 0.01$). **(d)** Within-subject classification accuracy for leave-one-run-out (LORO) multiclass SVM and **(e)** between-subject classification accuracy for LOSO multiclass SVM. Bars show classification accuracy for each vignette condition, averaged across subjects. Dots represent classification accuracy for each participant. Error bars indicate 95% CI of the mean based on 5,000 bootstrap iterations. Dashed line denotes chance level (0.125). amPFC: anterior mPFC; mInsula: mid Insula; daInsula: dorsal anterior Insula; vaInsula: ventral anterior Insula.

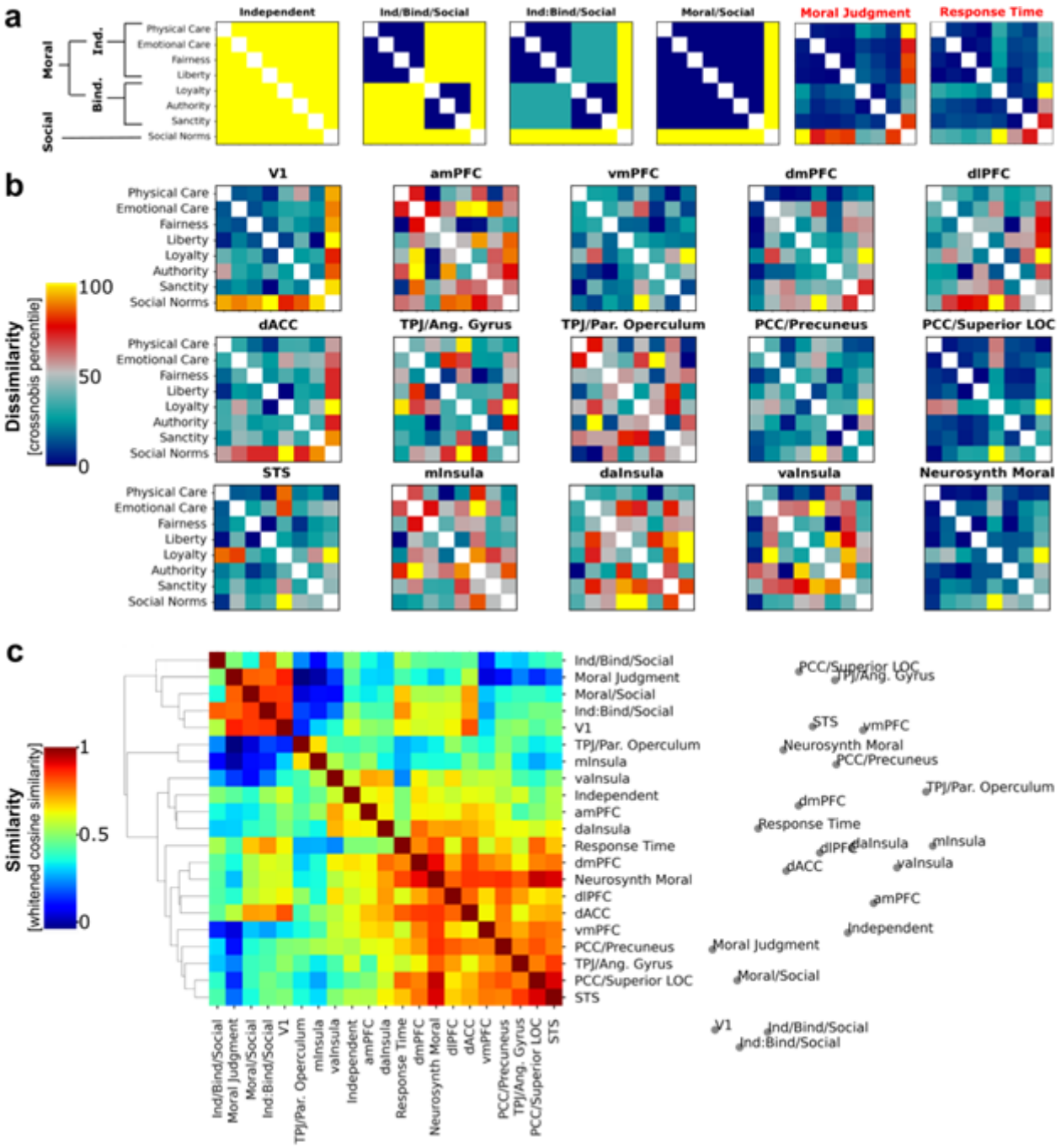

Figure 4

Neural representational mapping of moral foundations. **(a)** Candidate models. Representational dissimilarity matrices (RDM) for categorical models (black) predict vignette condition structures. RDMs

for behavioral models (red) denote moral judgment (wrongness ratings) and response times. **(b)** Reference models. Neural reference RDMs were computed using runwise beta maps for each condition derived from a first-level GLM. All RDMs show group-averaged crossnobis distances. **(c)** Left: Hierarchically-clustered heatmap of RDM similarities for group-averaged RDMs based on whitened cosine similarity. Right: t-distributed Stochastic Neighbor Embedding (t-SNE) of RDM similarities. Each point represents an RDM, and distances between the points approximate the similarities (whitened cosine similarity) among the RDMs.

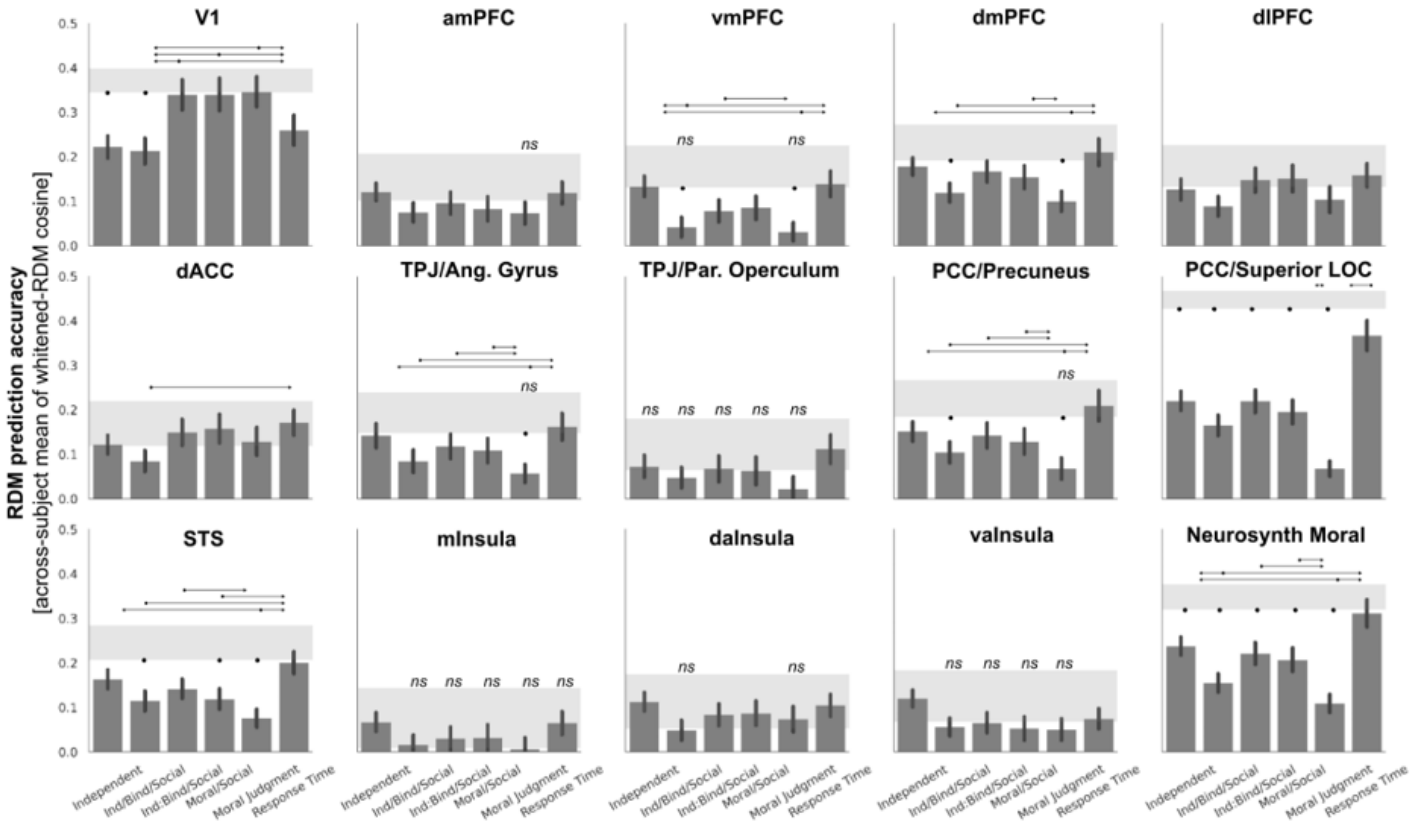

Figure 5

Ability of each candidate RDM to predict reference RDMs. Bars show across-subject means of whitened cosine similarity between each neural reference RDM and candidate model. Gray rectangles represent the noise ceiling, which indicates the expected performance of the true model given the noise in the data. Error bars show the SEM (95% confidence interval over 2,000 bootstrap samples). Pairwise differences are summarized by arrows (FDR $q < 0.01$), indicating that the model marked with the dot performed significantly differently than the model the arrow points at and all models further away in the direction of the arrow. Dots along the noise ceiling mark models that are significantly different from the noise ceiling. Models marked with *ns*were not significantly different from chance performance (Bonferroni-corrected for 6 models).

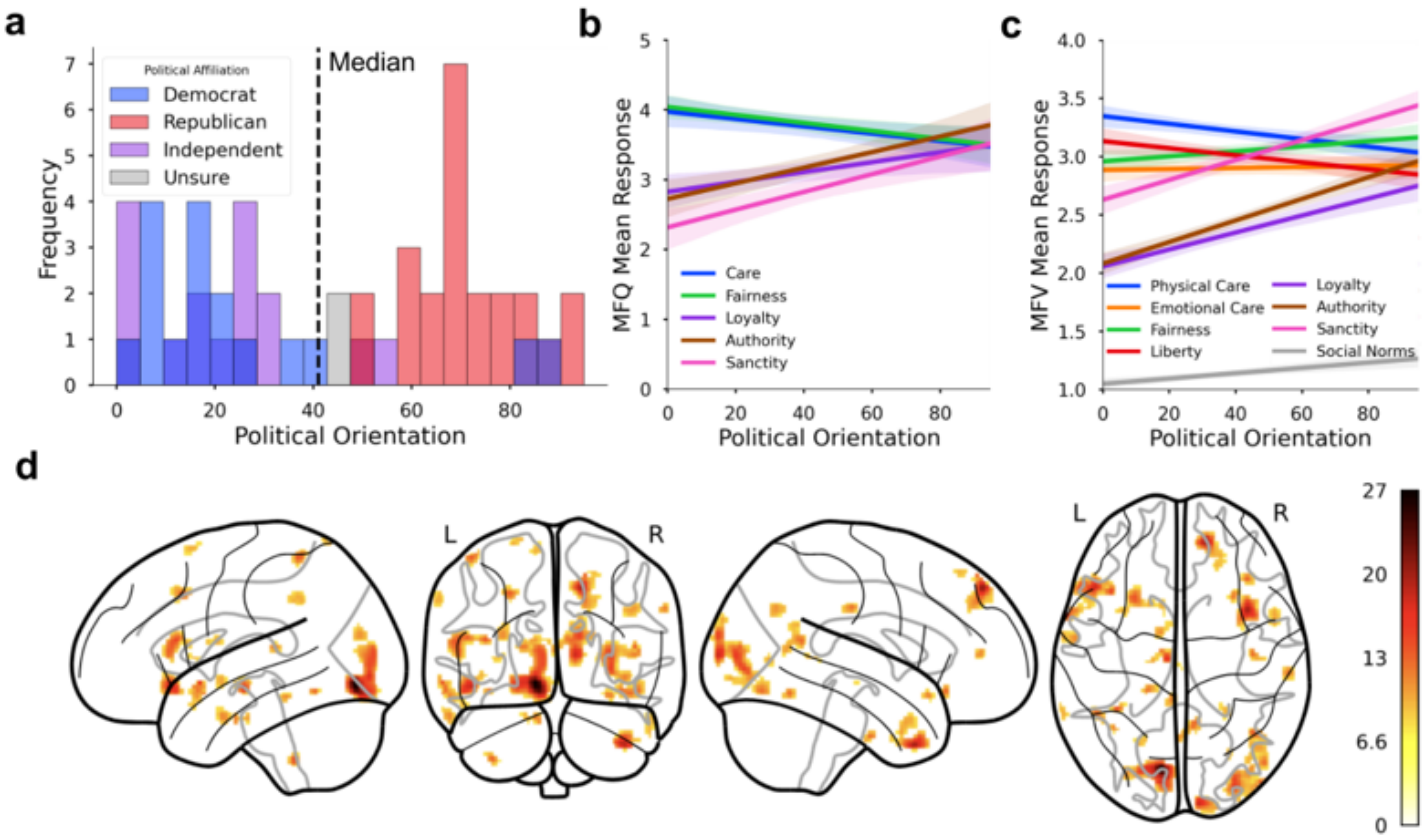

Figure 6

Political ideology shapes processing of moral foundations. (**a**) Sample distribution of self-reported political orientation and affiliation. Black dashed line reflects the median of political orientation (41). Slope estimates for regressing political ideology onto averaged responses to categories of the Moral Foundations Questionnaire (**b**) and Moral Foundation Vignettes (**c**). Shaded areas in **b** and **c** reflect 95% confidence intervals based on 1,000 bootstrap iterations. (**d**). Group × condition interaction map displaying *F*-values thresholded at $p < .005$ (uncorrected), $k > 10$.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementalmaterials.docx