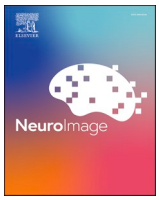
NeuroImage 320 (2025) 121485

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



Mapping the mind's landscape: Common neural encoding for spatial and morality concepts

Jing Wang [a,b,c,1,*], Miao Qian [a,b,c,1], Qing Cai [a,b,c]

[a] Shanghai Key Laboratory of Brain Functional Genomics (Ministry of Education), Affiliated Mental Health Center (ECNU), Institute of Brain and Education Innovation, School of Psychology and Cognitive Science, East China Normal University, Shanghai 200062, China
[b] Shanghai Changning Mental Health Center, Shanghai 200335, China
[c] Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, China

ARTICLE INFO

Keywords:
Semantics
fMRI
Decoding
Abstract concept
Neural representation
Multivoxel pattern classification

ABSTRACT

Abstract concepts such as *justice* are not directly tied to our sensory or motor experiences, yet they constitute an essential part of our knowledge. A longstanding question is how the brain, shaped by survival pressures, encodes these abstract concepts. This study investigated how vertical spatial representations relate to moral concept encoding in the brain ("good is up; bad is down"). We found that vertical positional processing and moral semantics elicited characteristic activation patterns, which enabled the learned neural distinctions between *up* and *down* to be generalized to decode the neural signatures of *moral* and *immoral* concepts, and vice versa, suggesting shared neural signatures between the two concept domains. Most of the vertical metaphorical representations of morality were independent of the encoding of *pleasant* vs. *unpleasant* affect, indicating the specificity of the vertical spatial representation that could not be attributed to the generic representation of arbitrary magnitude or polarity. Nonetheless, morality encoding did not rely solely on vertical spatial information, in that the morality of a word could also be decoded from neural signatures in non-spatial areas. These findings highlight both the spatial metaphorical associations and the domain-specific information in the neural representation of moral concepts.

1. Introduction

The concepts of right and wrong are universally encoded in human mind even though being abstract: they are not bound to specific objects, perceptual experiences, or actions, and get low scores on the scale of concreteness according to human raters (Brysbaert et al., 2014). Yet we say *Someone has taken the moral "high ground"* or *Humans are morally "fallible"* as if the abstract moral and immoral concepts were objects located at the two ends of the vertical dimension of the experienced physical space.

How human brains encode such abstract knowledge, namely concepts that do not have a specific and directly perceivable referent, is a longstanding question (Borghi et al., 2017). Abstract concepts are processed more slowly, less remembered (Paivio, 1991), and acquired later (Frank et al., 2017) than concrete concepts. It is proposed that the representation of abstract knowledge only relies on language information, whereas concrete concepts may resort to both language and sensorimotor information (Paivio, 1986). Recent evidence has identified the role of language experience in the neural representation of abstract knowledge (Bi, 2021; Wang et al., 2023). Nonetheless, as a system evolved for communication, language does not explicitly represent logic and is unlikely a solely qualified symbolic system for representing thought (Pinker, 2007) or its building block, concept. "Language information" itself derives from multiple sources, including the sensorimotor modalities. Attributing abstract knowledge representation to language is an unthorough account because the information content in language is not independent of sensorimotor experiences. Cognitive neuroscience evidence has shown the remarkable ability to use corpus-derived vector representations to decode or predict neural signatures of semantics, regardless of the abstractness (Mitchell et al., 2008; Huth et al., 2016; Wang et al., 2017, 2018; Pereira et al., 2018; Vodrahalli et al., 2018; Goldstein et al., 2022; Heilbron et al., 2022; Caucheteux et al., 2023; Tang et al., 2023). However, it has been challenging to uncover the neural representational principles for semantics by aligning the black

* Corresponding author at: School of Psychology and Cognitive Science, East China Normal University, 3663 North Zhongshan Road, Shanghai 200062, China.
E-mail address: wangjing@psy.ecnu.edu.cn (J. Wang).
[1] These authors contributed equally to this article.

https://doi.org/10.1016/j.neuroimage.2025.121485
Received 7 April 2025; Received in revised form 23 September 2025; Accepted 24 September 2025
Available online 24 September 2025

boxes of the brain and the language model. In short, it is still necessary to characterize the mechanism of abstract knowledge representation in the brain in addition to the generic appeal to language.

One possible explanation is that we understand abstract concepts by mapping them onto familiar, tangible domains of concrete concepts. According to conceptual metaphor theory (Lakoff and Johnson, 1980), phenomena such as using spatial words to describe morality are not mere rhetorical expressions, but manifestations of how we construe the world and construct concepts (Clark, 2024). Humans use concepts in a source domain that is tied to physical experiences (e.g. *visual perception*) to make sense of a more abstract target domain (e.g. *thought,* as in "your *point of view*" or "I *see*."). Mappings between various domains of abstract and concrete concepts have been found systematic and pervasive both in the form of languages (Lakoff and Johnson, 1980, 1999) and in nonlinguistic behaviors. Participants more frequently pair abstract concepts with the hypothetically consistent conceptual metaphor frame, process abstract concepts faster, and consider them aesthetically more pleasing when the consistent conceptual metaphor frame is presented (see Gibbs, 2011 for a review; Zhang et al., 2022; Janczyk et al., 2023). Even infants with limited language abilities and non-human primates present certain cross-domain concept mapping, such as linking a socially dominant agent to a higher spatial position (Dahl and Adachi, 2013; Meng et al., 2019). For the "moral is up" metaphor, various behavioral tasks have also found the explicit and implicit associations of the good, divine, or honorable concepts with a physically higher position and the bad with a lower position (Meier et al., 2007a, 2007b; Hill and Lapsley, 2009; Lin and Oyserman, 2021).

To examine the conceptual metaphor theory as a hypothesis of thought, it is essential to identify whether the brain represents cross-domain common semantics. Studies on rhetorical metaphors have found that processing metaphorical sentences recruits brain regions associated with sensorimotor information processing (Desai, 2022). Regarding conceptual metaphor, the existing neural evidence is loosely relevant and constrained to the common coding of abstract features such as magnitude. Judgment of space, time, and social distance has been found to recruit not only overlapped brain areas (Peer et al., 2015) but also common activation patterns (Parkinson et al., 2014). Common electrophysiological signatures have been found for coding social rank and reward value in monkeys (Munuera et al., 2018), magnitude differences in number and reward value (Luyckx et al., 2019), number across modalities (Gennari et al., 2023), and object distances in semantic and episodic memory in humans (Park et al., 2023). Findings of these and other studies (Sheahan et al., 2021; Riemer et al., 2022) suggest the existence of a "mental scale" on any conceptual dimension. The domain-general representation of relations also applies to two-dimensional mental space: The spatial coding mechanisms of the hippocampal-entorhinal cortices, namely grid-like coding and Euclidean distance coding, are found to be repurposed to enable virtual navigation through a two-dimensional space constructed from arbitrary conceptual features (Constantinescu et al., 2016; Bao et al., 2019; Bottini and Doeller, 2020; Theves et al., 2020; Park et al., 2021; Viganò et al., 2021, 2023). The cross-domain common properties discovered by these two lines of evidence are not necessarily grounded in physical experience, but rather the domain-general coding of quantitative magnitude (Walsh, 2003) and latent state (Whittington et al., 2022). Straightforward evidence is therefore still required for examining the neural implementation of the conceptual metaphor hypothesis.

The present study investigated one instantiation of the conceptual metaphor hypothesis, the spatial metaphor of moral concepts ("moral is up"), at the level of neural representation. First, we examined whether processing vertical position and moral concepts recruits overlapped neural substrates and elicits common activation patterns, so that the learned distinction of words such as "up" and "down" can be generalized to distinguish the neural signatures of words such as "justice" and "evil" and vice versa. In the aforementioned mental navigation studies (Parkinson et al., 2014; Constantinescu et al., 2016; Bao et al., 2019; Luyckx et al., 2019; Park et al., 2021; Viganò et al., 2021, 2023), the paradigms explicitly displayed the change of stimuli or a contrastive stimulus pair, in order to measure the neural responses to the virtual movement along the featural dimension or navigation in the featural space. To examine metaphorical representation of semantics, this study required participants to process real words in isolation. If conceptual metaphor has its place in the neural representation of abstract semantic knowledge, the shared coding should be observed when the task only requires passive semantic processing of single concept that has already been encapsulated in the form of words.

Second, regarding the necessity of the metaphorical representation, we investigated whether the neural representational distinction between moral and immoral concepts fully relies on the metaphorical vertical positions. Third, we investigated whether the metaphorical vertical representation of morality was driven by the affective valence in moral concepts, that is, whether the "moral is up" is one instantiation of the metaphor "positive is up". By looking for neural signatures that were specific to the vertical-moral representation, we addressed whether there is metaphorical representation of abstract concepts that is specific to certain pairs of concept domains, in addition to the domain-general structural mapping.

2. Results

2.1. Decoding morality in brain regions of vertical spatial processing

The first question was whether the brain regions that activated differently for upper vs. lower locations also encoded morality of words. In the first fMRI task, thirty-two human adults were guided to generate mental imagery about a classroom building and its surroundings. One classroom was used as the fixed referent. The task was to judge the position of a list of objects as being above, below, inside, or outside the reference classroom (Fig. 1a). Mass uni-voxel general linear models were performed to locate voxels showing different activation magnitudes when processing the above vs. below relations. Significant differences were found in the parahippocampal gyrus, hippocampus, inferior parietal cortex, precuneus, and several lateral and medial frontal cortices (cluster-wise corrected $p = 0.05$; Fig. 2a). These voxels were considered to encode vertical spatial positions and then investigated for the encoding of moral word concepts.

In the second fMRI task (Fig. 1b), participants read a list of 64 words four times and made occasional judgments on whether the word's meaning was pleasant or not. These words were chosen to represent eight semantic categories: up, down, moral (e.g., 高尚, the Chinese equivalent of *nobleness*), immoral (e.g., *vileness*), happy, unhappy, inside, and outside. The question was addressed using a within-domain (where the domain was morality) classification within the spatial processing regions. The neural signatures of semantic processing were estimated for each participant for each trial of word in the voxels identified by the spatial relation task.

To learn the multivoxel patterns associated with moral or immoral words, logistic regression classifiers were trained on all but one morality word and tested on the left-out word to label it as being moral or immoral based on its multivoxel patterns (see *"Within-domain decoding"* in *Methods*). The procedure was iterated in a cross-validation protocol until each morality concept was tested once. The mean accuracy across participants of classifying an unseen word as moral or immoral concepts based on its fMRI signature was 0.73 (95 % CI [0.46, 0.54] based on random permutation test), significantly above chance-level accuracy ($p < 0.05$; Fig. 2b). The effect was robust at the individual participant level: 78 % (25/32) participants showed above-chance accuracy (Figure S1).

Within these vertical spatial brain areas, it was also possible to classify up vs. down (mean accuracy = 0.71) and happy vs. unhappy concepts (accuracy = 0.67) by training the corresponding classifiers (Fig. 2b). These areas do not encode the semantics of enclosure (mean accuracy of classifying inside vs. outside = 0.54, not being significantly

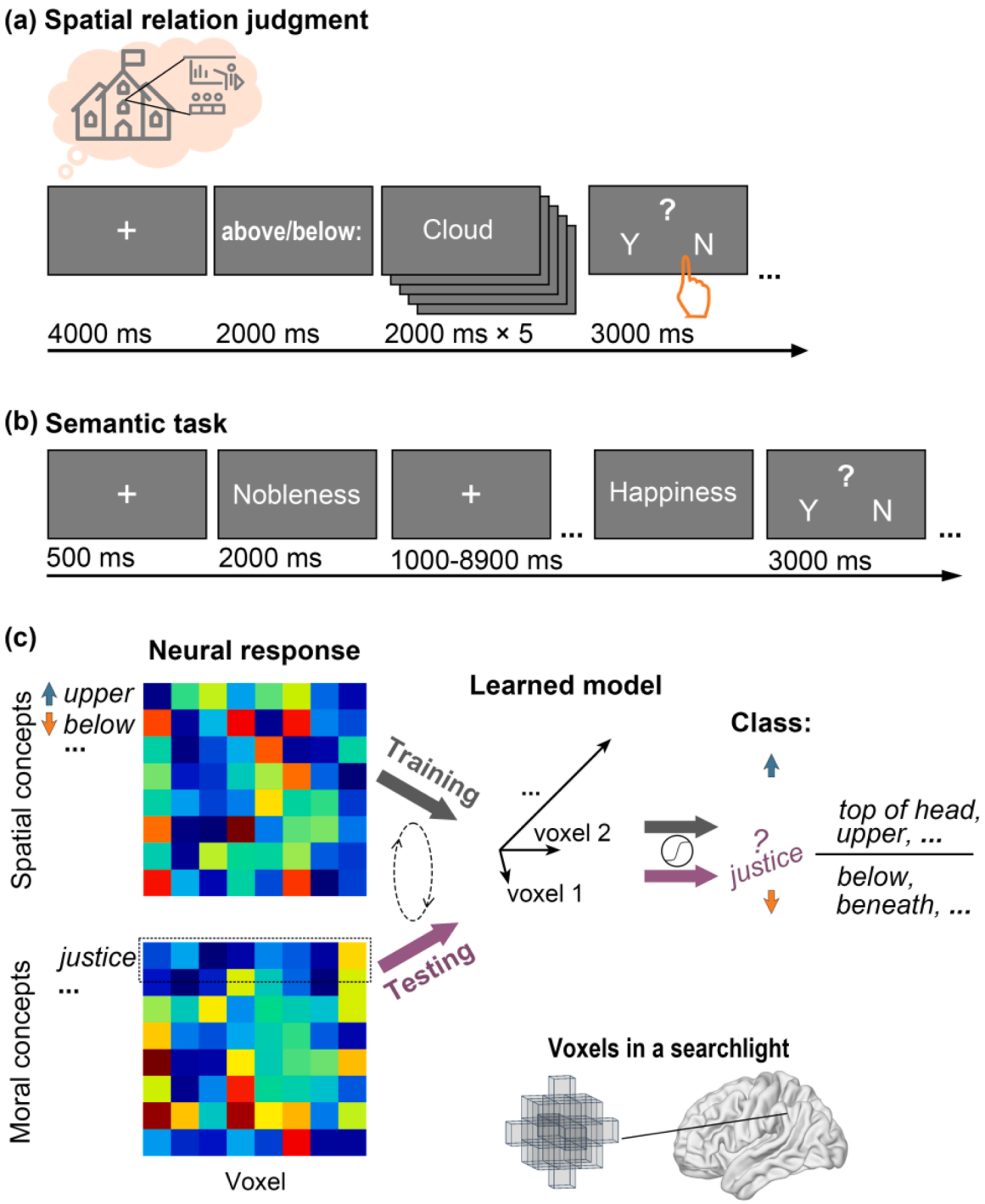


Fig. 1. Experimental paradigm and analytic approach. **(a)** A block in the spatial relation judgment task. **(b)** A trial in the semantic processing task. **(c)** Classification across the concept domains of space and morality based on multivoxel activation pattern per searchlight.

different from chance-level accuracy and significantly different from other classification accuracies). The result suggested that these areas specifically encoded the semantics of vertical spatial concepts and moral concepts rather than the generic semantics of polarity.

To investigate the low-dimensional representation of the neural semantic space, we applied the DISTATIS method (Hervé Abdi et al., 2005), an extension of principal component analysis that allowed the integration of distance matrices from multiple sources. Inter-word pairwise dissimilarity of activation patterns was computed for each participant. Compromise was computed to represent a weighted aggregation of neural dissimilarity matrices from multiple subjects. Eigen decomposition on the compromise matrix extracted the common structure underlying all participants' dissimilarity matrices of words. The first component opposed the concrete physical concepts (e.g., *over the sky, below*) to the abstract mental concepts (e.g., *hypocrisy, happiness, honesty*; Figure 2c; Figure S2). The second component captured the *up* vs. *down* difference in the literal sense for the spatial words and in the metaphorical sense for the moral words. Words at the upper end of this dimension included *justice, mountain top*, etc., whereas words at the lower end included *obscene, below*, etc. (Fig. 2c). By contrast, the non-moral affective words did not present a consistent metaphorical representation of valence in the two-dimensional space (Figure 2c;

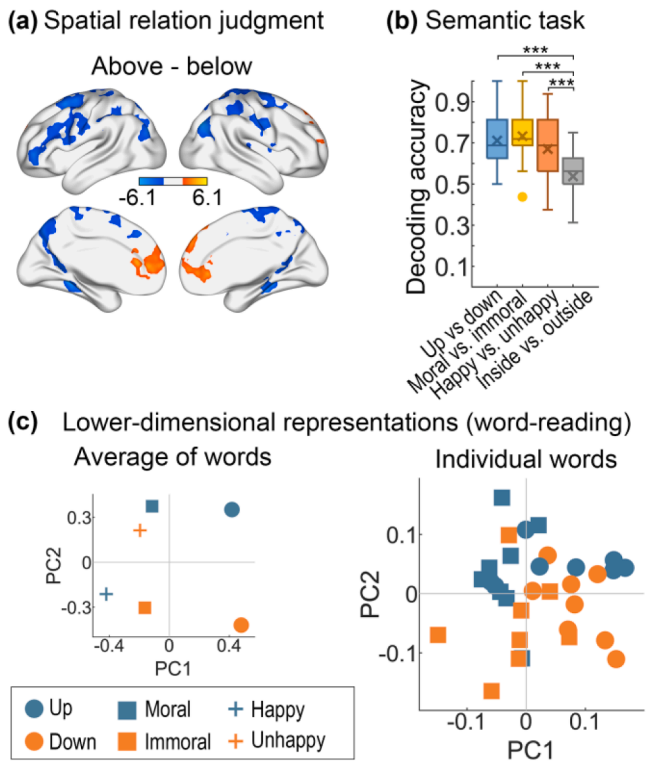


Fig. 2. Concept representation in spatial relation processing areas. **(a)** Clusters where the activations showed significant effect in the contrast of *above – below* conditions during the imaginary spatial relation processing task. Red: *above* > *below*. Blue: *above* < *below*. **(b)** Boxplots of the accuracy distribution of classifying individual words' semantic category in the spatial processing regions as shown in (a). The " × " indicated the mean over participants. The dot indicated the outlier that was 1.5 interquartile range away from the non-outlier minimum/maximum. ***: $p < 0.01$. The contrasts were explained in the main text. **(c)** Two-dimensional representations of the neural activation patterns of words in regions identified by the spatial relation task. The first dimension distinguished the concrete from the abstract concepts. The second dimension represented the "up" and "down" for the spatial and moral concepts but not the affective concepts. In the left figure, each marker represented the average of 8 words. In the right figure, each marker represented one word. The affective words were not plotted on the right figure for visualization purpose (see Figure S2 for plots with affective words and identity of each word).

Figure S2).

The task relied on mental imagery of specific scenarios. The difference in activations was likely to involve the representational difference between different categories of objects. The activations elicited by this task were likely to involve the mental imagery of objects. Moreover, the *above - below* contrast did not exclude voxels that are generally sensitive to spatial relations. The *in - out* contrast revealed substantial overlap with the *above - below* contrast (Figure S3a). Therefore, we performed the classification analyses again after excluding the overlapped areas (Figure S3b). The mean accuracy of *moral* vs. *immoral* classification dropped to 0.72, which was significantly lower than the original result ($M = 0.75$), but still above chance-level ($p < 0.001$). The pattern of all the classification results (Figure S3c) was similar to the classifications performed in the non-specific areas (Fig. 2b).

2.2. Shared multivoxel patterns between spatial and moral concept representations

The further question was whether vertical position and morality words used not only the same neural substrates, but also the same activation patterns to represent the literal and metaphorical "up" vs. "down". To identify such representation across the whole brain, searchlight multivoxel pattern classifications were implemented in a bidirectional cross-domain manner (Fig. 1c). One type of classifiers was trained on data associated with (spatial) up vs. down words, and applied to classify the neural signature of each morality word. If a moral word was labeled as "up" or an immoral word was labeled as "down", the classification was considered correct. The other type of classifiers was trained on moral vs. immoral words and tested on the vertical position words. In each spherical searchlight, the mean accuracy over both training-testing directions over all test words indicated whether this cluster of voxels encoded the morality of a concept in the same way as it encoded vertical position concepts. Random permutation test was performed to estimate the empirical null-hypothesis distribution and to evaluate the significance of the accuracy.

Widely distributed areas revealed common encoding patterns for vertical position and morality concepts at the group level. These areas included those robustly identified for semantic representation and other regions, namely the anterior temporal lobe (ATL), superior and inferior parietal lobule, hippocampus, *pars opercularis* and *pars triangularis* of inferior frontal gyrus (IFG), fusiform gyrus, superior temporal sulcus (STS), left dorsomedial prefrontal cortex (dmPFC), anterior cingulate cortex (ACC), precuneus and occipital areas (FDR-corrected $p = 0.05$; Figs. 3a, 3b; Figure S4).

In addition to the clusters with significant mean accuracy, some brain regions bore larger variability across participants, so that highly informative clusters in one participant were not aligned with others. To identify these areas, the searchlight decoding accuracy map of each participant was thresholded at $p = 0.05$ (FDR corrected) and examined at the level of atlas-defined regions (see Methods). Monte Carlo simulations were applied to evaluate how likely at least that many voxels would be found in a given region if the significant voxels were randomly located across the brain. The resulting p values were FDR-corrected for multiple comparisons of all regions. One hundred and eighteen out of 384 areas were found to have a reliable number of informative voxels (Figs. 3c, 3d). The vast majority of these regions had already been identified by the voxelwise-matched analyses, except the anterior section of the left hippocampus, bilateral medial ATLs, and the bilateral thalami (Figure S5). Within each region, the inter-subject dissimilarity of the accuracy map was calculated to indicate its individual variability (see *Methods*). These additionally identified regions exhibited comparable mean and peak performances with those identified by the voxelwise-matched analyses (Fig. 3e), but the inter-subject dissimilarities were among the highest and the proportions of informative voxels were among the lowest (Figs. 3f, 3g, 3h). Overall, the medial ATL, anterior hippocampus, and thalamus had small informative clusters and high individual variability for vertical moral representation.

When all the significant searchlight center voxels were aggregated to train one classifier on the neural signatures of vertical words and test it on moral words, we obtained an overall neural representation of the moral concepts along the vertical dimension (Fig. 4a). On the dimension expanded by the up vs. down classifiers, the probability of each moral word being "upper" was greater than 0.5 and that of each immoral word was below 0.5. The predictions of individual participants' models on individual words were more confident (i.e., extreme probability scores) but not perfect (Fig. 4b).

2.3. The specificity of spatial metaphor to moral representation

"Happy is up, sad is down" is a well-documented conceptual metaphor. The moral words were affectively more positive than the immoral (Table S1). Therefore, it was possible that the shared encoding patterns of moral concepts with vertical position in the searchlight areas were not specific to morality, but rather part of the more general mental association between spatial and affective concepts. To identify the vertical-moral representations that could be accounted for by valence, we performed cross-domain searchlight classifications between the 16 spatial and 16 affective words (*up = happy, down = unhappy*), and classifications

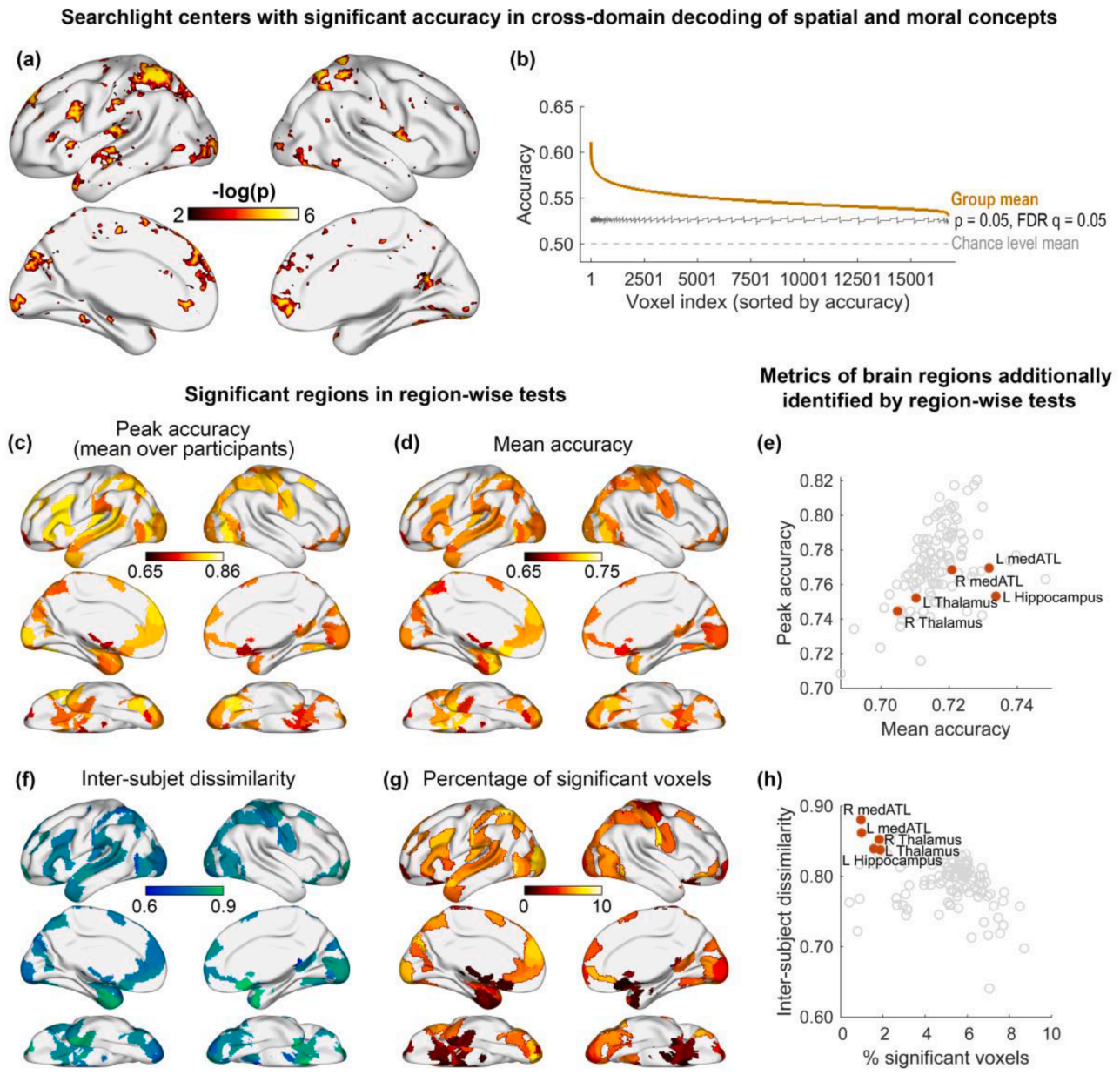


Fig. 3. Regions with common representations for vertical spatial words and moral words. **(a)** Center voxels of searchlight with common patterns for spatial and moral words, so that a classifier trained to classify *up* vs. *down* could be generalized to decode a moral word as *up* and immoral word as *down*, and vice versa. The color indicated the significance of accuracy by -log10(p). **(b)** Mean accuracy across participants for each group-level significant voxel. **(c)** Mean of the maximal accuracy within each atlas-defined region. **(d)** Mean accuracy across voxels in each region. **(e)** Mean and peak accuracy of the regions additionally identified in the regional tests as compared to the regions identified in the voxelwise-matched analysis. The red dots highlighted the 5 regions that were found significant in the region-wise analysis but did not show significant voxels in the voxel-wise matched analysis. The gray circles indicated all the other regions. The additionally identified regions showed a range of performances that were comparable to others. **(f)** Inter-subject dissimilarity (1-cosine) of accuracy map per region. **(g)** Percentage of significant voxels in each region. **(h)** All the newly identified regions showed great inter-subject variability and small proportion of significant voxels compared to the other regions.

between the affective and morality words (*happy* = *moral, unhappy* = *immoral*), within the identified vertical-moral regions. The affective-moral classifications showed reliable accuracies in two small clusters (51 voxels, i.e., 0.41 cm^3) in the dorsomedial prefrontal cortices (dmPFC) and left insula. Voxels surrounding the two clusters survived at a more lenient threshold that did not employ the FDR correction (p = 0.05; Fig. 5a). Moreover, no voxel showed reliable accuracy in the vertical-valence cross-classification in the vertical-moral region. These results suggested that the dmPFC and insula clusters encoded the valence of affective and moral concepts in a different way from how they encoded the metaphorical verticality of morality.

The whole brain was then searched for common activation patterns between the affective and morality words, using the same procedure as the whole-brain searchlight vertical-moral decoding. Significant (FDR-corrected $p = 0.05$) accuracies were found again in the dmPFC, and two small clusters in the left putamen and postcentral gyrus (Fig. 5b). The results suggested the small proportion of affective-moral representation in the vertical-moral areas was not due to a biased pre-selection of voxels.

We also checked if the small number of affective-moral representation voxels at the group level was due to a lack of inter-subject consistency. The decoding accuracy maps of each participant were

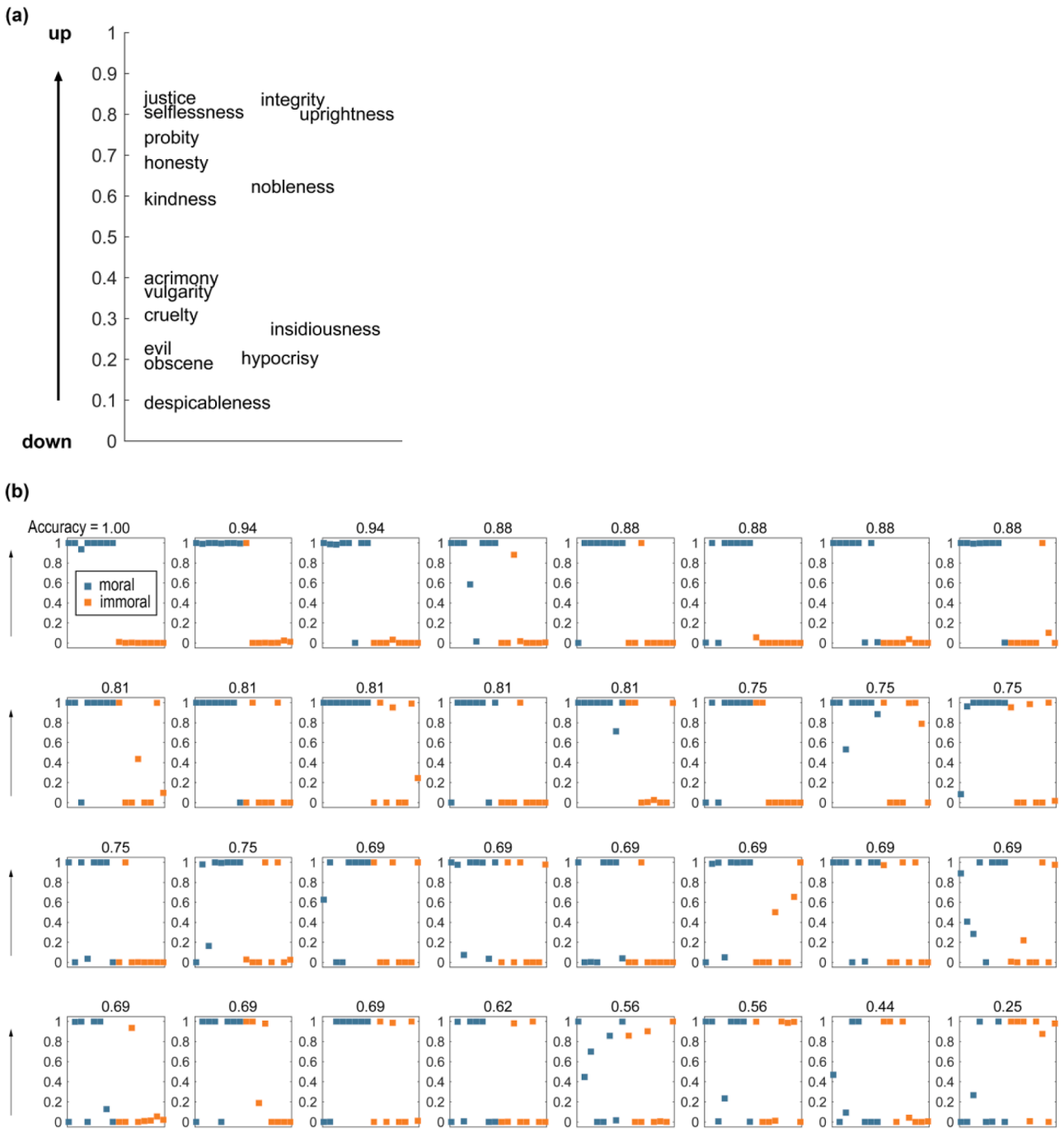


Fig. 4. Post-hoc visualization of the moral value line in the brain. **(a)** Classifier's predicted probability of each morality concept being *up* vs. *down,* averaged across participants. **(b)** The predicted probability of each morality concept being *up* vs. *down* in each participant's model.

thresholded using the random permutation tests based on participant-specific data. For any participant, among all the voxels with significant vertical-moral classification accuracy, over 80 % of the voxels are specific to the vertical moral representation, while few voxels showed common representation of vertical, affective, and moral concepts (Fig. 5c). Across the whole brain, voxels that showed significant accuracy only for vertical-moral classification were significantly more than the voxels with significant accuracy only for affective-moral classification, only for the vertical-affective classification, or for all three kinds of cross-domain decoding (paired-sample *t*-test on the mean number of voxels across participants, $p = 0.02$, $p < 0.001$ and $p < 0.001$ respectively; Fig. 5d). In short, the group-level pattern held at the single-subject level: the majority of brain areas for vertical moral representation were not accounted for by the postulated link between vertical position and affective valence.

2.4. *The necessity of vertical position metaphor for morality identification in limited regions*

Despite the presence of common encoding patterns between moral and spatial concepts, does the neural representation of whether a concept is moral fully rely on metaphorical verticality? The finding of common patterns for encoding moral and affective words (Fig. 5b) had already suggested the verticality-independent representation in dmPFC. We further investigated this question from two aspects. First, the activation patterns of "up" and "down" were respectively regressed out from

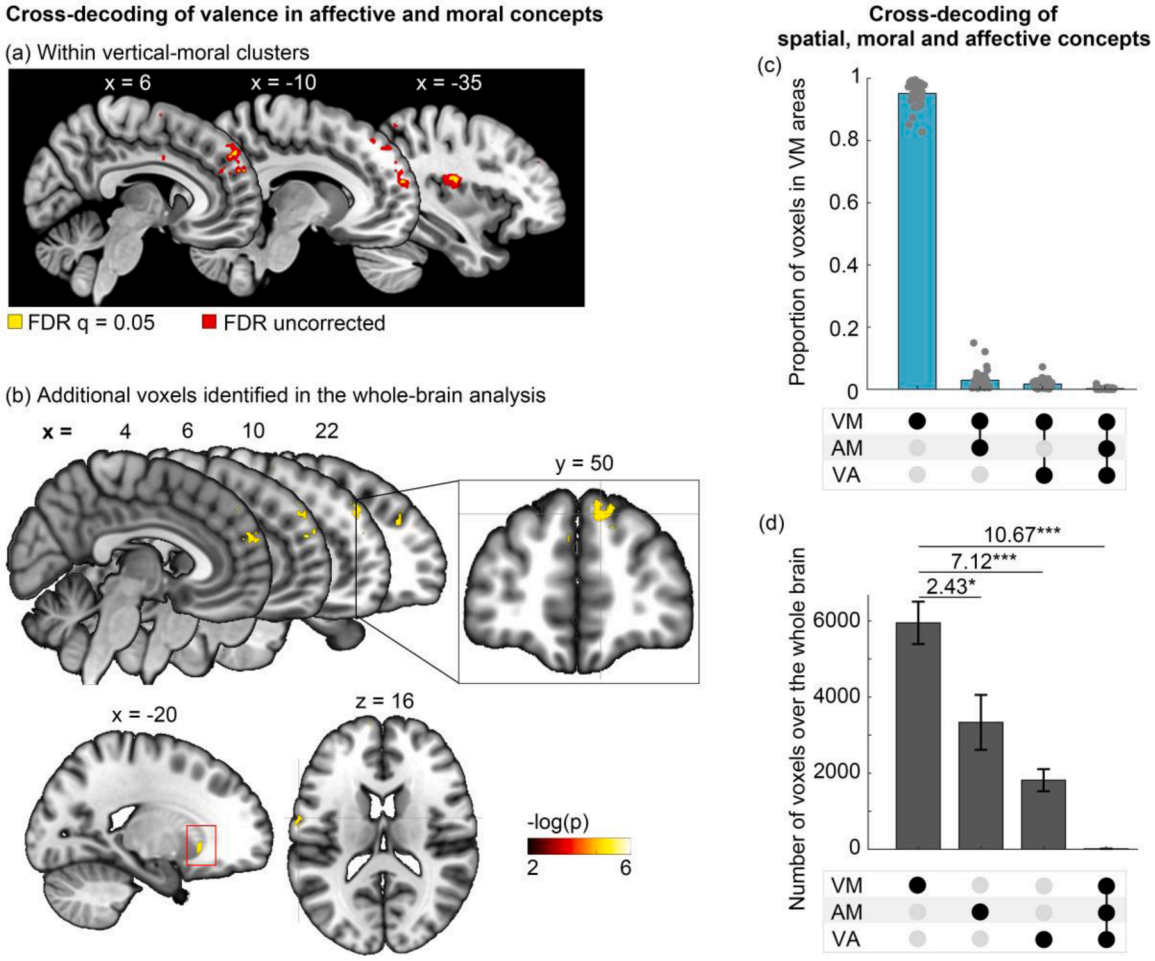


Fig. 5. Limited effect of affective valence in the representation of moral concepts. **(a)** Center of searchlight that showed common patterns for affective and moral words, within the areas showing common representation for vertical and moral concepts. **(b)** Center of searchlight that showed common patterns for affective and moral words in the non-vertical-moral areas. **(c)** Within vertical-moral areas, the proportion of voxels with significant accuracy in other cross-domain classifications at the individual participant level. The four colored bars represented the mean proportion over participants; the dots represented individual participants. The x-axis indicated the pattern of classification results. For example, the first column referred to voxels with significant accuracy for VM decoding (black circles) but not for AM or VA decoding (gray circles). VM: cross-domain classification of vertical and moral concepts. AM: cross-classification of affective and moral concepts. VA: cross-classification of vertical and affective concepts. **(d)** Number of voxels revealed by the whole-brain searchlight cross-classification for individual participants. The t values of the paired-sample tests were labeled. The gray bars represented the mean number of voxels across participants. The error bars indicated the standard error. The x-axis indicated the type of classification. VM, AM, and VA: same as Fig. 5c.

the activation patterns of moral and immoral concepts in each searchlight sphere identified by the vertical-moral cross-classification. Within-domain classifications were performed on the residuals to identify the morality of each moral/immoral concept per searchlight, using the same procedure as the aforementioned within-domain classifications. If a center voxel still showed significant accuracy, the spherical area was considered able to represent the morality of a concept using neural signatures that could not be explained by the vertical spatial concepts. These voxels accounted for 8.3 % of all the vertical-moral voxels, sporadically located in the bilateral inferior parietal lobule (IPL), precuneus, and left inferior frontal gyrus (Figure S6). In other words, for the rest majority of the regions that were revealed by the vertical-moral cross-classification, the metaphorical vertical representation was necessary for decoding the morality of a concept. The results by far still spoke for the necessity of the metaphorical vertical position in representing moral concepts.

Second, we examined the morality representation in areas other than those identified in the vertical-moral cross-classification. The within-domain classification was performed on these remaining voxels by training classifiers on all but one morality word to decode the left-out morality words. The mean classification accuracy across all the participants' models was 0.72, being significantly above the chance-level accuracy ($p < 0.001$; 78.1 % of participants' models showed above-chance accuracy at individual level; Figure S7a). Even when a broader area was excluded by removing voxels showing cross-domain accuracy that was significant before FDR correction for multiple comparisons, the mean accuracy remained 0.72 ($p < 0.001$). For the same purpose, we also decoded morality after removing all the voxels identified by the imaginary spatial relation processing task. For the morality words, it was still possible to decode morality from these remaining voxels with a mean accuracy of 0.73 ($p < 0.001$; 71.9 % of participants' models showed above-chance accuracy; Figure S7b). By contrast, for the literal spatial words, the accuracy of decoding vertical position from the remaining voxels was 0.68 ($p < 0.001$; Figure S7c), with only 46.9 % of participants' accuracy being above chance level. These classification results suggested there was additional neural representation of the core semantics of morality beyond the metaphorical vertical position.

The follow-up question was how much information the non-metaphorical voxels provided for representing morality, in addition to the metaphorical vertical representation. A straightforward test was to examine if adding the non-metaphorical voxels to the metaphorical model would increase the accuracy of morality classification. To obtain

the baseline accuracy, the training of within-domain classifiers used all the vertical-moral encoding voxels (i.e., voxel locations in Fig. 3a). When the classifiers were applied to decode morality words that were unseen by the models, the mean accuracy across cross-validation folds across participants was 0.74 ($p < 0.001$). Adding all the remaining voxels to the classification resulted in an accuracy of 0.72, which was numerically lower but not statistically different from the baseline (paired-sample t-test, $p > 0.05$). Thus, the non-vertical representation of morality did not add to the neural differentiation between the moral and the immoral concepts.

We computed Kullback–Leibler divergence (KLD) to quantify the extra information gained by adding the non-metaphorical voxels. Compared to the baseline (vertical-moral) model, KLD of the predicted probability distribution over the two classes increased when a subset of non-metaphorical voxels with stable response profiles over repeated presentations of words were included, and then reached a plateau (Fig. 6a). The entropy of the prediction decreased first and then recovered to the baseline level (Fig. 6b), suggesting the increase of models' prediction confidence was diluted when more remaining voxels were included. Overall, the non-metaphorical voxels provided new information content additional to the vertical moral representation, although it did not lead to a more accurate classification of morality.

3. Discussion

This study investigated the metaphorical representation of moral concepts in human brain. We summarize the main results as follows. The brain areas involved in spatial relation judgment also encoded the semantics of morality and affective valence during word processing. The literal or metaphorical "position" of spatial or moral concepts was consistently represented in these brain areas. Common activation patterns elicited by vertical spatial words denoting "up" and moral concepts systematically differed from those elicited by words denoting "down" and immoral concepts in multiple brain areas. Such encoding was not shared with words expressing the happy or unhappy affective state. Additionally, whether a concept was moral or not could also be decoded from the neural signatures in non-metaphorical regions.

Taken together, our results reveal that a set of neural signatures encoding vertical spatial locations are generalizable to the encoding of morality but not any arbitrary polarity concepts. This finding suggests a possible neural representational principle of abstract semantic knowledge and a possible neural realization of the conceptual metaphor.

Mounting evidence has shown that the brain processes abstract knowledge differently from the concrete (Wang et al., 2010). This difference can be attributed to the differential reliance on language system (Wang et al., 2020; Bi, 2021; Wang and Bi, 2021). While the embodied view emphasizes the role of sensorimotor information in addressing the explanatory limitations of the language account (Barsalou, 2008), empirical evidence is needed to test how this account might apply to, and whether it is sufficient for, the neural representation of abstract semantic knowledge that lacks a necessary link to sensorimotor features. We postulate that the metaphorical use of embodied information may function as a scaffold between perceptual experiential and domain-general coding, whether it be linguistic or nonlinguistic symbols and structures. We base our prediction of the neural representation of abstract concepts on the conceptual metaphor theory. The study uses cross-domain decoding to set aside the philosophical debate on whether the "nature" or coding format of certain representation is embodied, and to identify the neural signatures that distinguish concepts in both the concrete domain and the abstract domain. It enhances our understanding of how the brain can possibly encode abstract concepts, a large proportion of which are evolutionarily recent inventions.

3.1. Specificity of metaphorical vertical representation to moral concepts

The neural metaphorical representation of abstract concepts exhibited duality: We identified the neural signatures that were sensitive and specific to vertical positions in the literal and the metaphorical sense, and we also found the possibility of decoding morality without the neural code of vertical position. Most of the common activation patterns were specific to the vertical and moral concepts and independent of the coding of affective valence (Fig. 5). The neural codes of magnitude and two-dimensional feature space identified by previous studies (Parkinson et al., 2014; Constantinescu et al., 2016; Munuera et al., 2018; Bao et al., 2019; Luyckx et al., 2019; Viganò et al., 2021, 2023; Gennari et al., 2023; Park et al., 2023) suggest the existence of domain-general representation of structure or relation that is abstracted away from physical properties. Although the conceptual metaphor is also proposed as a cross-domain mapping, the hypothesis emphasizes its grounding in the experience of our interactions with the world (Lakoff and Johnson, 1980). In other words, not all the concepts with similar abstract structures are necessarily represented in the same way. In the present results, the dissociation of non-affective, vertical-specific representation from the affective-specific representation of moral concepts suggests the existence of a specific metaphorical representation that is not a domain-general representation of polarity or magnitude of any arbitrary feature.

The regions independently bearing the metaphorical representation include those consistently identified as part of the general semantic system (Binder et al., 2009), namely the left IFG, dmPFC, STS, and IPL.

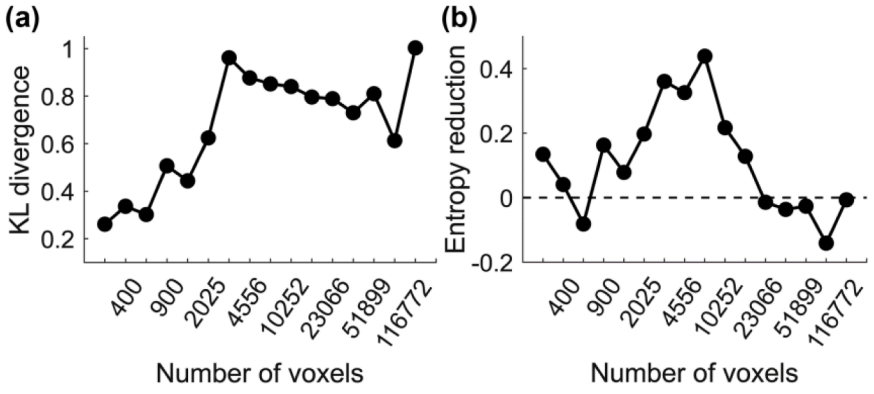


Fig. 6. Change of information-theoretic measures of the classifiers of morality decoding as the number of non-metaphorical voxels included in the model increased, averaged across participants. **(a)** Kullback–Leibler divergence of the predicted probability distribution of the model when extra voxels were included as compared to the distribution of the baseline model. The baseline model was the classifier trained on all the vertical-moral voxels. The number of extra voxels being selected was set to the rounded values of $400 \times 1.5^{n-1}$, where n was a natural number ranging from 1 to 16. **(b)** Reduction of entropy of the predicted probability distribution when extra voxels were included as compared to that of the baseline model.

ATL is known for representing abstract semantics and coding concepts in a sensory-independent way (Wang et al., 2020). The parietal cortex, particularly the intraparietal cortex and supramarginal gyrus, has also been found to code cross-modal distance (Parkinson et al., 2014; Riemer et al., 2022; Park et al., 2023; Viganò et al., 2023) and considered crucial for representing concepts in the self-centered frame of reference (Bottini and Doeller, 2020). In our case, the semantic categories (e.g., up vs. down) do not differ in distance. If the neural signatures in parietal areas code concepts in an egocentric perspective, the information coded in this experiment is likely the body-centered orientation. The hippocampus is not typically found in studies that investigate a large scope of the semantic space at once, but the hippocampal-entorhinal areas, along with the human medial prefrontal cortex, are known for representing concept relations in the two-dimensional feature space (Constantinescu et al., 2016; Bao et al., 2019; Theves et al., 2020; Park et al., 2021; Viganò et al., 2021, 2023). The current result further shows that such domain-general coding of relations also exists in the semantic knowledge of real, known words and can be detected during the word reading task. It suggests the representation of semantic knowledge and non-semantic conceptual "objects" in artificial feature space may share the same coding of concept relations.

3.2. Necessity of metaphorical vertical representation to moral concepts

For over 90 % of the voxels within the vertical-moral regions, the neural coding of verticality was necessary for coding morality, in that regressing out the former resulted in a chance-level accuracy for decoding the latter (in its within-domain classification). On the other hand, morality could still be decoded from the activation patterns of the rest non-vertical-moral areas in its within-domain classification. This result and the information-theoretic measures (Fig. 6) jointly suggested there was neural representation of morality in addition to the code of metaphorical verticality. Thus, if we operatively define the "decodability" of morality as the statistical significance of the accuracy of classifying concepts as being moral or immoral as compared to the chance-level accuracy, the tentative conclusion is that the metaphorical vertical representation is necessary for moral representation in multiple brain regions, but unnecessary at the whole-brain level. There has been ongoing debate on whether conceptual metaphor is necessary for the understanding of concept and metaphorical expression in any form, or whether it is epiphenomenal (Gibbs, 2011). Previous behavioral studies have already provided evidence of the interference between the source (concrete) and target (abstract) domain on participant's reaction time or subjective judgment (Meier et al., 2007a, 2007b; Hill and Lapsley, 2009; Dahl and Adachi, 2013; Meng et al., 2019; Lin and Oyserman, 2021; Zhang et al., 2022). This study examines this question from the perspective of neural representation, showing the unique neural signatures of metaphorical representation and the unique information they provide to morality identification.

It might seem paradoxical that adding non-vertical voxels to the model did not improve the performance of morality decoding. That is, the vertical and non-vertical neural signatures represent different aspects of morality, but the joint information does not help with the guess. An illustration of such a possibility is that when two non-sighted people were to guess the identity of an animal by touching it, one touched a horn and the other touched a slender leg. Regardless of them knowing the information from the other person, both concluded that the animal was likely a gazelle, even though the guess was not perfect (it happened to be a deer) and the two features of body parts are not fully coupled in the population of species (i.e., provide additional information to each other). Future study is required to delineate the non-metaphorical representation and explore when it might provide extra information for identifying the abstract semantics.

3.3. The intrinsic metaphorical coding in semantic knowledge

Two interrelated differences in design between this and previous studies on the representations of conceptual metaphor or domain-general structure were that (1) we focused on the neural representation of known words, i.e. semantic knowledge, instead of novel objects represented in temporarily built feature space, and (2) the task for obtaining neural responses to words required very limited cognitive operations in addition to passive word reading. This paradigm was chosen to tap into the natural, intrinsic representation of abstract concepts in our semantic knowledge system. Even in occasional trials (the data from which were not included in the analyses), participants were asked to judge the pleasantness of a word, the results still distinguish the pleasantness-independent vertical moral representation from the valence-specific representation. We attempt to illustrate that the metaphorical representation is not just an effect imposed by specific tasks of learning, navigation, or decision making, but is inherently present when human accesses familiar concepts that have already been packed into and are retrieved through words.

3.4. The limited and independent role of affect in the semantic representation of moral concepts

Affective valence has been found represented by both domain-specific and domain-general neural codes across stimuli and sensory modalities (Chikazoe et al., 2014; Kragel et al., 2023). In this study, words in the *moral* and *happy* categories were not statistically different in the mean subjective rating on either valence or arousal; so were the *immoral* and *unhappy* words. Moral words were rated more positive than immoral words (see Methods). Considering the common linguistic evidence for the "happy is up" metaphor, we had expected that a substantial proportion of metaphorically vertical representation of morality would have been accounted for by the vertical representation of valence in the brain. In fact, the dmPFC clusters for the valence coding across moral and affective words (Figs. 3a, 3b) have been identified as part of the common appraisal system (Bo et al., 2024). However, only a small proportion of voxels had a common valence coding for the moral and the affective words. This was consistent with what the lower-dimensional representation in all the regions of spatial processing revealed (Fig. 2c). The lower-dimensional representation result on affective words (Fig. 2c) seemed counterintuitive, in that the mean of positive concepts was in the same quadrant as the immoral concepts, and the mean of negative concepts was in the same quadrant as the moral ones. A careful look into the effect showed large variability in the locations of individual affect concepts and little consistency within the positive or negative category (Figure S2). Taken together with other findings, we infer that the results suggest little association between the affective and vertical concepts. Moreover, although some participants showed considerable vertical-affective representations (Fig. 5d), these effects were not reliable at the group level.

The fact that the within-domain decoding accuracy of affective words was comparable to the moral and spatial words (Fig. 2b) indicated that the valence information in the affective words could be captured by the present data. Thus, the limited role of affect in moral representation was not due to the lack of sensitivity to valence. These results suggested that the affective aspect of moral concepts was neurally represented in a different way from the spatial metaphorical representation. Affect as an independent factor can influence moral judgment depending on the context. However, the present study focused on the core semantics of moral concepts. Thus, the paradigm was unlikely to capture the effect of emotion during various kinds of moral thinking. The mapping between the semantics of morality and vertical position might reflect the semantics of rule, status, or transcendence that is not related to pleasantness. We did not assume, nor did the results suggest that the up-down metaphor encoded only moral concepts, or that verticality was the only basis for moral representation. For example, the IPL, precuneus, and

ACC that were identified by the cross-domain searchlight decoding were part of the theory of mind network (Saxe and Kanwisher, 2003; Schurz et al., 2021). It was possible that moral representation was not directly grounded in the metaphorical spatial concept, but contributed by the social relation schema, which was associated with both spatial cognition and moral judgement.

One concern about the conceptual metaphor theory has been the ambiguous boundary between a conceptual metaphor and a "dead" metaphor or simply polysemy (Keysar et al., 2000). What the present results remind us is that the extent to which a concept is represented via the hypothetical metaphorical semantics is an empirical cognitive and neuroscience question. Future study is required to examine the hypothetical source domains of conceptual metaphor jointly and quantify their contributions to specific abstract concepts.

3.5. Shortcomings of the study

Conceptual metaphors are considered a fundamental mechanism for abstract thoughts, but the interplay between thought and language complicates the problem. This study only examined speakers of a single language, thus it could not dissociate the effect of the specific language from that of the universal concept representation. Future study is required to investigate to what extent the difference in the expression of different languages affects semantic representation in addition to the conceptual metaphor.

The group-level effect of metaphorical representation does not suggest it applies to every participant. Although human beings have shared experience on many of the proposed source domains of conceptual metaphors, cultural factors and personal experiences may still affect the extent to which certain abstract concepts are metaphorically represented by certain source domains. Future studies are required to systematically investigate the association between physical experiences and conceptual metaphor at the individual level.

One limitation in the design was the potential pollution between tasks: the spatial mental imagery task in the first experiment might implicitly activate participants' spatial representation schema, which might carry over to the word reading task in the second experiment. It was possible that the vertical moral representation was amplified by such design. On the other hand, the probe task of experiment 2 was a pleasantness judgement, which might artificially elicit valence representation.

In addition to stroke count, word frequency, abstractness, valence, and arousal, other potential psycholinguistic confounders to the classification analyses included imageability and familiarity of words. We did not collect imageability ratings primarily due to practical reasons and partly because of the high correlation between imageability and concreteness/abstractness (Paivio, 1986; Yao et al., 2017). We did not collect familiarity ratings primarily due to practical reasons. Moreover, familiarity was a comprehensive subjective evaluation that is well explained by or correlated with word frequency and concreteness (Brysbaert and Cortese, 2011; Yao et al., 2017; Su et al., 2023). Nonetheless, imageability and familiarity may account for neural signatures additional to abstractness and frequency, and thus should have been balanced between semantic categories.

4. Methods

4.1. Data and code availability

The neuroimaging data and the code are deposited at https://osf.io/7egaz/. Any additional information required to analyze the data reported in the manuscript is available from the lead contact upon reasonable request.

4.2. Participants

Thirty-two adults were recruited from the East China Normal University community (24 females, mean age = 23 years, SD = 1.55). All participants were right-handed native speakers of Chinese. They had normal or corrected-to-normal visual acuity and no history of neurological or psychiatric disorders. All participants provided written informed consent approved by the East China Normal University Institutional Review Board (HR2-0015-2021).

4.3. Experimental design and procedure

4.3.1. Spatial relation judgment task

Participants were asked to imagine one typical classroom on the third floor of a five-floor school building. This classroom (the space enclosed by the walls, ceiling, and floor) was used as the referent for the entire task. In the block-designed fMRI experiment, participants saw five nouns sequentially in each block. Each noun was a two-character name of a familiar object (e.g., blackboard, school gate). After the last word, a question mark with two response options was presented. Participants pressed a button to indicate whether the typical spatial relations of the five objects to the referent were consistent – for example, all being outside a classroom. The spatial relation to be considered was restricted to either vertical position (above or below the referent) or enclosure (inside or outside the classroom). At the beginning of each block, a cue word was presented in bold font to specify the relation of interest for that block. The spatial relations of the objects to the referent were consistent for 16 blocks and inconsistent for 4 blocks. The 16 consistent blocks were composed of 4 types of relations (above, below, inside, and outside) × 4 blocks per relation. The presentation order of the blocks was randomized and ensured that no three consecutive blocks belonged to the same condition. Each block started with a 4000 ms period during which the participants were asked to fixate on the "+" at the center of the screen. Each word, including the cue word, was presented for 2000 ms. The response trial lasted for 3000 ms. Four long fixation periods, each lasting 19,000 ms, followed random blocks. The response button for the "consistent" option was on the left side for half of the trials in each condition.

4.3.2. Semantic processing task

Sixty-four words were selected from four semantic domains: vertical spatial position, morality, affect, and spatial enclosure. Each domain consisted of two levels (categories): up vs. down, moral vs. immoral, happy vs. unhappy, and inside vs. outside. Each category contained eight words. Thus, there were 4 domains × 2 categories/domain × 8 words/category = 64 unique words. The vertical position and morality domains were the concrete source domain and abstract target domain of interests. The affective domain was the potential confounding variable, and the enclosure domain was included as the filler. We did not choose concepts referring to other obvious spatial relations, such as those along the lateral axis (front-back or left-right) or distance, as the fillers, because those concepts had been found to associate with affective valence (Lakoff and Johnson, 1999; Meier and Robinson, 2004; Elizabeth Crawford et al., 2006; Eyal et al., 2008; Bender et al., 2020). All words were two-character nouns or flexible words with noun being one of the frequent parts of speech. Means of stroke counts of words and word frequency (Cai and Brysbaert, 2010) were matched between the two categories of each domain (Table S1). Words in the morality and affect domains were more abstract (Xu and Li, 2020) than those in the two spatial domains ($t = -7.86$, $p < 0.001$). Affective valence ratings (Xu et al., 2022) were significantly higher for moral than for immoral words ($t = 39.39$, $p < 0.001$); and significantly higher for pleasant than unpleasant words ($t = 20.93$, $p < 0.001$). Affective arousal (Xu et al., 2022) and word abstractness were respectively matched between moral and immoral words, between happy and unhappy words, between moral and happy words, and between immoral and unhappy words. Valence

was balanced between moral and happy words, and between immoral and unhappy words. Because the valence and arousal ratings for some words in the categories of *up, down, inside,* or *outside* were not available in the existing norm (Xu et al., 2022), we collected ratings for these 32 words from a separate group of 28 participants. Words in the up category were significantly more positive than those in the down category ($t = 3.26, p < 0.05$). The mean arousal or abstractness ratings did not differ significantly between the *up* and *down* categories, nor between the *inside* and *outside* categories (Table S1).

We further measured the semantic distance between different domains in a "cross-domain class-wise matched" way. For example, the distance between words in the "up" and "moral" categories and the distance between words in the "down" and "immoral" categories were concatenated to represent class-matched verticality-morality distance. Semantic distance was calculated by 1 – cosine similarity between the vectors of word pairs. The vectors were embeddings of a pre-trained word2vec model on Chinese words (https://github.com/Embedding/Chinese-Word-Vectors). The means of cross-domain distances were then pairwise compared. To our most interest, the mean of verticality-morality distance was not significantly different from the mean of verticality-valence distance (two-sample *t*-test, $t = 0.61, p = 0.54$). Thus, any performance difference between verticality-morality classification and verticality-valence classification is unlikely to be driven by the difference in the semantic similarity. Other contrasts indicate that words in the two concrete domains, verticality and enclosure, were more similar to each other than to the abstract domains, morality or valence. The abstract domains were also more similar to each other than to either abstract domain (Table S2). These results suggested that the decodability difference in the main analyses was unlikely due to the difference in the cross-domain semantic distance.

This event-related design was composed of 4 runs of fMRI scans, each beginning with a 1000 ms fixation period. In each run, a complete list of unique words was presented in a randomized order, so that each word was presented 4 times in 4 runs. The sequence and inter-trial interval were optimized using optseq2 (surfer.nmr.mgh.harvard.edu/optseq). A word trial began with a 500 ms fixation cross ("+"), followed by a 2000 ms word presentation period, during which participants were asked to think about the meaning of the word. The trial ended with another fixation cross, the duration of which was jittered (mean = 2954 ms, range = [1000 ms, 8900 ms]). Sixteen additional probe events were randomly inserted into the sequence. A response trial was presented for 3000 ms after a regular word trial. Participants pressed a button to indicate whether the word on the previous screen was pleasing or not, based on their subjective liking. After this judgment, participants fixated on the cross for 4500 ms. The response button for the "pleasing" option was on the left side of the screen for half of the trials.

4.4. MRI protocol

Magnetic resonance imaging data were collected using a Siemens Prisma 3-T scanner with a 64-channel head coil (Siemens, Erlangen, Germany). Structural images were acquired using a T1-weighted MPRAGE pulse sequence: Repetition time (TR) = 2300 ms, echo time (TE) = 2.25 ms, flip angle (FA) = 8°, acquisition matrix size = 224 × 224, field of view (FoV) = 224 × 224 mm^2, voxel size = 1 × 1 × 1 mm^3. T2*-weighted functional images were acquired using a multiband echo planar imaging (EPI) sequence: Acceleration factor = 6, TR = 1000 ms, TE = 32 ms, FA = 55°, acquisition matrix size = 96 × 96, voxel size = 2 × 2 × 2 mm^3. A pair of field maps was acquired using a gradient recalled echo sequence with TR = 413 ms, TE_{short} = 4.92 ms, TE_{long} = 7.38 ms, and flip angle = 60°.

4.5. Preprocessing

Image preprocessing was performed using SPM12 (Wellcome Department of Cognitive Neurology, London). Voxel displacement maps were calculated based on the field mapping images and applied to the EPI images for distortion correction. Functional image volumes were adjusted for slice timing difference and realigned to the first volume in each run. The T1 image of each participant was coregistered to their mean EPI image and normalized to the Montreal Neurological Institute (MNI) template. The deformation information of the normalization was then applied to normalize the functional images to the MNI template. Normalized images were spatially smoothed using a 6-mm FWHM Gaussian kernel.

4.6. Locating spatial relation processing regions

Voxelwise multiple linear regression was performed on the preprocessed images from the spatial relation processing task. For each participant, the weight image of each condition was estimated by fitting the time course with the timing function of each condition convolved with the standard hemodynamic response function. The onset time of each spatial relation block was the onset of the first word in that block. The duration of a block was the duration for presenting the five words, namely 10 s. The six rigid-body parameters of head motion were modeled as covariates. The individual contrast maps of above > below and below > above were calculated using independent samples *t*-tests and respectively submitted to the group-level one-sample *t*-test against 0. The group-level contrast maps were thresholded at a cluster-wise corrected α of 0.05 using the AlphaSim algorithm implemented in NeuroElf (https://neuroelf.net/). Significance of the clusters was determined jointly by the voxel-wise p of 0.05 and the minimum cluster size determined by a 2000-iteration simulation. The resulting map was used as the mask of regions associated with the spatial relation processing.

4.7. Estimate of single-trial response of semantic processing

Subject-specific responses to each word trial in the semantic processing task were estimated using multiple linear regressions using SPM. Each trial was estimated in one model, in which one regressor represented the target trial and all the rest of the trials within the same run were estimated as another regressor (Mumford et al., 2012). The six parameters of head motion were modeled as covariates. The resulting weight image of the target trial was used as the response per trial per participant. Further analyses were performed using in-house scripts on MATLAB R2022a (Mathworks, MA, USA). A gray matter template was applied to mask out the non-gray-matter voxels.

4.8. Within-domain decoding

The within-domain and cross-domain decoding was applied in multiple analyses. The following two sections described the common procedures. The methods that were specific to each analysis were described along with the corresponding results.

The classifications and statistical tests were performed for each individual participant. The within-domain decoding trained and tested classifiers on concepts within the same concept domain, i.e., within the domain of vertical spatial position, morality, affect, or spatial enclosure, in a leave-one-word-out cross-validation protocol. In each cross-validation fold, the test data were the mean neural response image over the four trials of a single word. A logistic regression classifier (Bishop, 2006) was trained on the multivoxel patterns of the rest 60 trials of 15 words. The target variable was the category of the word trial (e.g., *up* or *down* for vertical spatial concepts). The parameters were estimated using maximum likelihood estimation. This procedure was iterated until each word had been tested once. The accuracies of identifying labels for the test data across all folds were averaged to indicate the overall performance of each type of classification.

To evaluate the performance of each type of classification, a corresponding null distribution of accuracy was generated by a 2000-

iteration random permutation. In each iteration, all the training and testing procedures were the same as the main analysis, except that the labels of the entire set of data were randomly shuffled beforehand. The density function of the null distribution was estimated based on the permuted accuracies. The statistical significance was estimated by the p-value of the actual accuracy under the corresponding null distribution.

4.9. Cross-domain decoding

Logistic regression classifiers were trained on neural responses to words in one domain (e.g., *up* and *down*) and tested on the data from another domain (e.g., *moral* and *immoral*). The test data were the mean neural response image over the four trials of a single word. The cross-domain label was assigned based on the hypothesized conceptual metaphorical semantics: the *up, moral,* or *happy* concepts were of the same label, which differed from the label of the *down, immoral,* or *unhappy* concepts that were labeled as the same category. For any interested pair of domains, each domain served as the training data in one classification and as the test data in the other. The accuracies were averaged across the two directions of classifications.

For classifications performed by searchlight, the radius of the isotropic searchlight was 8 mm (3 voxels). Classification was performed based on the multivoxel patterns in each searchlight sphere. The searchlight moved with a step length of 1 voxel through all the gray matter voxels.

The performance of each type of classification was evaluated by the corresponding permutation test, in which the label of test data was randomly shuffled for 2000 iterations. To obtain the significance map for the searchlight analysis, p-values were FDR-corrected for the multiple comparisons across voxels. To obtain the significance map at the group level, each voxel was thresholded using the mean of the actual accuracy across participants against the mean accuracy of the permuted models.

4.10. Region-wise aggregation of the individual searchlight decoding results

The results related to this section were displayed in Figs. 3c-3h. To characterize the cross-domain representation at the regional level, the voxel-wise accuracies of the searchlight analysis were integrated within each region defined by the atlas AICHA v2 (Joliot et al., 2015). The accuracy map of each participant was first individually thresholded using the same permutation test procedure as the group-level one, but using the participant-specific permutation data. For the resulting accuracy maps, the number of significant voxels in each AICHA region was then averaged across participants. This number of voxels was tested against an empirically generated null distribution over 2000 iterations. The null hypothesis was that given the total number of significant voxels across the brain, the actual number of significant voxels in a given region was not greater than the number when voxels were randomly located over the brain.

4.11. Information-theoretic measures

This section described the methods specific to the results in Fig. 6. The goal was to examine whether non-metaphorical voxels provided information additional to the metaphorical voxels in terms of classifying moral vs. immoral concepts. The baseline model was the within-domain classification of morality built on all the voxels with significant accuracy in the vertical-moral classifications (locations shown in Fig. 3a). Sixteen additional sets of classifications were performed using the same procedure with increasingly more extra voxels. These voxels were selected based on their representational stability over stimulus presentations. Pearson's correlation of the activation profile over pairs of presentations (a complete list of stimulus words) indicated the stability of a voxel's response to word concepts. The pairwise correlations were averaged to represent the stability for each voxel. The voxels were gradually added to the classification models in a descending order of their stability scores. The number of extra voxels being added each time was set to the rounded values of $400 \times 1.5^{n-1}$ for simplicity, where n was a natural number ranging from 1 to 16.

For each classification model, the predicted probability distributions over the two classes of each test exemplar were compared against the probability distributions of the baseline model by the measures of Kullback–Leibler divergence (KLD) and the entropy reduction. The KLD quantified the information gain on morality representation provided by the non-vertical-metaphorical voxels:

$$D(p \parallel q) = -\sum_{c \in C} p_c * log_2\left(\frac{p_c}{q_c}\right) \quad (1)$$

Where p was the predicted probability of the classification built on the metaphorical voxels, q was the predicted probability of the classification when each set of additional voxels was included, and c indicated the classes. The KLD was computed for each classification and summed over all the cross-validation folds.

The entropy reduction (ER) measured the model's decrease of uncertainty, or confidence, in the prediction. The entropy of each model was calculated as

$$H = -\sum_{c \in C} p_c * log_2(p_c) \quad (2)$$

And ER was

$$ER(n) = H(0) - H(n) \quad (3)$$

where $H(0)$ was the entropy of the baseline model and $H(n)$ was the entropy of nth extra model. A larger ER value indicated greater confidence as compared to the baseline.

Data and code availability

The neuroimaging data and the code are deposited at https://osf.io/7egaz/. Any additional information required to analyze the data reported in the manuscript is available from the lead contact upon reasonable request.

CRediT authorship contribution statement

Jing Wang: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Miao Qian:** Writing – review & editing, Validation, Software, Investigation, Conceptualization, Formal analysis. **Qing Cai:** Writing – review & editing, Supervision, Resources, Conceptualization.

Declaration of competing interest

Authors report no conflict of interest.

Acknowledgements

We thank Ziyi Ding and Ming Song for their helpful discussion. Funding Information: This work was funded by the National Natural Science Foundation of China (32100857 to JW).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2025.121485.

References

Bao, X., Gjorgieva, E., Shanahan, L.K., Howard, J.D., Kahnt, T., Gottfried, J.A., 2019. Grid-like neural representations support olfactory navigation of a two-dimensional odor space. Neuron 102, 1066–1075.e5.

Barsalou, L.W., 2008. Grounded Cognition. Annu Rev Psychol 59, 617–645.

Bender, A., Teige-Mocigemba, S., Rothe-Wulf, A., Seel, M., Beller, S., 2020. Being in front is good—but where is in front? Preferences for spatial referencing affect evaluation. Cogn Sci 44, e12840.

Bi, Y., 2021. Dual coding of knowledge in the human brain. Trends Cogn. Sci. (Regul. Ed.) 25, 883–895.

Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cerebral Cortex 19, 2767–2796.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, New York, NY, USA.

Bo, K., Kraynak, T.E., Kwon, M., Sun, M., Gianaros, P.J., Wager, T.D., 2024. A systems identification approach using Bayes factors to deconstruct the brain bases of emotion regulation. Nat. Neurosci. 27, 975–987.

Borghi, A.M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., Tummolini, L., 2017. The challenge of abstract concepts. Psychol Bull 143, 263–292.

Bottini, R., Doeller, C.F., 2020. Knowledge across reference frames: cognitive maps and image spaces. Trends Cogn. Sci. (Regul. Ed.) 24, 606–619.

Brysbaert, M., Cortese, M.J., 2011. Do the effects of subjective frequency and age of acquisition survive better word frequency norms? Quarter. J. Experiment. Psychology 64, 545–559.

Brysbaert, M., Warriner, A.B., Kuperman, V., 2014. Concreteness ratings for 40 thousand generally known english word lemmas. Behav Res Method. 46, 904–911.

Cai, Q., Brysbaert, M., 2010. SUBTLEX-CH: chinese word and character frequencies based on film subtitles. PLoS One 5, e10729.

Caucheteux, C., Gramfort, A., King, J.-R., 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. Nature Human Behaviour. 7, 430–441.

Chikazoe, J., Lee, D.H., Kriegeskorte, N., Anderson, A.K., 2014. Population coding of affect across stimuli, modalities and individuals. Nat. Neurosci. 17, 1114–1122.

Clark, K.M., 2024. Embodied imagination: lakoff and Johnson's experientialist view of conceptual understanding. Rev. General Psychol. 28, 166–183.

Constantinescu, A.O., O'Reilly, J.X., Behrens, T.E.J., 2016. Organizing conceptual knowledge in humans with a gridlike code. Science 352, 1464–1468.

Dahl, C., Adachi, I., 2013. Conceptual metaphorical mapping in chimpanzees (Pan troglodytes). eLife 2.

Desai, R.H., 2022. Are metaphors embodied? The neural evidence. Psychol. Res. 86, 2417–2433.

Elizabeth Crawford, L., Margolies, S.M., Drake, J.T., Murphy, M.E., 2006. Affect biases memory of location: evidence for the spatial representation of affect. Cognit. Emot. 20, 1153–1169.

Eyal, T., Liberman, N., Trope, Y., 2008. Judging near and distant virtue and vice. J. Exp. Soc. Psychol. 44, 1204–1209.

Frank, M.C., Braginsky, M., Yurovsky, D., Marchman, V.A., 2017. Wordbank: an open repository for developmental vocabulary data. J. Child. Lang. 44, 677–694.

Gennari, G., Dehaene, S., Valera, C., Dehaene-Lambertz, G., 2023. Spontaneous supra-modal encoding of number in the infant brain. Curr. Biol. 33, 1906–1915.e6.

Gibbs, R., 2011. Evaluating conceptual metaphor theory. Discourse Proc. 48, 529–562.

Goldstein, A., et al., 2022. Shared computational principles for language processing in humans and deep language models. Nat. Neurosci. 25, 369–380.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., de Lange, F.P., 2022. A hierarchy of linguistic predictions during natural language comprehension. Proceed. Nation. Acad. Scienc. 119, e2201968119.

Abdi, Hervé, O'Toole, A.J., Valentin, D., Edelman, B., 2005. DISTATIS: the analysis of multiple distance matrices. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops, 42–42.

Hill, P.L., Lapsley, D.K., 2009. The ups and downs of the moral personality: why it's not so black and white. J. Res. Pers. 43, 520–523.

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453.

Janczyk, M., Koch, I., Ulrich, R., 2023. Is there a cognitive link between the domains of deictic time and number? J. Experiment. Psychol. 49, 493–507.

Joliot, M., Jobard, G., Naveau, M., Delcroix, N., Petit, L., Zago, L., Crivello, F., Mellet, E., Mazoyer, B., Tzourio-Mazoyer, N., 2015. AICHA: an atlas of intrinsic connectivity of homotopic areas. J. Neurosci. Methods 254, 46–59.

Keysar, B., Shen, Y., Glucksberg, S., Horton, W.S., 2000. Conventional language: how metaphorical is it? J. Mem. Lang. 43, 576–593.

Kragel, P.A., Treadway, M.T., Admon, R., Pizzagalli, D.A., Hahn, E.C., 2023. A mesocorticolimbic signature of pleasure in the human brain. Nat. Human Behaviour. 7, 1332–1343.

Lakoff, G., Johnson, M., 1980. Metaphors We Live By. University of Chicago Press, Chicago.

Lakoff, G., Johnson, M., 1999. Philosophy in the flesh: the embodied mind and its challenge to western thought. Basic Books, New York.

Lin, Y., Oyserman, D., 2021. Upright and honorable: people use space to understand honor, affecting choice and perception. Pers. Soc. Psychol. Bull. 47, 3–19.

Luyckx, F., Nili, H., Spitzer, B., Summerfield, C., 2019. Neural structure mapping in human probabilistic reward learning Lee D, Gold JI, Lee D, Chafee M, eds. eLife 8, e42816.

Meier, B.P., Hauser, D., Robinson, M., Friesen, C., Schjeldahl, K., 2007a. What's "up" with god? Vertical space as a representation of the divine. J. Pers Soc. Psychol. 93, 699–710.

Meier, B.P., Robinson, M.D., 2004. Why the sunny side is up: associations between affect and vertical position. Psychol. Sci. 15, 243–247.

Meier, B.P., Sellbom, M., Wygant, D.B., 2007b. Failing to take the moral high ground: psychopathy and the vertical representation of morality. Pers. Individ. Dif. 43, 757–767.

Meng, X., Nakawake, Y., Nitta, H., Hashiya, K., Moriguchi, Y., 2019. Space and rank: infants expect agents in higher position to be socially dominant. Proceed. Royal Soc. B. 286, 20191674.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.-M., Malave, V.L., a, Mason R, Just, M.A., 2008. Predicting human brain activity associated with the meanings of nouns. Science 320, 1191–1195.

Mumford, J.A., Turner, B.O., Ashby, F.G., Poldrack, R.A., 2012. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. NeuroImage 59, 2636–2643.

Munuera, J., Rigotti, M., Salzman, C.D., 2018. Shared neural coding for social hierarchy and reward value in primate amygdala. Nat. Neurosci. 21, 415–423.

Paivio, A., 1986. Mental representations: a dual coding approach. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195066661.001.0001. Available at:

Paivio, A., 1991. Dual coding theory: retrospect and current status. Can. J. Psychol. 45, 255–287.

Park, S.A., Miller, D.S., Boorman, E.D., 2021. Inferences on a multidimensional social hierarchy use a grid-like code. Nat. Neurosci. 24, 1292–1301.

Park, S.-E., Lee, J., Lee, S.A., 2023. Domain-general and domain-specific electrophysiological markers of cognitive distance coding for what, where, and when memory retrieval. J Neurosci 43, 4304.

Parkinson, C., Liu, S., Wheatley, T., 2014. A common cortical metric for spatial, temporal, and social distance. J. Neurosci. 34, 1979–1987.

Peer, M., Salomon, R., Goldberg, I., Blanke, O., Arzy, S., 2015. Brain system for mental orientation in space, time, and person. Proc. Natl. Acad. Sci. U.S.A. 112, 11072–11077.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S.J., Kanwisher, N., Botvinick, M., Fedorenko, E., 2018. Toward a universal decoder of linguistic meaning from brain activation. Nat. Commun. 9.

Pinker, S., 2007. The Language Instinct. Harper Perennial Modern Classics, New York, NY.

Riemer, M., Achtzehn, J., Kuehn, E., Wolbers, T., 2022. Cross-dimensional interference between time and distance during spatial navigation is mediated by speed representations in intraparietal sulcus and area hMT+. NeuroImage 257, 119336.

Saxe, R., Kanwisher, N., 2003. People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind. NeuroImage 19, 1835–1842.

Schurz, M., Radua, J., Tholen, M.G., Maliske, L., Margulies, D.S., Mars, R.B., Sallet, J., Kanske, P., 2021. Toward a hierarchical model of social cognition: a neuroimaging meta-analysis and integrative review of empathy and theory of mind. Psychol. Bull. 147, 293–327.

Sheahan, H., Luyckx, F., Nelli, S., Teupe, C., Summerfield, C., 2021. Neural state space alignment for magnitude generalization in humans and recurrent networks. Neuron 109, 1214–1226.e8.

Su, Y., Li, Y., Li, H., 2023. Familiarity ratings for 24,325 simplified chinese words. Behav. Res. Meth. 55, 1496–1509.

Tang, J., LeBel, A., Jain, S., Huth, A.G., 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. Nat. Neurosci. 26, 858–866.

Theves, S., Fernández, G., Doeller, C.F., 2020. The Hippocampus maps concept space, not feature space. J. Neurosci. 40, 7318.

Viganò, S., Bayramova, R., Doeller, C.F., Bottini, R., 2023. Mental search of concepts is supported by egocentric vector representations and restructured grid maps. Nat. Commun. 14, 8132.

Viganò, S., Rubino, V., Soccio, A.D., Buiatti, M., Piazza, M., 2021. Grid-like and distance codes for representing word meaning in the human brain. NeuroImage 232, 117876.

Vodrahalli, K., Chen, P.-H., Liang, Y., Baldassano, C., Chen, J., Yong, E., Honey, C., Hasson, U., Ramadge, P., Norman, K.A., Arora, S., 2018. Mapping between fMRI responses to movies and their natural language annotations. NeuroImage 180, 223–231.

Walsh, V., 2003. A theory of magnitude: common cortical metrics of time, space and quantity. Trends Cogn. Sci. (Regul. Ed.) 7, 483–488.

Wang, J., Cherkassky, V.L., Just, M.A., 2017. Predicting the brain activation pattern associated with the propositional content of a sentence: modeling neural representations of events and states. Hum Brain Mapp 38, 4865–4881.

Wang, J., Conder, J.A., Blitzer, D.N., Shinkareva, S.V., 2010. Neural representation of abstract and concrete concepts: a meta-analysis of neuroimaging studies. Hum. Brain Mapp. 31, 1459–1468.

Wang, X., Bi, Y., 2021. Idiosyncratic tower of Babel: individual differences in word-meaning representation increase as word abstractness increases. Psychol. Sci. 32, 1617–1635.

Wang, X., Men, W., Gao, J., Caramazza, A., Bi, Y., 2020. Two forms of knowledge representations in the Human brain. Neuron 107, 383–393.e5.

Wang, X., Wang, B., Bi, Y., 2023. Early language exposure affects neural mechanisms of semantic representations Peelle JE, de Lange FP, Reilly J, Fairhall SL, eds. eLife 12, e81681.

Wang, X., Wu, W., Ling, Z., Xu, Y., Fang, Y., Wang, X., Binder, J.R., Men, W., Gao, J.-H., Bi, Y., 2018. Organizational principles of abstract words in the human brain. Cerebral Cortex. 28, 4305–4318.

Whittington, J.C.R., McCaffary, D., Bakermans, J.J.W., Behrens, T.E.J., 2022. How to build a cognitive map. Nat. Neurosci. 25, 1257–1272.

Xu, X., Li, J., 2020. Concreteness/abstractness ratings for two-character Chinese words in MELD-SCH. Plos One 15, e0232133.
Xu, X., Li, J., Chen, H., 2022. Valence and arousal ratings for 11,310 simplified chinese words. Behav Res. Meth. 54, 26–41.
Yao, Z., Wu, J., Zhang, Y., Wang, Z., 2017. Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1100 Chinese words. Behav. Res. Meth. 49, 1374–1385.
Zhang, L., Atari, M., Schwarz, N., Newman, E.J., Afhami, R., 2022. Conceptual metaphors, processing fluency, and aesthetic preference. J. Exp. Soc. Psychol. 98, 104247.