Four-dimensional neural space for moral inference

Jinglu Chen[1,2*], Severi Santavirta[1,2], Vesa Putkinen[1,2,3], Paulo Sérgio Boggio[4,5], Lauri Nummenmaa[1,2,6]
1 Turku PET Centre, Turku University Hospital and University of Turku, Turku, Finland
2 Turku University Hospital, Turku, Finland
3 Turku Institute for Advanced Studies, University of Turku, Finland
4 Social and Cognitive Neuroscience Laboratory, Mackenzie Presbyterian University, São Paulo, Brazil
5 National Institute of Science and Technology on Social and Affective Neuroscience (INCT-SANI), São Paulo, Brazil
6 Department of Psychology, University of Turku, Turku, Finland

*Corresponding author: Jinglu Chen

Postal address: Turku PET Centre, 4-8 Kiinamyllynkatu, 20520 Turku

Contact email: jinglu.chen@utu.fi

Abstract

Intuitive moral inference enables us to evaluate moral situations and judge their rightness or wrongness. Although Moral Foundations Theory provides a framework for understanding moral inference, its underlying neural basis remains unclear. To capture spontaneous neural activity during moral inference, participants were instructed to watch a film rich in moral content without making explicit judgments while undergoing fMRI scanning. Independent participants evaluated the moment-to-moment presence of twenty moral dimensions in the film. Correlation and consensus cluster analyses revealed four independent main moral dimensions: virtue, vice, hierarchy, and rebellion. While each dimension exhibited unique neural activation patterns, the temporoparietal junction and inferior parietal lobe were activated across all types of moral inference. These findings establish the low-dimensional nature for the neural basis of intuitive moral inference in everyday settings.

Introduction

Every day, people engage in complex social interactions that require moral evaluation of their own and others' actions. Should we donate money to charity, should we follow our father's advice, or should we return the wallet we found on the street to the police station? These decisions and moral judgments are driven by quick, automatic intuitions and emotions (Chen et al., 2024; Haidt, 2001). Because there is no objective right or wrong in the nature, cultures have shaped various sets of moral values for promoting their integrity, and the differences in the moral value codes are primarily products of disparate developmental environments (Haidt et al., 1993; Park et al., 2022; van Hemert et al., 2007). These codes are engaged automatically to evaluate the moral value of others' actions, leading to emotional responses supporting subsequent affiliative or rejective behavior (Haidt, 2001). Moral reasoning is an intuitive process guided by abstract moral principles and specific life experiences that individuals use to make decisions regarding right and wrong (Ellemers et al., 2019; Haidt & Joseph, 2004). At the individual level, morality promotes personal growth, as those with more advanced moral reasoning are more likely to maintain cooperative behaviors. At the societal level, moral reasoning motivates individuals to advocate for justice and challenge inequities (Killen & Dahl, 2021; Miranda-Rodríguez et al., 2023).

The Moral Foundations Theory (MFT) postulates six core dimensions for moral evaluation, and these universal "intuitive ethics" shape moral judgments across populations (Haidt & Joseph, 2008). These dimensions include care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, sanctity/degradation, and liberty/oppression. Care vs. harm pertains to actions that nurture and support others' well-being versus those that cause physical, emotional, or psychological damage. Fairness vs. cheating focuses on evaluating whether actions promote justice, equality, and reciprocity. Loyalty vs. betrayal relates to steadfast commitment and dedication to others versus violating or disregarding established trust and obligations. Authority vs. subversion serves as a foundation for evaluating the moral importance of social hierarchy, strong leadership, and tradition. Sanctity vs. degradation describes the moral weighting of purity, sacredness, and good manners. Liberty vs. Oppression describes the ability to act freely and independently. The relative importance of these moral foundations to individuals or societies influences people's moral choices. For instance, individuals with differing political ideologies utilize moral foundations in distinct ways when making moral judgments, which can lead to divergent political views and social orientations (Cornwell & Higgins, 2013; Crowson & DeBacker, 2008; Federico et al., 2013; Graham et al., 2009).

Most neuroimaging studies on moral reasoning have however focused on the neural basis of the moral decision-making process, rather than the neural basis of the moral foundations. The prevalent paradigm involves presenting moral dilemmas and evaluating moral violations or moral decision-making (FeldmanHall et al., 2014; Greene et al., 2001, 2004; Han et al., 2014, 2016; Harenski et al., 2008; Hopp et al., 2023; Parkinson et al., 2011; Reniers et al., 2012; Schneider et al., 2013; Shenhav & Greene, 2014). The pivotal fMRI study on the brain basis of moral reasoning demonstrated that emotion plays a

significant role in moral judgment. The participants were presented with two types of moral dilemmas—those involving personal emotional engagement and those that did not—and non-moral dilemmas during brain scanning. The results revealed that emotionally charged moral reasoning activated the brain's emotion circuit more effectively than the other two dilemmas, in the medial frontal gyrus, posterior cingulate gyrus, and angular gyrus. Conversely, when the dilemmas involved less personal emotion, regions associated with working memory, including the middle frontal gyrus and parietal lobe, were more active (Greene et al., 2001).

Subsequent studies have further investigated moral reasoning under similar conditions, revealing the involvement of a distributed network of cortical and subcortical brain regions. The orbitofrontal cortex is involved in theory of mind and reward learning (Greene & Haidt, 2002; Moll et al., 2002; Yoder & Decety, 2018), while the dorsolateral prefrontal cortex is crucial for cognitive control and analyzing moral situations (Greene et al., 2004; Mendez, 2009; Reniers et al., 2012). The anterior cingulate cortex is activated during moral decision making and inference (FeldmanHall et al., 2012; Han et al., 2016). In the temporal lobe, the temporal pole is activated in response to moral transgressions (Finger et al., 2006), and the superior temporal sulcus is important for emotion cognition and theory of mind (Cáceda et al., 2011; Greene et al., 2004; Harenski et al., 2008; Mendez, 2009; Moll et al., 2002; Prehn et al., 2008). The precuneus plays a role in the moral appraisal processes underlying moral reasoning (Harenski et al., 2008; Reniers et al., 2012; Schneider et al., 2013; Young & Saxe, 2008). Additionally, the thalamus is activated by morally unpleasant pictures (Moll et al., 2002), the striatum is involved in moral value evaluation (Shenhav & Greene, 2010; Yoder & Decety, 2018), and the amygdala participates in various aspects of moral reasoning, especially in affective assessment (Berthoz et al., 2006; FeldmanHall et al., 2012; Mendez, 2009; Moll et al., 2002; Schneider et al., 2013; Shenhav & Greene, 2014; Sommer et al., 2010). The insula is engaged in emotion processing and outcome probability assessment (Greene et al., 2004; Mendez, 2009; Shenhav & Greene, 2010), and the temporoparietal junction is crucial for understanding others' mental states in moral reasoning (FeldmanHall et al., 2014; Finger et al., 2006; Greene et al., 2004; Harenski et al., 2010; Mendez, 2009; Schneider et al., 2013; Sommer et al., 2010; Wasserman et al., 2017; Young & Saxe, 2008).

These fMRI studies demonstrate that neural systems supporting both social cognition and emotions play crucial roles in moral judgments (Greene & Haidt, 2002). However, controlled studies based on artificial moral dilemmas cannot fully capture the natural variation of moral evaluations that occur frequently in everyday life (Wilhelm Hofmann et al., 2015). Because many of the stereotypical situations presented in the moral dilemmas tasks are such that may never be encountered in an average person's life, researchers have begun to suggest that more ecologically valid approaches should be employed in future research (Graham, 2014).

The current study

Here, we investigated the neural basis of moral inference during natural vision. Given the

inherently spontaneous nature of morality, we departed from traditional self-report paradigms that rely on explicit moral reasoning judgments (Khoudary et al., 2022). Instead, we employed an approach that captured participants' haemodynamic activation during spontaneous moral inference. Participants viewed a feature film with morally complex and engaging scenes while undergoing functional magnetic resonance imaging (fMRI). The film was dynamically annotated for the moment-to-moment presence of the core moral dimensions postulated in MFT. Modelling the haemodynamic activation parametrically with the time series of the intensities of moral dimensions allowed us to map the brain basis of spontaneous moral processing during complex and contextually rich social scenarios. We further adopted a data-driven approach to examine whether spontaneous moral inference involves six distinct dimensions or whether it is represented by higher-level concepts. The results indicated that moral dimensions in the data can be reduced to four distinct evaluative main dimensions: Virtue, Vice, Hierarchy, and Rebellion. Each of these main dimensions was associated with overlapping, yet distinct neural activation patterns across large-scale cortical and subcortical networks.

Materials and methods

104 Finnish-speaking volunteers with normal or corrected to normal vision were recruited for the fMRI study. The exclusion criteria were BMI below 20 or above 30, current use of medications affecting the central nervous system, neurological or psychiatric disorders, mood or anxiety disorders, substance abuse, and standard MRI exclusion criteria. Clinically relevant structural brain abnormalities were ruled out by a consultant neuroradiologist based on anatomical MR images. Altogether, eight participants were excluded from the analysis: two due to gradient coil malfunction, two due to abnormalities in structural MRI, and four because of visible motion artefacts in the preprocessed fMRI data. The final sample comprised 96 participants (47 females, mean ± SD age: 31.5±9.36 years, range 20-57). The study was approved by the ethics board of the hospital district of Southwest Finland, and it was conducted according to Good Clinical Practice and the Declaration of Helsinki. An independent sample of 43 healthy Finnish volunteers with normal or corrected-to-normal vision were recruited (35 females, mean ± SD age: 25.3 ± 7.21 years, range 18–53) to evaluate the moment-to-moment presence of moral dimensions from the movie stimuli. All subjects signed a written informed consent and were informed of their right to withdraw at any time without providing a reason. The participants were compensated for their time and travel expenses.

Stimulus

The Finnish feature film Käsky (*Tears of April* in English, 70min11s) was used as the stimulus since it displays an emotionally intense and morally complex story set during the Finnish Civil War (Louhimies, 2008). This kind of natural audio-visual stimulation has previously been established as a powerful tool for mapping various social and emotional functions (Karjalainen et al., 2017; Lahnakoski et al., 2012; Nummenmaa et al., 2021; Santavirta et al., 2023). In the behavioural experiment, the stimulus was presented in

consecutive eight-second-long chunks while the neuroimaging data was acquired in two sessions (32 min 44 s and 37 min 27 s) with a short break in between.

Evaluation of moral and emotional dimensions

The evaluated moral dimensions: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Liberty/Oppression, Authority/Subversion, and Sanctity/Degradation were based on the MFT (Haidt & Joseph, 2008; Iyer et al., 2012). We also incorporated additional dimensions, including Individualism (harm versus benefit to the individual) and Collectivity (harm versus benefit to the group), which serve as overarching features of moral reasoning (Graham et al., 2009). Because moral inference is influenced by basic emotions, we included dimensions for pleasure, displeasure, arousal, and calmness, which capture a substantial proportion of variance in basic emotion ratings (Russell, 1980). Furthermore, we included simple dimensions of "moral righteousness" and "moral injustice" as umbrella terms to capture the rudimentary "gut feeling" of moral evaluations. In total, we collected data for 20 dimensions (**Table 1**).

Dimensions	Definition	Example
Care	Actions are considered caring when they involve helping, protecting, caring for, or showing compassion for another.	A woman tenderly takes care of her dog.
Harm	Actions are considered harm when they involve physical, psychological or emotional harm to another.	A woman drives a car over a squirrel on purpose.
Fairness	Actions are considered just when they maintain fairness, reciprocity and individual rights.	The headmaster ensures that all students have the right to run for president of the student organization.
Cheating	Actions are considered unjust when they result in unfairness, one-sidedness or violate individual rights.	A student copies a classmate's answers in an exam.
Loyalty	Actions express loyalty when they communicate commitment, a sense of duty or loyalty to others.	An employee refuses a high-paying job offer from a competing company because he is committed to his current employer.
Betrayal	Acts express betrayal if they violate or ignore commitments or obligations to others.	An employee jokes with a representative of a competing company about his employer's poor performance last year.
Liberty	Liberty refers to actions that maintain the ability of individuals to act freely and independently.	Spouses respected each other's different religions.
Oppression	Oppression refers to actions that reduce the ability of individuals to act freely and independently.	A man forces his girlfriend to convert to his own religion.
Authority	Authority refers to actions that convey respect or esteem for persons in a leadership or superior position.	A soldier salutes a general who walks past him.

Subversion	Subversion refers to actions that convey disrespect or that tend to undermine the position of the person in charge.	A student repeatedly interrupts the teacher and speaks to him disrespectfully.
Sanctity	Actions express sanctity when they uphold good manners or support a sacred cause.	Even though it's a hot summer day, a man wears a white shirt and a suit for his friend's wedding.
Degradation	Acts are considered degrading when they threaten good manners or defile something sacred.	A drunk old man announces that he is ready to have oral sex with any customer in the bar.
Moral Righteousness	Actions that are good and right are considered moral.	The girl returns the things she borrowed from her roommate at the promised time.
Morally injustice	Actions that are evil or wrong are considered immoral.	A girl steals her roommate's property.
Individualism	An act is harmful to an individual if it targets one person at a time (but not directly to a larger group of people)	A man steals an old woman's handbag.
Collectivity	An act is harmful to a group if it primarily targets a group of people.	A politician reveals his own country's secret security information to an agent of a foreign government.

Table 1. Moral dimensions, as well as their definitions and examples.

In the behavioral experiment, participants were instructed as follows: Your task is to assess whether the film depicts aspects related to the dimensions important for moral evaluation. The definitions and examples for each dimension were also presented to the participants. The movie was split into eight-second clips. Two interleaved splits were created with 4-second difference in starting times, thus yielding 4-second temporal resolution when the rating data were combined across the splits. These splits were presented to different participants to provide mean moral evaluations of the whole movie at a four-second temporal resolution (**Figure S1**). Each participant was randomly assigned to evaluate one of the two sets of interleaved movie clips for 10 of the 20 evaluated dimensions. After viewing a movie segment, participants rated the presence of elements related to the predefined dimensions on a scale from 0 (not at all) to 10 (very much) (See **Figure S3** for the ratings corresponding to the 20 dimensions). The participants evaluated the moral contents of the movie in the laboratory and the ratings were collected using the Gorilla platform (https://gorilla.sc/). The ratings from all participants were then averaged to produce the final moral inference scores. See **Figure 1** for the data collection and analysis procedures (for more details, see section *Behavioral data collection and analysis* in the supplementary materials).

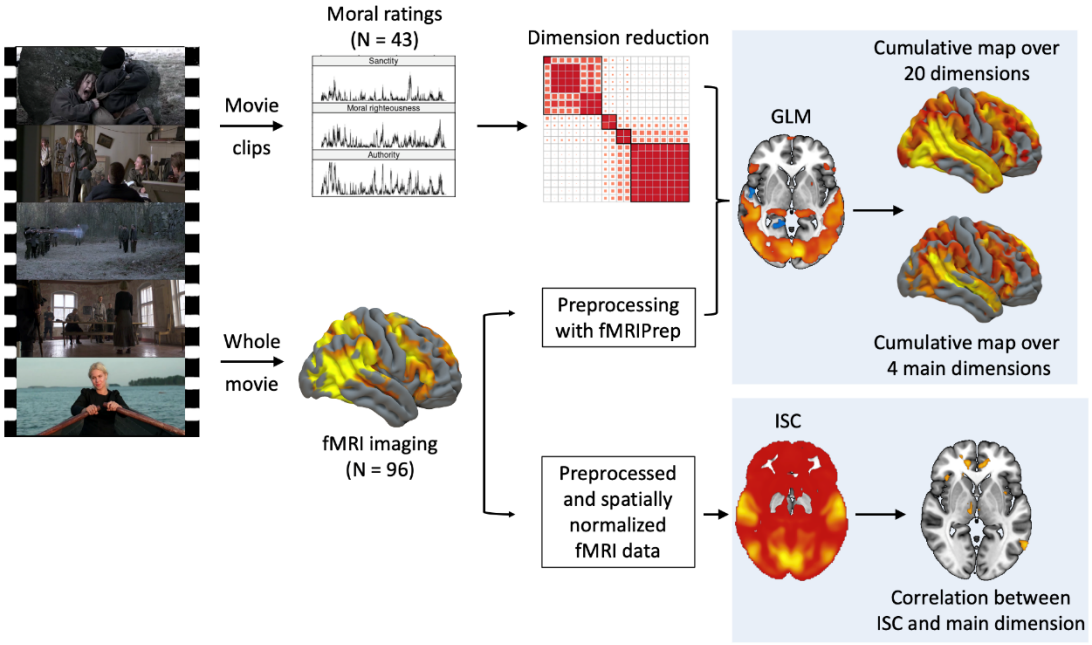

Figure 1. Data acquisition and processing.

Dimension reduction analysis for the moral dimensions

To determine whether the theoretical dimensions of moral inference form independent evaluative dimensions or larger evaluative categories in the movie stimuli, we conducted a consensus clustering analysis. First, a correlation matrix of the mean time series of 20 moral evaluative dimensions was calculated. Temporal correlation quantifies the co-occurrence of dimensions in the movie, and high correlations would suggest shared or overlapping processes when making the evaluations. Consensus clustering analysis was then conducted on the correlation matrix using the DiceR package (Chiu & Talhouk, 2018). Consensus approach ensured stable clustering structure by running 1,000 independent iterations with unique subsamples of the data and over different numbers of clusters (from 2 to 8 clusters).

fMRI acquisition and preprocessing

During the fMRI experiment, the participants were instructed to stay still and watch the movie attentively. The experiment was run using Presentation software (https://www.neurobs.com/). Visual stimuli were presented with NordicNeuroLab VisualSystem binocular display. Sound was conveyed with Sensimetrics S14 insert earphones. Before the functional run, sound intensity was adjusted for each subject so that it could be heard over the gradient noise.

Functional MRI data were acquired with the Phillips Ingenuity TF PET/MR 3-T whole-body scanner at the Turku PET Centre. T1-weighted high-resolution structural images ($1mm^3$ resolution, TR = 9.8 ms, TE = 4.6 ms, flip angle = 7°, FOV = 250 mm, 256 × 256 reconstruction matrix) and T2*-weighted functional images (TR = 2600 ms, TE = 30 ms, flip angle = 75°, FOV = 240 mm, 80 × 80 reconstruction matrix, 62.5 kHz bandwidth in EPT direction, 3.0 mm slice thickness, 45 interleaved axial slices acquired

in ascending order without gaps) were collected and preprocessed with fMRIPrep 1.3.0.post2 (Esteban et al., 2019).

Preprocessing of the T1-weighted images included correction for intensity non-uniformity (INU), distributed with ANTs 2.2.0 and used as a reference later. The images were skull-stripped and reconstructed using recon-all for brain surfaces. The brain mask was refined with a custom variation of the method to reconcile ANTs-derived segmentations of the cortical gray matter of Mindboggle. Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c and brain tissue segmentation were also performed. The following preprocessing was performed on the functional data: slice-time correction was done using 3dTshift, and the data were resampled to fsaverage5. They were then resampled to their native space for head-motion and susceptibility distortions. Spatial smoothing was applied with an isotropic Gaussian kernel of 6mm FWHM. Automatic removal of motion artifacts was performed using ICA-AROMA.

General linear model analysis for the fMRI

FMRI data were analyzed in SPM12 (Wellcome Trust Center for Imaging, London, UK, https://www.fil.ion.ucl.ac.uk/spm/). To map the regions associated with moral inference, we first standardized (0 mean, 1 SD) the evaluative rating time series. Next, we averaged the ratings of moral dimensions within each cluster identified in the dimension reduction analysis to get a cluster-level time series of the prevalence of major moral events in the films. These time series were convolved with canonical HRF and re-standardized before incorporating them into first-level general linear models (GLM). Additional details on the variance inflation factors (VIF) between the convolved cluster time series can be found in the supplementary materials (*Correlations and VIF values for the movie*). In the full-volume analysis, contrast images were first generated separately for each participant to examine the main effects of the four clusters for moral inference. These contrast images were then subjected to a second-level analysis (one-sample t-test) to estimate the population-level effects. Finally, the voxel-wise results were corrected for multiple comparisons using family-wise error (FWE) at the cluster level. Specifically, statistical maps were thresholded at $p_{unc} < 0.001$, identifying the FWEc minimum cluster-size value for FWE correction at the cluster-size level, and then thresholded again at $p_{unc} < 0.001$ and k = FWEc.

In the ROI analysis, 56 bilateral anatomical ROIs were extracted from the AAL2 atlas (Rolls et al., 2015). A one-sample t-test on the mean ß-weights of each ROI was used to assess statistical inference on the group-level. ROI-analysis results were considered significant with $p < 0.05$ (Bonferroni-corrected for multiple comparisons).

Cumulative maps for moral inference

To identify brain regions activated during different types of moral inference, we summarized the results by calculating a cumulative map (binarized sum of significant activations) over all 20 original moral dimensions and four main dimensions separately to reveal a fine-grained gradient in the brain basis of moral inference. Modelling and

statistical thresholding for individual moral dimensions were conducted similarly to the main analysis. Subsequently, two cumulative maps of the binarized BOLD beta maps were computed for original dimensions and four main dimensions, highlighting regions with the most pronounced BOLD responses associated with moral inference.

Moral inference occurs in social situations and is inherently based on socioemotional factors. To identify how the brain responses to moral inference overlap with general social perception networks, we conducted similar cumulative analysis for 44 consistently evaluated social perceptual features that were previously evaluated for the same stimulus (Santavirta et al., 2023). Subsequently, we utilized 56 bilateral anatomical regions of interest (ROIs) from the AAL2 atlas to calculate the average cumulative proportion within each ROI for two moral inference cumulative maps and social perception map. Finally, we computed correlations at the ROI level between the social perception cumulative map and the cumulative maps of two moral inference dimensions.

Finally, we examined whether moral inference synchronizes neural activations across participants. We conducted a supplementary time-window intersubject correlation analysis (ISC) and modelled the time-varying neural synchronization with the four dimensions for moral inference (see supplementary materials: *Inter-subject correlation analysis* for more detail).

Results

Main dimensions of moral evaluation

Consensus clustering analysis identified a four-cluster structure that was both theoretically meaningful and supported by the data-driven clustering for the 20 evaluated moral dimensions. These four clusters were chosen for further analysis of the fMRI data (**Figure 2**). They were labeled as labeled as '**virtue**' (sanctity, care, loyalty, fairness, moral righteousness, liberty, pleasure, calmness), '**hierarchy**' (authority, collectivity), '**rebellion**' (subversion, individualism), and '**vice**' (betrayal, arousal, degradation, oppression, displeasure, moral injustice, harm, cheating). The cluster-level time-series are shown in **Figure 3**. The correlations between the resulting four cluster time series ranged from -0.34 to 0.54, and the VIF values ranged from 1.16 to 2.40. This shows that the regression coefficients for the four main dimensions are stable in the linear model estimations, and that they can be included in the same regression model given their independence.

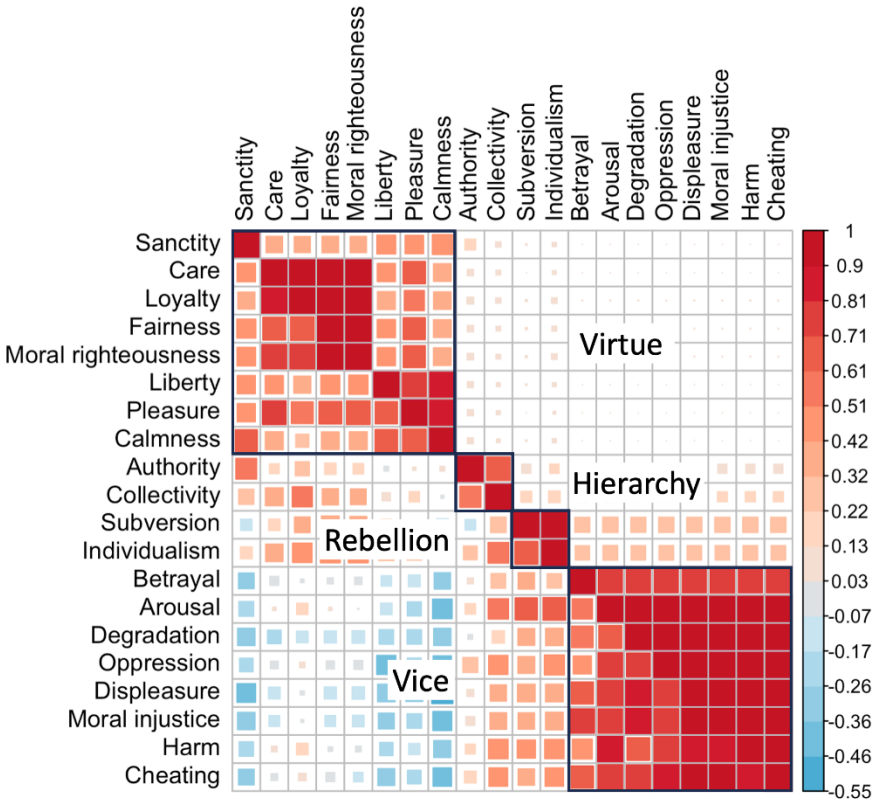

Figure 2. Correlation (lower left) and consensus (upper right) matrix.

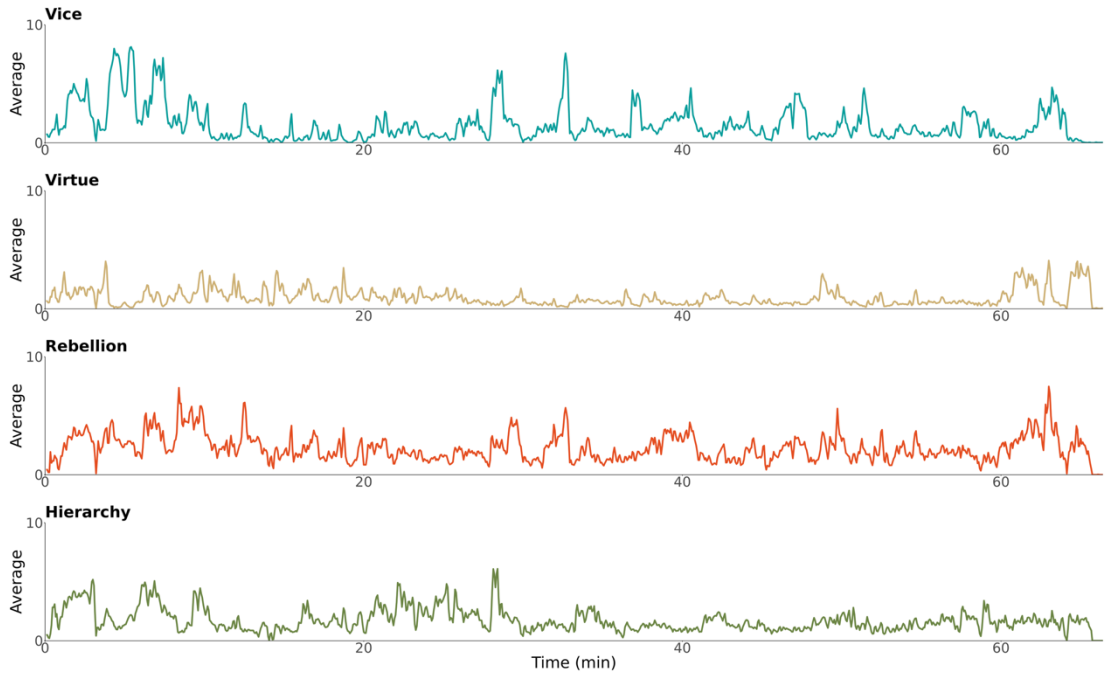

Figure 3. Temporal distribution of mean ratings for four main dimensions throughout the movie.

Cerebral topography of moral inference

The whole-brain general linear model (GLM) revealed distinct neural activation patterns

across four main moral dimensions (**Figure 4**). The **vice** dimension engaged large-scale cortical and subcortical networks, including the medial frontal gyrus, anterior cingulate cortex, precuneus, posterior cingulate cortex, dorsolateral prefrontal cortex, middle frontal gyrus, temporoparietal junction, superior temporal gyrus, orbitofrontal cortex, insula, amygdala, caudata, putamen, thalamus, supramarginal gyrus and angular gyrus. In contrast, the neural activation patterns for **virtue** spanned a smaller area compared to the vice cluster. However, temporal regions, including the superior temporal sulcus and the temporal pole, were responsive to the virtue dimension rather than to the vice dimension. The **rebellion** dimension was associated with both positive and negative neural activations. Positive activation was observed in the medial frontal gyrus, precuneus, posterior cingulate cortex, dorsolateral prefrontal cortex, temporoparietal junction, superior temporal sulcus, superior temporal gyrus, temporal pole, caudate, thalamus, supramarginal gyrus, and angular gyrus. Conversely, the anterior cingulate cortex exhibited negative activation. Finally, the **hierarchy** dimension differed markedly from the other main dimensions, showing primarily negative associations with neural responses. Only few regions, including the amygdala, temporoparietal junction, superior temporal sulcus, superior temporal gyrus, supramarginal gyrus, and angular gyrus showed positive activation.

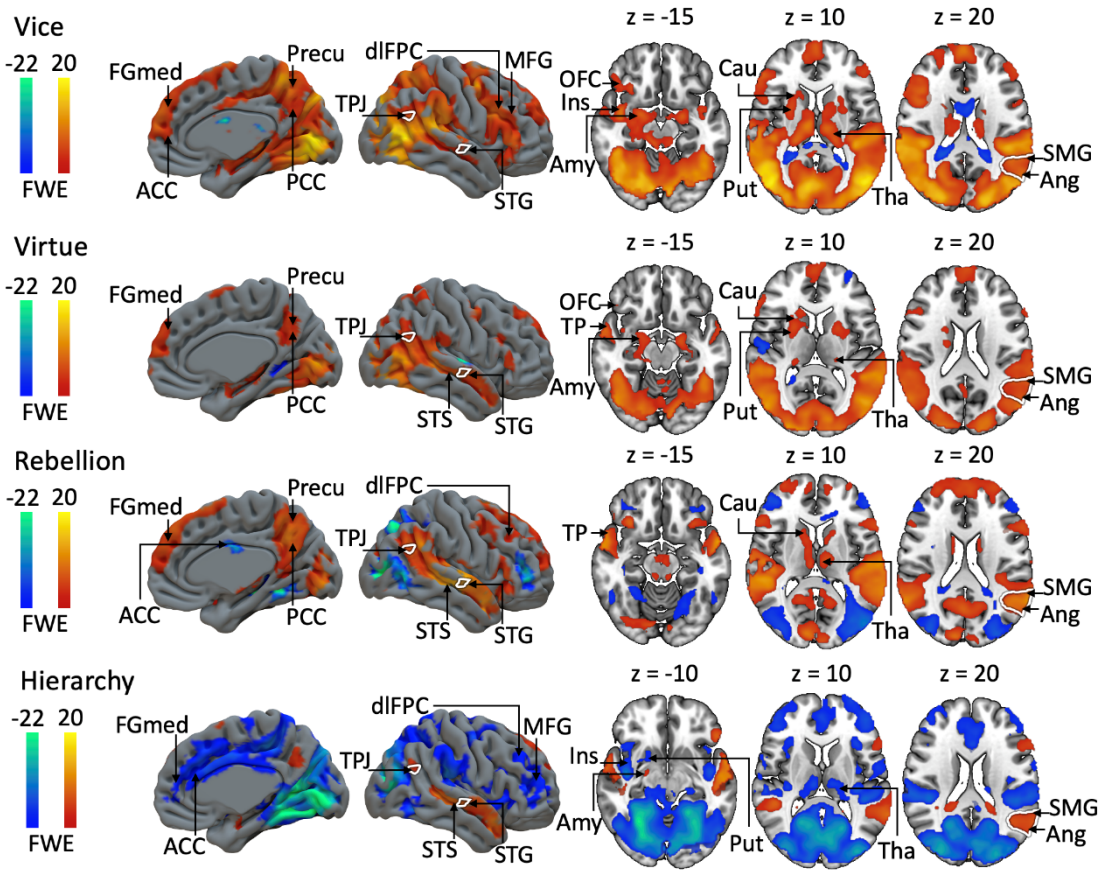

Figure 4. Neural activation patterns associated with the four main dimensions of moral inference (Cluster FWE-corrected at $p < 0.05$. Initial cluster-forming threshold $p < 0.001$). Color bars indicate the t-statistic range. FGmed = medial frontal gyrus, Precu =

precuneus, ACC = anterior cingulate cortex, PCC = posterior cingulate cortex, dlPFC = dorsolateral prefrontal cortex, MFG = middle frontal gyrus, TPJ = temporoparietal junction, STS = superior temporal sulcus, STG = superior temporal gyrus, OFC = orbitofrontal cortex, Ins = insula, Amy = amygdala, TP = temporal pole, Put = putamen, Cau = caudate, Tha = thalamus, SMG = supramarginal gyrus, Ang = angular gyrus.

Results from the ROI analysis supported the full-volume analysis results (**Figure 5**). Overall, the regional effects were mostly positive for vice, virtue, and rebellion, and negative for hierarchy, except for temporal positive associations. Spatially, the associations between moral inference and haemodynamic activity were most consistently observed in the temporal lobe. Frontoparietal activations were most prominent for vice and rebellion dimensions, while occipital responses were consistently positive for vice and virtue, negative for hierarchy, and mostly absent for rebellion. Finally, subcortical activity was most consistently associated with vice.

Figure 5. The heatmap displays the t values for regression coefficients corresponding to each ROI and moral inference cluster. Statistically significant ROIs ($p < 0.05$, Bonferroni corrected independently for each moral cluster) are indicated with an asterisk.

Domain-general responses during moral inference

The cumulative analysis of all 20 original moral dimensions and the four main dimensions revealed extensive subcortical-cortical activation patterns. The occipital lobe and subcortical regions demonstrated the most consistent activations across both calculation methods. For comparison, the cumulative map of social perception also

exhibited widespread activation patterns (**Figure 6**).

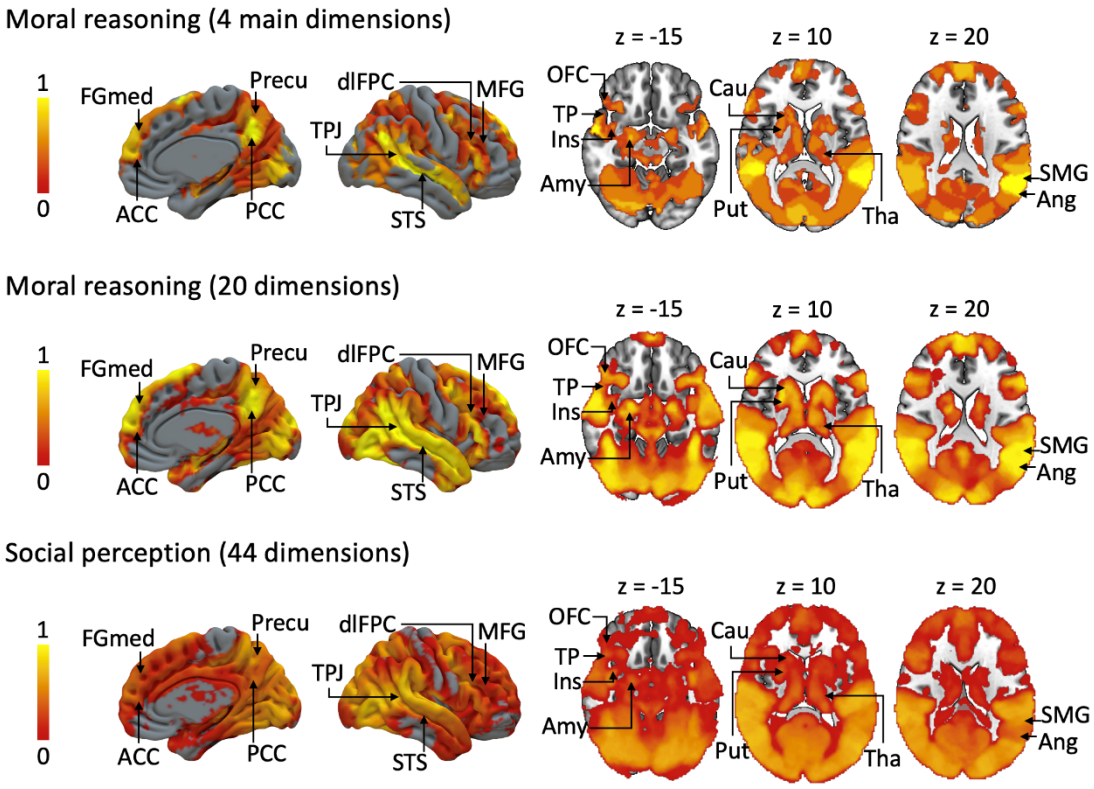

Figure 6. Cumulative activations for moral inference and social perception. The colorbar represents the proportion of dimensions that associated significantly (Cluster FWE-corrected at $p < 0.05$. Initial cluster-forming threshold $p < 0.001$) with the neural responses. FGmed = medial frontal gyrus, Precu = precuneus, ACC = anterior cingulate cortex, PCC = posterior cingulate cortex, dlPFC = dorsolateral prefrontal cortex, MFG = middle frontal gyrus, TPJ = temporoparietal junction, STS = superior temporal sulcus, OFC = orbitofrontal cortex, TP = temporal pole, Ins = insula, Amy = amygdala, Cau = caudate, Put = putamen, Tha = thalamus, SMG = supramarginal gyrus, Ang = angular gyrus.

The correlation analysis between moral inference and social perception revealed substantial overlap in activation patterns between these two processes. The temporal, occipital, and parietal lobes demonstrated the most consistent activation across both moral inference and social perception during movie viewing. Notably, subcortical regions exhibited more consistent activation in moral inference (**Figure 7**).

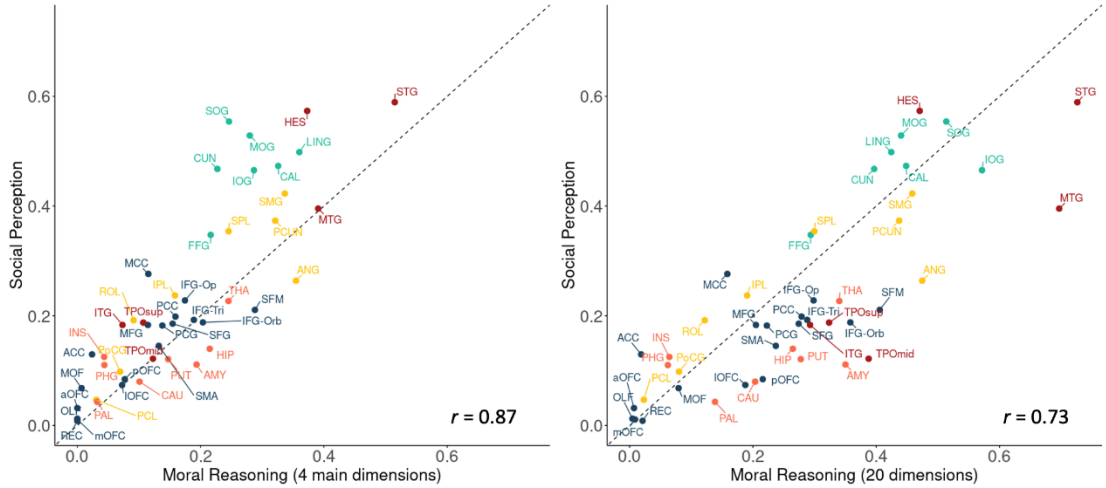

Figure 7. Scatterplot showing the relationship between the proportion of cumulative regional activations for moral inference and social perception. The left panel depicts the correlation using the cumulative map derived from the four main dimensions of moral inference, while the right panel shows the correlation based on all 20 dimensions. Each point represents an anatomical ROI (see **Table S3** for the full name of the regions). The position in relation to the diagonal indicates whether a ROI is more tuned to moral inference (below diagonal) or social perception (above diagonal).

Discussion

Our main finding was that moral inference during natural vision is based on four main dimensions which have both shared and independent neural bases. Moral inference is often a spontaneous and intuitive process that occurs without conscious deliberation (Haidt, 2001), yet many studies require participants to make decisions regarding moral dilemmas, which may not accurately reflect the underlying moral inference process (Graham et al., 2009; Greene et al., 2001; Hopp et al., 2023; Khoudary et al., 2022). Here, we presented participants with a full-length feature film that was previously annotated for the presence of different moral dimensions. This approach identified four main dimensions of moral inference: Virtue, Vice, Hierarchy, and Rebellion. The participants viewed the movie during fMRI without an explicit task, allowing us to quantify the brain basis of spontaneous moral inference in a naturalistic context. Although the brain activation patterns associated with the four main dimensions showed marked differences, they also shared common neural bases, particularly in the TPJ and inferior parietal lobe. These data highlight the low-dimensional nature of moral inference, with distinct but partially overlapping brain basis across dimensions.

Four foundations of moral inference

Moral Foundations Theory posits six psychological "systems" that individuals utilize to guide their moral judgments (Haidt & Joseph, 2008; Iyer et al., 2012). To test whether these dimensions are independently evaluated in the naturalistic movie stimulus, we first conducted correlation and cluster analyses on the behavioural moral ratings. This

revealed that many of the proposed dimensions are interrelated, suggesting shared or overlapping evaluative processes. Consensus clustering revealed four distinct clusters (**Figure 2**). Sanctity, care, loyalty, fairness, and liberty, from MFT, as well as moral righteousness, pleasure, and calmness, were clustered together. These dimensions indicate favourable or desired moral evaluations, and we thus labelled the cluster as **"Virtue".** Authority and collectivity were included in another cluster. Authority refers to actions that demonstrate respect or esteem for individuals in leadership or superior positions, while collectivity refers to actions that may impact the well-being of the group rather than just a single individual. These dimensions primarily emphasize the importance of respecting authority and prioritizing collective interests. Consequently, we designated this cluster as "**Hierarchy**," reflecting its focus on behaviours that promote the collective welfare and social hierarchy.

The third cluster encompassed the dimensions of subversion and individualism. Subversion refers to actions that express disrespect or undermine the authority of individuals in charge. The dimension of individualism pertains to actions that are detrimental to an individual. Together, these dimensions highlight behaviours that are either harmful or disrespectful toward individuals. Therefore, we designated this cluster as "**Rebellion**," reflecting its focus on behaviours that influence primarily the individual rather than the group. The last cluster includes the dimensions of betrayal, degradation, oppression, harm, and cheating, as identified in MFT, along with arousal, displeasure, and moral injustice. In contrast to the virtue cluster, the dimensions within this cluster were associated with negative or unfavourable moral evaluations. We labelled this cluster as "**Vice**," reflecting the negative behaviours identified in moral inference. The variance inflation factor values for these four clusters indicated that they were independent from one another. In contrast to the original Moral Foundations Theory, we dismantled the paired dimensions and considered them as independent constructs that might have involved different evaluative processes.

Virtue and Vice reflected how individuals perceived behaviours as either good or bad, morally righteous or wrong. Reasoning about right and wrong is a fundamental aspect of ethical decision making, which serves as a fundamental feature of human morality and plays a pivotal role in maintaining social order (Ellemers et al., 2019). The dimensions of Virtue and Vice represent possible distinct pathways through which individuals process morally right and wrong information, highlighting the complexity of moral cognition. As social animals, humans naturally engage in cooperation, leading to the emergence of groups (Tomasello, 2014). Given the importance of group dynamics to each individual, it is reasonable to conclude that people were particularly sensitive to behaviours that pertained to themselves or their groups. The Hierarchy and Rebellion main dimensions suggested that individuals might have perceived behaviours differently depending on whether they targeted an individual or a group.

Neural basis of automatic moral inference

Previous imaging research on brain basis of moral reasoning has employed controlled

paradigms where participants make deliberate moral judgments in artificial scenarios pertaining specific moral concerns (FeldmanHall et al., 2014; Greene et al., 2001, 2004; Han et al., 2014, 2016; Harenski et al., 2008; Hopp et al., 2023; Khoudary et al., 2022; Parkinson et al., 2011; Reniers et al., 2012; Schneider et al., 2013; Shenhav & Greene, 2014). In contrast, we investigated the automatic, intuitive processes involved in moral inference using a naturalistic movie viewing where participants were not required to provide explicit answers to moral-related questions during brain imaging. Instead, the potential moral questions emerged naturally from the flow of the action in the cinema (Karjalainen et al., 2017; Lahnakoski et al., 2012; Nummenmaa et al., 2021; Santavirta et al., 2023).

The analysis of both the initial 20 dimensions and 4 higher-order main dimensions revealed a widely distributed network of cortical and subcortical regions involved in encoding the moral behaviours depicted in the movie (**Figure 6**). Consistent activation was observed in the medial frontal gyrus, posterior cingulate cortex, temporoparietal junction, superior temporal sulcus, and amygdala – regions previously identified as key nodes of emotion processing in moral reasoning (FeldmanHall et al., 2012; Greene et al., 2001, 2004; Harenski et al., 2008; Mendez, 2009; Moll et al., 2002; Wasserman et al., 2017). Additionally, the dorsal striatum and inferior parietal lobe exhibited consistent activation, aligning with their established roles in working memory and value evaluation in prior studies (Greene et al., 2001; Han et al., 2016; Mendez, 2009; Reniers et al., 2012; Shenhav & Greene, 2010). These activation patterns resemble the brain's social perception network, and comparison of the presently observed moral judgment networks with previously established network for social perception revealed clear correspondence, as illustrated in **Figure 7**. The more social features a region was responsive to, the more moral dimensions it also responded to. However, there was a clear cortical-subcortical gradient in general-purpose social versus moral processing, in that subcortical regions involved in emotions and motivation exhibited wider tuning profiles for moral versus social perception, highlighting the affective nature of moral information processing and decision making.

The GLM analysis showed that the four main dimensions of moral inference were found to share a common cerebral basis while also exhibiting unique activation patterns. Consistent activation across all dimensions was observed in the temporoparietal junction (TPJ), superior temporal gyrus, supramarginal gyrus, and angular gyrus across these dimensions. The TPJ, a key region associated with Theory of Mind and understanding others' mental states (Saxe & Kanwisher, 2003), showed consistent activation in our study, suggesting its role in encoding and integrating mental state information as a foundational process for further moral judgments (Wasserman et al., 2017; Young & Saxe, 2008). The supramarginal gyrus and angular gyrus, both components of the inferior parietal lobe (IPL), have been shown to be active during moral decision-making tasks (Greene & Haidt, 2002; Han et al., 2016; Mendez, 2009; Reniers et al., 2012). The IPL is recognized as a critical hub underlying diverse basic and social cognitive processes. The consistent activation of the IPL across the four dimensions of moral inference suggests a unified function in processing cognitive and social information essential for moral

inference. Although the superior temporal gyrus was not frequently reported in previous studies, it has consistently been shown to activate during social perception (Santavirta et al., 2023). Compared to prior controlled experiments, the naturalistic paradigm used in the present study may have more effectively elicited activation of the superior temporal gyrus.

There were also notable differences in the activation patterns across the dimensions. The Vice main dimension showed the most pronounced positive activation, particularly in subcortical regions like the thalamus, striatum, and amygdala (**Figures 4 and 5**). The widespread brain activation observed in response to the Vice main dimension suggests a complex interplay between cognitive and emotional processes in the brain's reaction to morally questionable or potentially harmful stimuli. Voxel-wise analysis revealed selective activation of the anterior cingulate cortex (ACC), insula, and middle frontal gyrus in response to vice. The ACC, known for its role in cognitive control (Botvinick et al., 2001), demonstrated positive activation, which may indicate that cognitive control systems are particularly engaged when behaviours are perceived as morally wrong or conflict with internal values. The middle frontal gyrus, previously implicated in working memory during moral reasoning tasks (Greene et al., 2001), likely supports this process by maintaining relevant information for moral evaluation. The insula, a critical hub for integrating cognitive and emotional information (Namkung et al., 2017), showed activation that underscores its integrative role in processing information during the detection of moral violations. Additionally, the insula and anterior cingulate cortex-key nodes of the salience network-were observed (Menon & Uddin, 2010). This activation may indicate that situations involving moral violations are particularly salient, requiring the integration of complex information and heightened attention.

Fewer brain regions were activated in response to Virtue compared to Vice, particularly in subcortical areas, the frontal lobe, and the parietal lobe. This pattern of reduced activation may suggest that the detection of morally right behaviour requires less cognitive and emotional effort compared to the processing of morally questionable or harmful actions. The caudate nucleus and posterior cingulate cortex were activated consistently in response to the Rebellion in both analyses, but not for other main dimensions (**Figure 4** & **Figure 5**). Previous research has shown that these areas exhibit greater activation for personal moral judgments compared to impersonal moral judgments (Greene et al., 2004). In this study, brain activation was notably higher when the film depicted behaviours that caused more harm to individuals. These results suggested that the caudate nucleus and posterior cingulate cortex are involved in detecting immoral behaviours directed at individuals. Responses to the hierarchy dimension exhibited extensive negative activation (**Figure 4** & **Figure 5**). The stimuli, drawn from a war movie featuring disturbing scenes such as shootings, likely evoked feelings of fear among participants. Notably, these disturbing scenarios predominantly involved groups of people. Based on participant ratings, these parts of the movie received relatively higher scores in the authority and collectivity dimensions compared to other segments. The amygdala, which is known to be activated in response to fear-related emotions (Ressler, 2010), showed increased activation when greater harm was inflicted

on the group. In contrast, most other brain regions exhibited deactivation. This pattern may suggest that participants experienced a fear response, leading them to avoid processing behaviours that caused harm to the group.

Limitations

We deliberately wanted to study implicit moral judgments, so a separate sample participated in the rating and fMRI experiment. We do not thus know how well the participant-level moral processing matches with the population level average derived from the fMRI experiment. Despite this, the brain responses to the different dimensions were consistent across participants, indicative of convergence. Finally, although we acquired data from well over one hour of morally evocative scenarios, it is possible that all possible moral dimensions were not comparably present in the stimulus. However, the time series of the main moral dimensions revealed significant variation in all original as well as higher-order dimensions.

Conclusion

We conclude that a low-dimensional space effectively captures the subjective and neural-level variation in moral inference. Contrary to the Moral Foundations Theory, our findings revealed four distinct foundations that appear to process in parallel within the human brain: Virtue, Vice, Hierarchy, and Rebellion. While each of these main dimensions demonstrated unique neural signatures, they also shared common neural substrates in the TPJ and inferior parietal lobe. Our data suggest that automatic moral inference involves the detection of morally right and wrong information, as well as behaviours directed toward individuals or groups. These findings establish the low-dimensional nature for the neural basis of intuitive moral inference in everyday settings.

Acknowledgements

This work was supported by Aatos Erkon Säätiö, China Scholarship Council (202106040042), Alfred Kordelin Foundation, and Research Council of Finland (grant #350416).

Supplementary materials

Behavioral data collection and analysis

To get compact moral emotion intensity fluctuation during the whole movie, and decrease the fragmenting feeling of watching movie, we created two sets of movie clips. The first set began at the start of the film, with the first clip lasting 4 seconds and all subsequent clips lasting 8 seconds. The second set also started at the beginning but consisted entirely of 8-second clips. As shown in **Figure S1**, groups 1 and 2 viewed the first set of clips, while groups 3 and 4 viewed the second set. There are 10 raters in group1, 12 raters in group 2, 10 raters in group3, and 11 raters in group4.

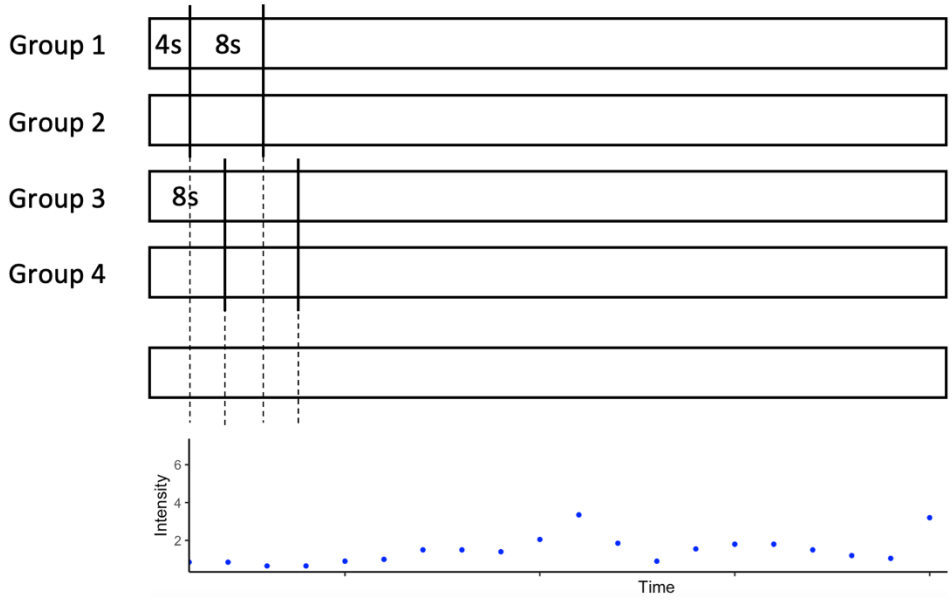

Figure S1. Videos clips viewing in the behavioural experiment. In the behavioral experiment, video clips were segmented differently for participant groups. Groups 1 and 2 initially viewed a 4-second clip, followed by 8-second clips for the remainder of the experiment. In contrast, Groups 3 and 4 consistently viewed 8-second clips throughout. Groups 1 and 3 rated the same set of 10 moral inference dimensions, while Groups 2 and 4 rated the remaining 10 dimensions. Final ratings for each dimension were calculated as the average of all participant ratings. By combining ratings across different time points, we generated ratings for each 4-second time window across all 20 dimensions.

The 20 dimensions of moral inference were divided into two subsets, with participants in each group rating only 10 dimensions per set. In **Figure S1**, groups 1 and 3 rated the first set of moral inference dimensions, while groups 2 and 4 rated the second set. Refer to **Figure S2** for the intraclass correlation coefficients (ICC) within each rating group.

To achieve more accurate estimates of moral emotion intensity that closely reflect participants' perceptions, we calculated the mean ratings from both relevant groups. For instance, for the first 12 seconds of the film, groups 1 and 3 provided three data points for emotion intensity across 10 dimensions, while groups 2 and 4 contributed three data points for the remaining 10 dimensions. This combined data provided a comprehensive measure of moral emotion intensity across all dimensions (see **Figure S1**).

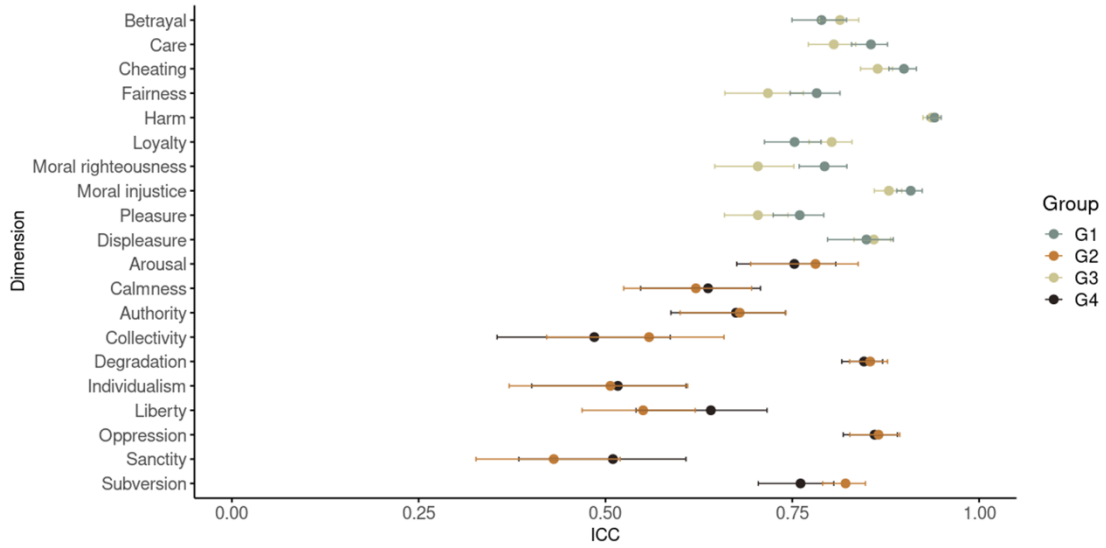

Figure S2. Intraclass correlation coefficients (ICC) for each rating group across all dimensions. Each point represents the ICC value, with the line on the left indicating the lower confidence interval and the line on the right indicating the upper confidence interval. The ICC was calculated using the R package psych. ICC(2k) was selected as the appropriate model because all raters within each group evaluated all video clips.

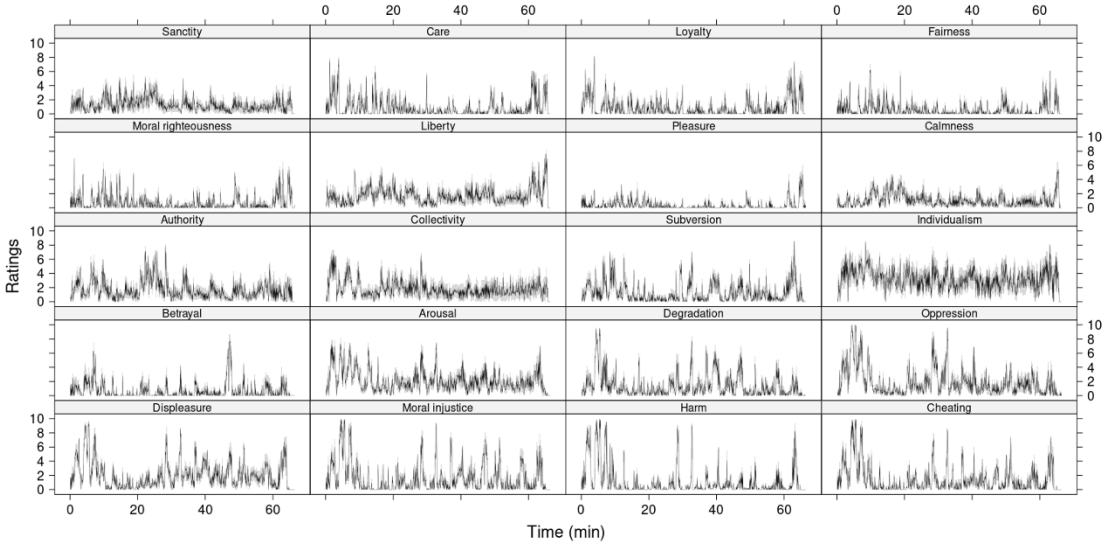

Figure S3. Moral inference ratings for 20 dimensions. The black line represents the mean ratings across participants, while the gray area indicates the standard error.

Correlations and VIF values for the movie

Table S1 presents the correlation matrices for each cluster during the first and second halves of the movie. The upper-right section of the table corresponds to the first half, while the lower-left section represents the second half.

	Virtue	Hierarchy	Rebellion	Vice
Virtue	1	0.2385	0.1650	-0.3428
Hierarchy	0.3608	1	0.2628	0.3552
Rebellion	0.4695	0.1873	1	0.5374
Vice	-0.3189	-0.0506	0.3917	1

Table S1. Correlation matrix of the clusters across the movie.

Table S2 provides the variance inflation factor (VIF) values for all clusters across the first and second halves of the movie.

	Virtue	Hierarchy	Rebellion	Vice
First	1.6971	1.3926	1.7663	2.3889
Second	2.2783	1.1571	2.2091	1.9265

Table S2. VIF values of the clusters across the movie.

Inter-subject correlation analysis

Viewing movies featuring social interactions has been shown to synchronize brain activity across individuals (Hasson et al., 2004). However, how neural synchronization relates with moral inference remains unclear. To investigate this, we calculated the inter-subject correlation (ISC) of neural activation patterns using the ISC-toolbox (Kauppi et al., 2014). Dynamic ISC was calculated in consecutive time windows (window length: 10 TR = 26 seconds) to allow correlating time-window ISC with moral evaluations of the movie stimuli.

Neural synchronization during moral inference

Whole-brain ISC analysis revealed extensive neural synchronization across participants during movie viewing (**Figure S4 & S5**).

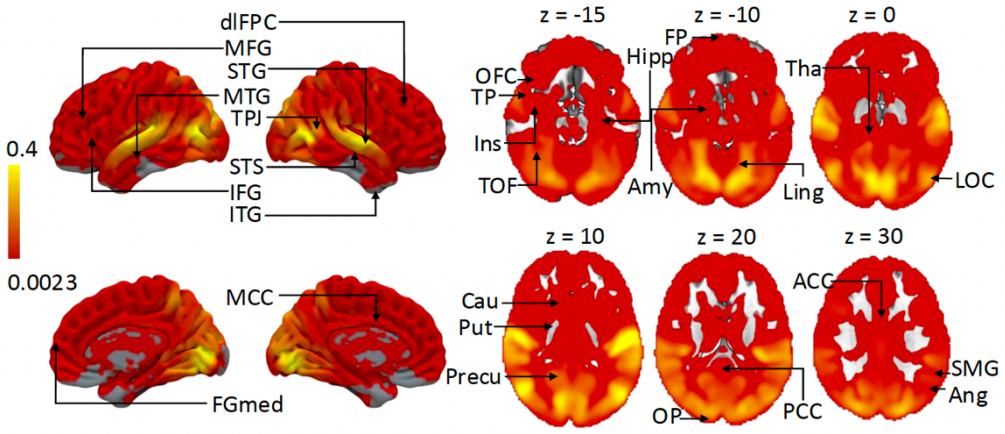

Figure S4. Significant ISC (FDR-corrected, q = 0.001) across subjects over the first half

of the movie. dlPFC = dorsolateral prefrontal cortex, MFG = middle frontal gyrus, STG = superior temporal gyrus, MTG = middle temporal gyrus, TPJ = temporoparietal junction, STS = superior temporal sulcus, IFG = inferior frontal gyrus, ITG = inferior temporal gyrus, FGmed = medial frontal gyrus, MCC = middle cingulate cortex, OFC = orbitofrontal cortex, TP = temporal pole, Ins = insula, TOF = temporal occipital fusiform cortex, Hipp = hippocampus, FP = frontal pole, Amy = amygdala, Ling = lingual gyrus, Tha = thalamus, LOC = lateral occipital cortex, Cau = caudate, Put = putamen, Precu = precuneus, OP = occipital pole, PCC = posterior cingulate cortex, ACC = anterior cingulate cortex, SMG = supramarginal gyrus, Ang = angular gyrus.

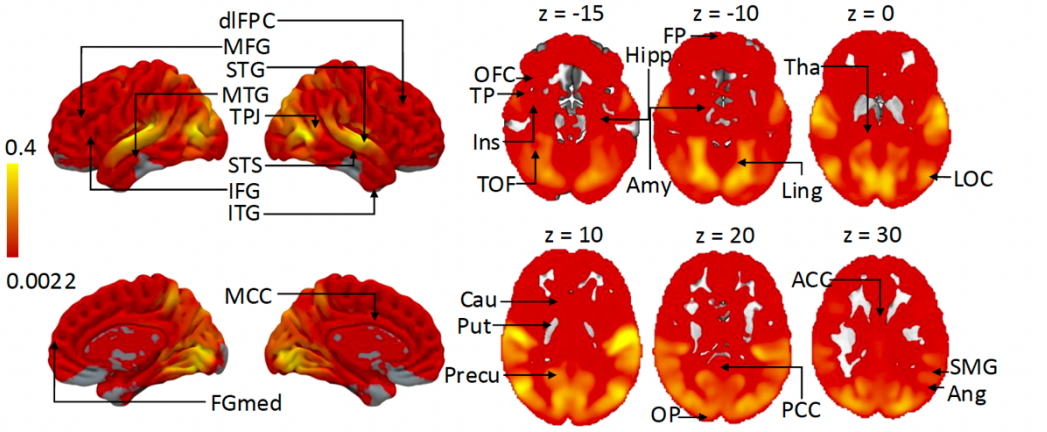

Figure S5. Significant ISC (FDR-corrected, $q = 0.001$) across subjects over the second half of the movie. dlPFC = dorsolateral prefrontal cortex, MFG = middle frontal gyrus, = superior temporal gyrus, MTG = middle temporal gyrus, TPJ = temporoparietal junction, STS = superior temporal sulcus, IFG = inferior frontal gyrus, ITG = inferior temporal gyrus, FGmed = medial frontal gyrus, MCC = middle cingulate cortex, OFC = orbitofrontal cortex, TP = temporal pole, Ins = insula, TOF = temporal occipital fusiform cortex, Hipp = hippocampus, FP = frontal pole, Amy = amygdala, Ling = lingual gyrus, Tha = thalamus, LOC = lateral occipital cortex, Cau = caudate, Put = putamen, Precu = precuneus, OP = occipital pole, PCC = posterior cingulate cortex, ACC = anterior cingulate cortex, SMG = supramarginal gyrus, Ang = angular gyrus.

Figure S6 illustrates the brain regions where synchronization among participants correlates with the Vice and Rebellion main dimensions during the movie.

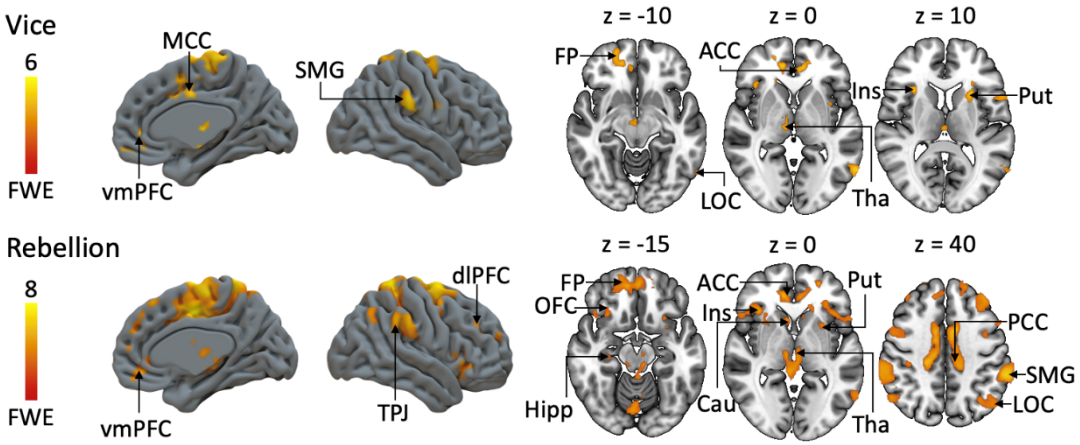

Figure S6. The significant associations between dynamic ISC and moral clusters (Cluster FWE-corrected at $p < 0.05$. Initial cluster-forming threshold $p < 0.001$). MCC = middle cingulate cortex, vmPFC = ventral medial prefrontal cortex, SMG = supramarginal gyrus, FP = frontal pole, LOC = lateral occipital cortex, ACC = anterior cingulate cortex, Tha = thalamus, Ins = insula, Put = putamen, dlPFC = dorsolateral prefrontal cortex, TPJ = temporoparietal junction, OFC = orbitofrontal cortex, Hipp = hippocampus, Cau = caudate, PCC = posterior cingulate cortex.

Abbreviation	Full name
PCG	Precentral Gyrus
SFG	Superior Frontal Gyrus
MFG	Middle Frontal Gyrus
IFG-Op	Inferior Frontal Gyrus (Opercular Part)
IFG-Tri	Inferior Frontal Gyrus (Triangular Part)
IFG-Orb	Inferior Frontal Gyrus (Orbital Part)
ROL	Rolandic Operculum
SMA	Supplementary Motor Area
OLF	Olfactory Cortex
SFM	Superior Medial Frontal Cortex
MOF	Middle Orbital Frontal Cortex
REC	Rectus Gyrus
mOFC	Medial Orbitofrontal Cortex
aOFC	Anterior Orbitofrontal Cortex
pOFC	Posterior Orbitofrontal Cortex
lOFC	Lateral Orbitofrontal Cortex
INS	Insular Cortex
ACC	Anterior Cingulate Cortex
MCC	Middle Cingulate Cortex
PCC	Posterior Cingulate Cortex
HIP	Hippocampus
PHG	Parahippocampal Gyrus
AMY	Amygdala
CAL	Calcarine
CUN	Cuneus
LING	Lingual Gyrus

SOG	Superior Occipital Gyrus
MOG	Middle Occipital Gyrus
IOG	Inferior Occipital Gyrus
FFG	Fusiform Gyrus
PoCG	Postcentral Gyrus
SPL	Superior Parietal Lobule
IPL	Inferior Parietal Lobule
SMG	Supramarginal Gyrus
ANG	Angular Gyrus
PCUN	Precuneus
PCL	Paracentral Lobule
CAU	Caudate Nucleus
PUT	Putamen
PAL	Pallidum
THA	Thalamus
HES	Heschl
STG	Superior Temporal Gyrus
TPOsup	Temporal Pole (Superior)
MTG	Middle Temporal Gyrus
TPOmid	Temporal Pole (Middle)
ITG	Inferior Temporal Gyrus

Table S3. Brain regions presented in Figure 7.

Reference

Berthoz, S., Grèzes, J., Armony, J. L., Passingham, R. E., & Dolan, R. J. (2006). Affective response to one's own moral violations. *NeuroImage*, *31*(2), 945–950. https://doi.org/10.1016/j.neuroimage.2005.12.039

Botvinick, M. M., Carter, C. S., Braver, T. S., Barch, D. M., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10.1037/0033-295X.108.3.624

Cáceda, R., James, G. A., Ely, T. D., Snarey, J., & Kilts, C. D. (2011). Mode of effective connectivity within a putative neural network differentiates moral cognitions related to care and justice ethics. *PLoS ONE*, *6*(2). https://doi.org/10.1371/journal.pone.0014730

Chen, J., Putkinen, V., Seppälä, K., Hirvonen, J., Ioumpa, K., Gazzola, V., Keysers, C., & Nummenmaa, L. (2024). Endogenous opioid receptor system mediates costly altruism in the human brain. *Communications Biology*, *7*(1), 1401. https://doi.org/10.1038/s42003-024-07084-7

Chiu, D. S., & Talhouk, A. (2018). DiceR: An R package for class discovery using an ensemble driven approach. *BMC Bioinformatics*, *19*(1), 17–20. https://doi.org/10.1186/s12859-017-1996-y

Cornwell, J. F. M., & Higgins, E. T. (2013). Morality and Its Relation to Political Ideology: The Role of Promotion and Prevention Concerns. *Personality and Social Psychology Bulletin*, *39*(9), 1164–1172. https://doi.org/10.1177/0146167213489036

Crowson, H. M., & DeBacker, T. K. (2008). Political identification and the defining issues test: Reevaluating an old hypothesis. *Journal of Social Psychology*, *148*(1), 43–60. https://doi.org/10.3200/SOCP.148.1.43-60

Ellemers, N., van der Toorn, J., Paunov, Y., & van Leeuwen, T. (2019). The Psychology of Morality: A Review and Analysis of Empirical Studies Published From 1940 Through 2017. *Personality and Social Psychology Review*, *23*(4), 332–366. https://doi.org/10.1177/1088868318811759

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Federico, C. M., Weber, C. R., Ergun, D., & Hunt, C. (2013). Mapping the connections between politics and morality: The multiple sociopolitical orientations involved in moral intuition. *Political Psychology*, *34*(4), 589–610. https://doi.org/10.1111/pops.12006

FeldmanHall, O., Dalgleish, T., Thompson, R., Evans, D., Schweizer, S., & Mobbs, D. (2012). Differential neural circuitry and self-interest in real vs hypothetical moral decisions. *Social Cognitive and Affective Neuroscience*, *7*(7), 743–751.

https://doi.org/10.1093/scan/nss069

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. https://doi.org/10.1093/scan/nss139

Finger, E. C., Marsh, A. A., Kamel, N., Mitchell, D. G. V., & Blair, J. R. (2006). Caught in the act: The impact of audience on the neural response to morally and socially inappropriate behavior. *NeuroImage*, *33*(1), 414–421. https://doi.org/10.1016/j.neuroimage.2006.06.011

Graham, J. (2014). Morality beyond the lab. *Science*, *345*(6202), 1242–1242. https://doi.org/10.1126/science.1259500

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

Greene, J. D., & Haidt, J. (2002). How (and where) does moral. *Trends in Cognitive Sciences*, *6*(12), 517–523. http://tics.trends.com1364-6613/02/$-seefrontmatter

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389–400. https://doi.org/10.1016/j.neuron.2004.09.027

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4), 55–65. https://doi.org/10.1162/0011526042365555

Haidt, J., & Joseph, C. (2008). 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*(January). https://doi.org/10.1093/acprof:oso/9780195332834.003.0019

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog? *Journal of Personality and Social Psychology*, *65*(4), 613–628. https://doi.org/10.1037/0022-3514.65.4.613

Han, H., Chen, J., Jeong, C., & Glover, G. H. (2016). Influence of the cortical midline structures on moral emotion and motivation in moral decision-making. *Behavioural Brain Research*, *302*, 237–251. https://doi.org/10.1016/j.bbr.2016.01.001

Han, H., Glover, G. H., & Jeong, C. (2014). Cultural influences on the neural correlate of moral decision making processes. *Behavioural Brain Research*, *259*, 215–228. https://doi.org/10.1016/j.bbr.2013.11.012

Harenski, C. L., Antonenko, O., Shane, M. S., & Kiehl, K. A. (2008). Gender differences in neural mechanisms underlying moral sensitivity. *Social Cognitive and Affective Neuroscience*, *3*(4), 313–321. https://doi.org/10.1093/scan/nsn026

Harenski, C. L., Antonenko, O., Shane, M. S., & Kiehl, K. A. (2010). A functional imaging investigation of moral deliberation and moral intuition. *NeuroImage*, *49*(3), 2707–2716. https://doi.org/10.1016/j.neuroimage.2009.10.062

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject Synchronization of Cortical Activity during Natural Vision. *Science*, *303*(5664), 1634–1640. https://doi.org/10.1126/science.1089506

Hopp, F. R., Amir, O., Fisher, J. T., Grafton, S., Sinnott-Armstrong, W., & Weber, R. (2023). Moral foundations elicit shared and dissociable cortical activation modulated by political ideology. *Nature Human Behaviour*, *7*(12), 2182–2198. https://doi.org/10.1038/s41562-023-01693-8

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). https://doi.org/10.1371/journal.pone.0042366

Karjalainen, T., Karlsson, H. K., Lahnakoski, J. M., Glerean, E., Nuutila, P., Jääskeläinen, I. P., Hari, R., Sams, M., & Nummenmaa, L. (2017). Dissociable roles of cerebral μ-opioid and type 2 dopamine receptors in vicarious pain: A combined PET-fMRI study. *Cerebral Cortex*, *27*(8), 4257–4266. https://doi.org/10.1093/cercor/bhx129

Kauppi, J. P., Pajula, J., & Tohka, J. (2014). A versatile software package for inter-subject correlation based analyses of fMRI. *Frontiers in Neuroinformatics*, *8*(JAN), 1–13. https://doi.org/10.3389/fninf.2014.00002

Khoudary, A., Hanna, E., O'Neill, K., Iyengar, V., Clifford, S., Cabeza, R., De Brigard, F., & Sinnott-Armstrong, W. (2022). A functional neuroimaging investigation of Moral Foundations Theory. *Social Neuroscience*, *17*(6), 491–507. https://doi.org/10.1080/17470919.2022.2148737

Killen, M., & Dahl, A. (2021). Moral Reasoning Enables Developmental and Societal Change. *Perspectives on Psychological Science*, *16*(6), 1209–1225. https://doi.org/10.1177/1745691620964076

Lahnakoski, J. M., Glerean, E., Salmi, J., Jääskeläinen, I. P., Sams, M., Hari, R., & Nummenmaa, L. (2012). Naturalistic fMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. *Frontiers in Human Neuroscience*, *6*(AUGUST), 1–14. https://doi.org/10.3389/fnhum.2012.00233

Louhimies, A. (2008). *Käsky*. Helsinki-filmi.

Mendez, M. F. (2009). The neurobiology of moral behavior: Review and neuropsychiatric

implications. *CNS Spectrums*, *14*(11), 608–620. https://doi.org/10.1017/S1092852900023853

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Structure and Function*, *214*(5–6), 655–667. https://doi.org/10.1007/s00429-010-0262-0

Miranda-Rodríguez, R. A., Leenen, I., Han, H., Palafox-Palafox, G., & García-Rodríguez, G. (2023). Moral reasoning and moral competence as predictors of cooperative behavior in a social dilemma. *Scientific Reports*, *13*(1), 1–11. https://doi.org/10.1038/s41598-023-30314-7

Moll, J., De Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourão-Miranda, J., Andreiuolo, P. A., & Pessoa, L. (2002). The Neural Correlates of Moral Sensitivity: A Functional Magnetic Resonance Imaging Investigation of Basic and Moral Emotions. *Journal of Neuroscience*, *22*(7), 2730–2736. https://doi.org/10.1523/jneurosci.22-07-02730.2002

Namkung, H., Kim, S. H., & Sawa, A. (2017). The Insula: An Underestimated Brain Area in Clinical Neuroscience, Psychiatry, and Neurology. *Trends in Neurosciences*, *40*(4), 200–207. https://doi.org/10.1016/j.tins.2017.02.002

Nummenmaa, L., Lukkarinen, L., Sun, L., Putkinen, V., Seppälä, K., Karjalainen, T., Karlsson, H. K., Hudson, M., Venetjoki, N., Salomaa, M., Rautio, P., Hirvonen, J., Lauerma, H., & Tiihonen, J. (2021). Brain Basis of Psychopathy in Criminal Offenders and General Population. *Cerebral Cortex*, *31*(9), 4104–4114. https://doi.org/10.1093/cercor/bhab072

Park, B. K., Vepachedu, S., Keshava, P., & Minns, S. (2022). Culture, theory-of-mind, and morality: How independent and interdependent minds make moral judgments. *Biological Psychology*, *174*(December 2020), 108423. https://doi.org/10.1016/j.biopsycho.2022.108423

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

Prehn, K., Wartenburger, I., Mériau, K., Scheibe, C., Goodenough, O. R., Villringer, A., Van der meer, E., & Heekeren, H. R. (2008). Individual differences in moral judgment competence influence neural correlates of socio-normative judgments. *Social Cognitive and Affective Neuroscience*, *3*(1), 33–46. https://doi.org/10.1093/scan/nsm037

Reniers, R. L. E. P., Corcoran, R., Völlm, B. A., Mashru, A., Howard, R., & Liddle, P. F. (2012). Moral decision-making, ToM, empathy and the default mode network. *Biological Psychology*, *90*(3), 202–210. https://doi.org/10.1016/j.biopsycho.2012.03.009

Ressler, K. J. (2010). Amygdala Activity, Fear, and Anxiety: Modulation by Stress. *Biological Psychiatry*, *67*(12), 1117–1119.

https://doi.org/10.1016/j.biopsych.2010.04.027

Rolls, E. T., Joliot, M., & Tzourio-Mazoyer, N. (2015). Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *NeuroImage*, *122*, 1–5. https://doi.org/10.1016/j.neuroimage.2015.07.075

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Santavirta, S., Karjalainen, T., Nazari-Farsani, S., Hudson, M., Putkinen, V., Seppälä, K., Sun, L., Glerean, E., Hirvonen, J., Karlsson, H. K., & Nummenmaa, L. (2023). Functional organization of social perception networks in the human brain. *NeuroImage*, *272*(March). https://doi.org/10.1016/j.neuroimage.2023.120025

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842. https://doi.org/10.1016/S1053-8119(03)00230-1

Schneider, K., Pauly, K. D., Gossen, A., Mevissen, L., Michel, T. M., Gur, R. C., Schneider, F., & Habel, U. (2013). Neural correlates of moral reasoning in autism spectrum disorder. *Social Cognitive and Affective Neuroscience*, *8*(6), 702–710. https://doi.org/10.1093/scan/nss051

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*(4), 667–677. https://doi.org/10.1016/j.neuron.2010.07.020

Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: Dissociating the roles of the amygdala and ventromedial prefrontal cortex. *Journal of Neuroscience*, *34*(13), 4741–4749. https://doi.org/10.1523/JNEUROSCI.3390-13.2014

Sommer, M., Rothmayr, C., Döhnel, K., Meinhardt, J., Schwerdtner, J., Sodian, B., & Hajak, G. (2010). How should I decide? The neural correlates of everyday moral reasoning. *Neuropsychologia*, *48*(7), 2018–2026. https://doi.org/10.1016/j.neuropsychologia.2010.03.023

Tomasello, M. (2014). The ultra-social animal. *European Journal of Social Psychology*, *44*(3), 187–194. https://doi.org/10.1002/ejsp.2015

van Hemert, D. A., Poortinga, Y. H., & van de Vijver, F. J. R. (2007). Emotion and culture: A meta-analysis. *Cognition and Emotion*, *21*(5), 913–943. https://doi.org/10.1080/02699930701339293

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*(March), 371–387. https://doi.org/10.1016/j.neuroimage.2017.07.043

Wilhelm Hofmann, Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2015). Morality in everyday life. *Science*, *345*(6202), 1340–1343. https://doi.org/10.1126/science.1251560

Yoder, K. J., & Decety, J. (2018). The neuroscience of morality and social decision-making. *Psychology, Crime and Law*, *24*(3), 279–295. https://doi.org/10.1080/1068316X.2017.1414817

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage*, *40*(4), 1912–1920. https://doi.org/10.1016/j.neuroimage.2008.01.057