Bipartite functional fractionation within the neural system for social cognition supports the psychological continuity of self *vs.* other

Rocco Chiou[1] ✉, Christopher R. Cox[2], & Matthew A. Lambon Ralph[1] ✉
[1] MRC Cognition & Brain Sciences Unit, University of Cambridge, UK
[2] Department of Psychology, Louisiana State University, USA

✉ Correspondence to this work can be addressed to either of the authors by email:
(**RC:** Rocco.Chiou@mrc-cbu.cam.ac.uk, **MALR:** Matt.Lambon-Ralph@mrc-cbu.cam.ac.uk).

Abstract:	225 words
Significance:	120 words
Main text:	6,992 words

ORCID: 0000-0002-3009-8075 (RC); 0000-0001-8287-3865 (CRC); 0000-0001-5907-2488 (MALR)

Acknowledgements: This research was funded by a Sir Henry Wellcome Fellowship (201381/Z/16/Z) to RC, and an MRC programme grant and intramural funding to MALR (MR/R023883/1; MC_UU_00005/18).

Conflict of interest: The authors declare no competing financial interests.

Abstract

In the present study, we integrated two crucial issues in cognitive neuroscience: (*i*) The field of brain connectome research has been characterising the fine-grained fractionation within of each identified functional network. (*ii*) The field of social neuroscience has been exploring how a network of social-affective cortical zones implement the representational codes to idiosyncratically represent personal identities. Here we performed a series of multivoxel decoding to investigate the neural representations of self- *vs.* other-referential concept, a pivotal psychological construct that permeates every aspects of social life and has been argued as a core function of the specialised system for social cognition. Across multiple analyses, we found a robust dyadic fractionation between the two subnetworks within the social system – multivoxel patterns reliably dissociated between regions that are affiliated with the 'default mode' network and those with the 'semantic cognition' network, evident both in the outcomes of supervised classification and representational similarity. Moreover, representational configuration of the neural responses to self- and other-referential thoughts mirrored the psychological continuum of social distance, from an integral sense of one's present self to a distant other, with the broad-stroke distinction of 'self-ness vs. otherness' being the principal axis in the representational space, and social distance being the auxiliary axis. Taken together, our findings inform the burgeoning endeavour of connectome research about brain's network architecture and how it impacts on social cognition.

Significance

The neural representations whereby the brain employs to differentiate self from other is complicated by interpersonal distance (compared to a distant other, a personally familiar other may be deemed more similar to self) and by the evolution of self-concept over time (our sense of past-self may be distant from present-self). This is further complicated by the heterogeneity within the brain's social system that its different subregions may contribute differentially. We untangled this complexity by demonstrating that (*i*) the distinction of self *vs.* other was a dominant axis in the representational space while interpersonal distance was the secondary axes; (*ii*) within the social brain, the default-mode network dissociated from the semantic network, with the former carrying more information about personal identities.

Introduction

Decades of research suggests that humans deduce other individuals' mental states by simulating how oneself would think and feel in similar situations (1). This ability ('theory of mind') is rooted in the fundamental awareness that self and others are distinct social beings. The rudimentary sense of 'self *vs.* other' emerges during infancy, while self-concept (a sophisticated understanding of ourselves in relation to others under social contexts) evolves over lifespan. Neuroimaging evidence (for review, see 2) has established that the brain's default-mode network (DN) is robustly engaged by various self-referential processes. This leads to the view that the DN is the key neural substrate that underpins the 'core self'/'ego' (3). Rather than a homogenous structure, converging results from seed-based connectivity, data-driven parcellation, and task-induced activation have further indicated that the DN comprises functionally distinct and anatomically separable subnetworks (e.g., 4, 5, 6). The division within the DN into subnetworks begs an important question – whether and how each subnetwork differentially contributes to the representation of self. In the present study, we investigated this issue by testing how the psychological continuity from 'self-ness' to 'otherness' is encoded in different DN subnetworks using multiple-voxel pattern analysis (MVPA). Our findings unravelled a systematic, graded pattern wherein the distributed neural coding for different social beings varied between brain regions and psychological constructs. Below we first discuss the neural basis of self-referential processes and the limitation of univariate approach. Next, we discuss evidence for a bipartite fractionation within the DN. We specifically focus on how such a fractionation constrains our scope of scrutiny for functional neural substrates, and how we can exploit MVPA to investigate the neural coding for 'self *vs.* other' within this bipartite structure.

Self-concept is a collection of beliefs about oneself, embodying the contents of contemplation about '*Who am I?*' (7). It is inextricably intertwined with many critical issues in cognitive and social psychology, ranging from perception of the aesthetics and attractiveness of one's own body (self-image), the capability for reflecting on one's own attitudes and dispositions (self-knowledge), to one's recollection of life experiences (autobiographical memory). Previous research has shown that self-concept depends, in a fundamental way, upon one's capacity to represent self as a psychologically coherent entity persisting through time, whose *past self* is construed as an entity closely related to yet partially separable from the *present self* (8). In addition to representing *present self* as a continuation of *past self*, research has shown that self-concept is impacted by the presence of other people in a social environment. Especially, during childhood and adolescence, self-concept has been found to be formulated through interaction with significant others or primary caregivers, which spawned numerous investigations of familial and peer influences on personality

and self-esteem (e.g., 9, 10). The significance of self-related processing is not only seen in the field of social psychology but also in cognitive neuroscience. Researchers have long understood that self-related processing is one of the major neurocognitive dimensions that characterise the functionality of the cortical midline structures, particularly the ventromedial prefrontal cortex (vmPFC) (for review, see 2). For example, compared to information regarding another individual, processing information with reference to self amplifies vmPFC activity. This has been robustly observed when one evaluates descriptions with respect to self (11), recalls personal experiences (12), or even recognises one's own name/address that is ostensibly unrelated to the task at hand (13). In addition, patients with vmPFC lesion fail to show self-referential memory advantage (superior mnemonic performance for information related to self that healthy people reliably show), despite those patients having otherwise intact memory for non-self information (14). Research also indicates that the vmPFC is sensitive to social distance with regard to self, with friends eliciting greater vmPFC activity than strangers, presumably due to one's construal of a 'friend' being inextricably linked to oneself (15). In a similar vein, vmPFC activity was found to reflect whether someone's political opinion is in agreement with oneself, with encountering someone holding concordant views eliciting greater vmPFC activity than someone holding opposite views (16). Taken together, multiple threads of evidence converge on the conclusion that the vmPFC is a key neural structure underlying self-related processing, and is involved in processing other individuals when they are deemed similar/related to self.

The univariate approach is useful for identifying the loci of maximally-responding clusters/regions for a certain cognitive process, giving researchers a bird's eye view about where the 'hotspot' is. In the case of self-related processing, researchers often adopt the 'cognitive subtraction' paradigm – the univariate response of each voxel (or averaged response within a region of interest) elicited by self-related processing is compared against the response driven by evaluating, imagining, or reminiscing about someone else. While this approach is useful in identifying conspicuous targets in the brain (i.e., contiguous voxels that form a cluster surpassing thresholds), it has limited use in unveiling nuances that are jointly encoded by distributed patterns of neural responses across voxels. With MVPA, however, researchers are able to reveal how distributed neural patterns represent the subtle distinction between self- and other-referential processes. For example, the patterns of vmPFC activity have been used to decode if one was entertaining self- or other-referential thoughts (e.g., 17, 18). In addition to deciphering the distinction of self *vs.* other using vmPFC patterns, MVPA has been applied to decode other facets of social cognition using neural patterns elsewhere in the brain (further discussion below), such as decoding the identities of fictitious characters (19), the identities of personally familiar people (20), and whether a behaviour shows goodwill or malice

(21). Together, these studies demonstrate how MVPA can be used to elucidate our understanding about the neural basis of social cognition, capturing the granularity in collective, distributed codes that often goes unnoticed by the univariate analysis (for review of recent progress, see 22).

While the vmPFC is heavily studied, it is not the only region involved in self-related processing. Alongside vmPFC activity, a group of widely distributed brain regions, which collectively form the brain's specialised system for social cognition, have been reliably found to be involved in various self-related/socio-cognitive processes (for review, see 23). For instance, vmPFC activity usually elevates in tandem with activity of other midline structures – the posterior cingulate cortex (PCC), retrosplenial cortex (RSC), and dorsomedial prefrontal cortex (dmPFC) – when the task requires assessing personality-traits about self/other, retrieving autobiographical episodes, or envisaging prospective and counterfactual scenarios that involve human activity. Apart from the brain's medial aspect, self-referential/social-cognitive tasks also recruit a set of regions on the brain's lateral surface (6, 23-25), including the bilateral anterior temporal lobes (ATL), the left inferior frontal gyrus (IFG), and the inferior parietal lobule (IPL). Resting-state research has shown that these areas, conjointly as the system for social cognition, tend to stay intrinsically connected during task-free resting moments, and overlap substantially with the cortical realms of the DN. Regions implicated in social cognition exhibit a similar functional profile that has been adopted to characterise the tendencies of DN regions (26, 27): Akin to the DN, social regions are more active when a situation necessitates integrating internally-constructed representations (e.g., memories or schemas) with external signals (e.g., 28). Also resembling the DN, activation of the social system abates when a situation requires externally-oriented sensorimotor processes that minimally involve any internal representations (for empirical results, see 6, for review of related evidence, see 29). Interestingly, converging evidence from separate investigations have shown that the extensive DN can be (at least) fractionated into two subsystems (4-6, 30, 31). While the exact nomenclature varies between studies, the two subsystems of the broad default-mode/social-cognitive network exhibit distinct functional profiles. One subsystem comprises the vmPFC, posterior sections of the midline structures (PCC and RSC), and bilateral IPL. These areas collectively are suggested to be the core/hub nodes of the DN (32) and referred to as 'Network-A' by more recent studies (e.g., 5). These core DN regions show a strong propensity to be *suppressed* by contexts that require externally-focused sensory-motoric processes, and *activated* by contexts that encourage introspective processes. The other subsystem consists of primarily the bilateral ATL and left IFG. These regions are situated within the realm of so-called 'Network-B', and also referred to as the semantic network (SN; owing to their robust participation in semantic tasks). DN and SN regions show a similar inclination to deactivate in outwardly-oriented sensorimotor tasks, although the

extent of 'aversion' to sensorimotor processing is significantly more moderate in SN regions (6). Crucially, the two subnetworks diverge on reaction to semantic processing – whereas SN activity intensifies for lexical-semantics, DN activity attenuates (30, 31, 33). Such dissociation is observed regardless of the methods used to define the scope of DN and SN – be it task-based contrasts or task-free resting-state connectivity, suggesting that, rather than a sheer methodological artefact, the two subnetworks are two functionally and anatomically distinct entities. It is noteworthy that there is less consensus on the taxonomy of dmPFC. The dmPFC shares similar functional profiles during socio-cognitive tasks to other DN regions (34), and is reliably linked with other core nodes of the DN, particularly the vmPFC and PCC/RSC (e.g., 35, 36, 37). However, the dmPFC is also affiliated with the SN – relative to the core DN nodes, the dmPFC has tighter functional coupling with SN regions in resting-state (38), and is more reactive to tasks that lay emphasis on extracting semantic meaning from lexical stimuli than other DN regions (6). Pertinent to our main question, while the DN and SN dissociate on their coarse-grained/univariate reaction to different contexts, it remains unclear whether the two subnetworks carry distinct fine-grained multiple-voxel patterns that convey different facets/amount of information about 'self *vs.* other'.

In the present study, we investigated the distributed neural coding the underlie the continuity of 'self *vs*. other' representations. Self-concept is known to evolve along one's lifetime. However, although one's beliefs about *present self* might be radically different from his/her opinions about *past self*, the human mind maintains the stability/continuity of self-identity, treating *present self* and *past self* as dissociable yet related entities that represent the same person (39). This continuity is applicable to the construal of self *vs.* others, with a close other (or personally familiar person) perceived as more affiliated with/ similar to self than a distant other (personally unrelated person). Such continuity is fundamental to social cognition; however, it remains unknown how it is encoded in the brain. We instigated this issue by gradationally manipulating social distance. We established four points of reference: the participant's sense of *present self*, the participant's sense of *past self* ten years ago, a personally familiar other – the participant's mother, and a personally unfamiliar but well-known other – Queen Elizabeth II. Behavioural testing confirmed that participants rated these 4 identities along a continuum, with *present self* being most distinct from *the Queen* with *past self* and *mother* situate serially somewhere in between. While undergoing fMRI, participants read personality descriptions (either positive or negative personality traits) in contexts of each of the four individuals and judged whether the description aptly characterises the person. We used two types of MVPA, machine-learning classification (40) and representational similarity analysis (41), to investigate how the two subnetworks encode the psychological continuity of self *vs.* other and the possible divergence between DN and SN (see *SI Appendix* for details of the paradigm).

Results

Machine-learning classification analysis. Foremost amongst the goals of our study is to examine whether the multivoxel patterns of DN and SN allowed deciphering the continuity and boundary of 'self *vs.* other' representations. We aimed to clarify (*i*) whether there is a gradient-like change in neural coding in relation to social distance and (*ii*) whether or not the two subnetworks were equally capable of representing such information. To localise regions of interest (ROI) unbiasedly in each participant's native-space brain, we used a localiser task to independently identify eight targeted areas sensitive to social cognition and semantic comprehension (see *Methods*). Based on the peak-coordinates derived from the localiser run, we defined eight equally-sized ROIs within the DN and SN – five ROIs are typically affiliated with the DN (the dmPFC, the vmPFC, the PCC, the left IPL, the right IPL), while three are associated with the SN (the left ATL, the right ATL, the left IFG). Support-vector-machine (SVM) classifiers were trained on the multivoxel patterns of each ROI. Using a standard supervised-learning cross-validation procedure, the algorithms were tested using 'quarantined' data to evaluate whether it was able to predict the mental content (i.e., a particular person one was reflecting on), and whether the chance of successful prediction varied systematically with social distance and network membership.

Results revealed that, in nearly all of the ROIs situated in both the default and semantic networks, the algorithm was able to extract the neural coding of 'self *vs.* other' from regional activity and made prediction. Adjustments for multiple comparisons were applied to every analysis by restraining false-discovery-rate (FDR) below 5% (we also used the stricter Bonferroni correction, which yielded entirely consistent results with FDR). First, we tested whether we could detect broad-stroke information about the aggregate sense of self-ness (*Present Self* and *Past Self*) *vs*. the aggregate sense of otherness (*Mother* and *Queen*). Significantly above-chance decoding was achieved in every ROI (Figure 1A), suggesting that this coarse-grained, binary distinction between self and other is discernible using the patterns of local DN/SN activity. Motivated by this result, next we investigated whether we could detect subtler distinction within the orbit of self-concepts (*Present Self vs. Past Self*) and within the other-concepts (*Mother vs. Queen*). As shown in Figure 1A, statistically reliable decoding was achieved in nearly all ROIs (all, except for the left IFG) for the differentiation between *Present Self* and *Past Self*, and in every ROI for the differentiation between *Mother* and *Queen*. This indicates that regional patterns of DN and SN activity enabled accurate classification not only *between* the domains of self *vs.* other, but also the finer-grained subgroups *within* the realm of self- and other-concepts. Successful decoding of socially proximal (*Present Self*) *vs.* distant concept (*Past Self*) within the 'self' domain suggests that information

about ‘social distance’ intersected with ‘self *vs.* other’. To untangle this intricacy, we performed four-way classification amongst the four individuals to understand how decoding fared when the algorithm had to consider ‘self *vs.* other’ and ‘social distance’ simultaneously. As shown in Figure 1A, while the chance-level dropped from 50% (binary) to 25% (four-way), reliable above-chance decoding was achieved in every ROI (also see *SI Appendix* for the results of confusion-matrix), indicating that information about each specific individual’s identity can be reliably differentiated from the remainder. Critically, a systematic trend was replicated across all of the four analyses: Although significantly above-chance decoding was obtained in every ROI, classification accuracy was reliably higher in the five ROIs that belong to the DN (with the PCC reliably reaching the highest accuracy), compared to the three ROIs that belong the SN. This ‘imbalance’ between the two subnetworks implies that the DN, as a while, carried greater amount of information about personal identities that could be extracted by the classifier to enable correct predictions.

After ascertaining that decoding was possible using information within the ROIs, we performed multiple whole-brain searchlight analyses to scrutinise whether the expansive swathe of regions outside the ROIs also carried information about the person-related concepts. In separate analysis, the algorithm was trained to classify (*i*) the difference between composite class of self-ness and otherness, (*ii*) within the ‘self’ domain – *Present vs. Past Self*, (*iii*) within the ‘otherness’ domain – *Mother vs. Queen*, (*iv*) four-way classification amongst all categories. Statistics was thresholded stringently using FWE-correction at $p < 0.05$ for whole-brain voxels. Results of the searchlight interrogations were highly coherent across the four analyses and revealed striking consistency with the ROI findings. Example results are illustrated in Figure 1B (*Present Self vs. Past Self*) and 1C (*Mother vs. Queen*). Above-chance classification was detected in broad expanses of the DN, including much of the rostro-medial prefrontal cortex (both the dmPFC and vmPFC), the PCC and its surrounding regions (dorsal: the precuneus; ventral: the retrosplenial cortex/RSC), and the bilateral IPL (encompassing both its anterior sector/temporoparietal junction and its posterior sector/angular gyrus). In addition to these three core clusters within the DN, successful decoding was found in the bilateral dorsolateral prefrontal cortices. The clusters of successful decoding expanded selectively to regions of the SN (i.e., the bilateral ATLs and the left IFG) when we applied a slightly more lenient threshold ($p < 0.001$ for voxel intensity, $p < 0.05$ FWE-correction for cluster-level multiple comparisons). Together, these data concur with the ROI results that regions of the DN contained overall more information than those of the SN, evident in their differential levels of successful decoding rate.

While the standard cross-validation approach unravelled whether the information of self *vs.* other existed in the multivoxel pattern of an ROI (e.g., whether the classifier succeeded in predicting if one was reflecting on *Mother* or *Queen* based on distributed activities), it lacked the ability to test whether the brain reinstates a reproducible and generalisable neural code across varying situations. A generalisable code indicates *abstraction* despite contextual variation. A hypothetical example that epitomises this generalisability and invariance would be that the brain summons reproducible neural patterns to represent one's identity no matter whether that person is seen, heard, or recalled. To overcome the limitation of standard cross-validation, we conducted cross-classification – the classifier was trained on data from one cognitive context and tested on data from another; with this procedure, we tested whether there was commonality in underlying neural codes that was invariant to varying experimental contexts. Two analyses were performed: First, we tested whether there is generalisable neural coding of self *vs.* other, irrespective of social distance. This was achieved by training the algorithm to classify *Present Self* from *Mother* (closer distance) and testing whether it succeeded in applying the knowledge to distinguish *Past Self* from *Queen* (farther distance), and *vice versa*. Second, we tested whether generalisable/reproducible patterns of neural representations were used to represent the four individuals, irrespective of the emotive valence of personality traits that were used to probe these person-related concepts. By virtue of our factorial design that crossed individual identities and emotive valence (see *Methods*), this cross-classification was achieved by training the algorithm to classify the four persons using data from the positive-valence context and testing the generalisability of the learnt patterns using data from the negative-valence context (and *vice versa*). Cross-classification between contexts provides a rigorous test to assay whether invariant/generalisable neural codes were reliably summoned to represent the mental concepts about particular individuals, unaltered by social distance or emotive valence.

As illustrated in Figure 2A, the analysis showed that, based on local neural patterns of four regions of the DN (the dmPFC, vmPFC, PCC, and left IPL), the classifier was able to extrapolate information (e.g., the neural codes elicited by the pair of near distance – *Present Self vs. Mother*) from one context and successfully applied it to cross-classify in another context with a different degree of social distance (e.g., far: *Past Self vs. Queen*). This suggests context-invariant patterns that encode the *essence* of 'self *vs.* other', generalisable across close and distant social relationships. The representational content of these regions is instantiated in the three inset-boxes of Figure 2A. Here we saw a consistent pattern across the three ROIs with highest cross-classification accuracy – when the ground truth was 'self' the classifier was more inclined to predict 'self' than 'other', and *vice versa* when the ground truth was 'other'. A coherent but more striking pattern was found in the four-way cross-classification which generalised across positive *vs.* negative emotive valence.

As illustrated in Figure 2B, based on the regional activity of an ROI, the algorithm was capable of extrapolating personal identities from one context and leveraging the codes to make predictions in another context of different emotive valence. Significantly above-chance cross-classification was achieved in every ROI of the DN and SN. The representational contents of such cross-valance neural coding were exemplified by the pattern of PCC activity: As illustrated in the inset box of Figure 2B, the percentage of the classifier's prediction clearly followed a gradient-like pattern (which was particularly conspicuous in the polar 'extreme' cases of *Present Self* and *Queen*). For instance, while the classifier correctly predicted *Present Self* most frequently when the truth was indeed *Present Self*, its erroneous responses conformed to social distance – *Present Self* was most confused with *Past Self*, followed by *Mother*, and least confused with *Queen*. A similar pattern (with the opposite order) was observed for *Queen*. Together, these indicated that the underlying neural representations that were employed to represent the distinction amongst personal identities were reproducible across contexts, invariant to changes of social distance and emotive valence, and were structured in a gradational manner that followed interpersonal distance. In *SI Appendix*, we report a complementary analysis wherein the algorithm was trained to cross-classify social distance across self *vs.* other (i.e., it learnt to discern Present Self *vs.* Past Self, was tested with Mother *vs.* Queen, and *vice versa*). Results showed that none of the ROIs contained statistically reliable information to sustain this cross-classification, indicating that the overarching distinction of 'self *vs.* other' might be a superordinate dimension that other subordinate attributes (e.g., social distance) hinged upon. When the superordinate information of 'self *vs.* other' was rendered unavailable (i.e., the set-up of this analysis deliberately confined the materials of training to be within one domain but not the other), the algorithm became incapable of cross-classifying labels of social distance between self-ness and otherness. To better navigate the principal and subsidiary dimensions of neural coding, we employed representational similarity analysis (RSA) to reveal the representational structure (see the subsequent RSA section for elaboration).

In both cross-validation and cross-classification analyses, we found a robust disparity that decoding accuracy was consistently better in DN regions than in SN regions. This implies that DN regions might contain more information about the persons than SN regions. To further investigate this gap in predictive power between the two subnetworks, we employed the combinatorial-ROI decoding analysis (for details, see 42, 43, 44). This analysis entailed a stepwise procedure wherein decoding was conducted on a joint ROI, which was crafted by merging an original ROI with another ROI; this was iterated for every pairwise combination of ROIs (see *Methods*). The outcome of joint-ROI decoding was subsequently compared with the baseline performance where the decoding was conducted using the patterns from the original ROI alone. This procedure revealed

whether adding a given ROI improved or impaired the classifier's performance through serially coupling this ROI with other ROIs and assessing how the joint-decoding altered relative to baseline. As has been demonstrated by previous research (42-44), this method bypassed the difficulty of directly comparing between ROIs on the amount of information they carried (i.e., in the serial joint-decoding, with the number of voxels being exactly matched, we were able to quantitatively evaluate if an ROI carried more/less information and was a reliable 'contributor'/'beneficiary'). Using this approach, we found that (*i*) adding any of the five DN regions to the combinatorial decoding robustly boosted accuracy whereas adding any of the three SN regions had little impact on decoding; (*ii*) the three SN regions reliably benefited more from the addition of another ROI whereas the DN regions benefited less. Shown in Figure 3 are the example results of joint-decoding for '*Self vs. Other*' (Figure 3A), '*Present Self vs. Past Self*' (Figure 3B), and '*Mother vs. Queen*' (Figure 3C). A highly reliable pattern was seen in these results (as well as in other joint-decoding). Dovetailing our earlier data, the joint-decoding revealed a robust 'imbalance' between the amount of information carried by the two subnetworks. Whereas DN regions carried more information about personal identities (making them 'givers' that reinforced classification accuracy), SN regions carried less information (making them reliable 'takers' in the combinatorial-ROI decoding).

Representational similarity analysis. Classification analysis gave tantalising signs regarding the configuration of neural coding: Personal identity might be structured in a continuous or gradient fashion along a single-dimension spectrum of social distance (as seen in Figure 2B and the confusion matrix in *SI Appendix*), with potentially an equidistant progression along the gradient. However, it might also be configured with a binary boundary that categorically separates self-ness from otherness, with finer-grained differentiation nested within the 'self' or 'other' domain. Information provided by the confusion matrix was limited, as it did not directly quantify the extent of the similarity between two representations. Unlike classification-style analysis that by nature discretises different categories via imposing a decision boundary, RSA quantifies the extent of similarity using continuous measures (e.g., correlation coefficient). Therefore, we exploited RSA to further investigate the representational 'geometry' that underlies the similarity between different categories and different brain regions.

We first verified whether there was good concordance between the neural representations of positive- and negative-valence conditions, given the fact that the SVM algorithm successfully cross-classified individual identities across valence. As illustrated in Figure 4, in every ROI we found a statistically significant correlation between the representational matrices of positive- and negative-valence contexts, suggesting a coherent motif that the classifier could extrapolate from

one situation and apply to another. Moreover, the outcomes of RSA and pairwise-classification concurred with one another. An example result from one ROI is illustrated in Figure 5: When the representational similarity was high between two conditions (which indicated a higher degree of overlap in neural coding), there was a corresponding decline in the accuracy scores of classification (the classifier was more prone to confusion). This resulted in negative correlations, reliably seen in both positive- and negative-valence contexts. Next, hierarchical clustering was used to visualise the representational distance between conditions, quantified by the branching of a dendrogram tree. As Figure 6 illustrates, in these four 'core' regions of the DN (in which we saw generally higher classification accuracy), there is a clearly dyadic structure that stratified neural representations into two major branches along the border of 'self *vs.* other' distinction. One branch comprised *Present Self* and *Past Self*, while the other branch comprised *Mother* and *Queen*. Interestingly, in the vmPFC, the representations of *Present Self* formed a distinct cluster that was separable from all other conditions. This is consistent with the literature that the vmPFC has a particularly important role in representing the essence of self-concept. It is also clear that emotive valence has minimal impact on the segregation between neural representations, which further validates a reproducible encoding of personal identities unvarying across valence. Whereas a clear bipartite split between self and other was found in the representations of DN regions, in the three SN regions we did not see any systematic clustering that demarcated the difference between conditions. Taken together, these results emphasise the convergence of evidence that we obtained from the two approaches.

Replicated across multiple classification analyses, decoding accuracy was found to be robustly higher in DN regions compared to SN regions. Furthermore, combinatorial-ROI decoding also showed that addition of DN regions reliably boosted accuracy of joint-ROI decoding whereas SN regions caused no improvement, suggesting an asymmetry of information quantity held by the two networks. To investigate this issue farther, we used RSA to examine whether this difference in classification accuracy translates into representational similarity between the two networks. Hierarchical clustering showed that the principal factor that characterised the heterogeneity between neural representations was the separation between DN and SN. As the configuration of the dendrogram illustrates (Figure 7A), the initial bifurcation between regions was driven by the network they belong to. Within the cluster of DN, there were two sub-clusters – the two medial prefrontal regions (the dmPFC and vmPFC) formed a subgroup separable from the posterior regions (medial-parietal: the PCC; lateral-parietal: the left/right IPL). Within the cluster of SN, there was a pattern highly consistent with contemporary theories and findings of semantic cognition (e.g., 45) – the areas representing semantic meaning *per se* (the left/right ATL) formed a sub-cluster separable from the area that controls the retrieval of semantic meaning (the IFG).

This structure is also evident in the similarity matrices whereby we colour-coded the strength of correlation/similarity (Figure 7B): Close inspection of the layout of correlation matrix revealed that the three SN regions congregated to form a cluster that was separable from all of the remaining DN regions; within DN regions, the dmPFC and vmPFC formed a sub-cluster while other posterior regions formed another one. These two visualisation methods provide complementary yet consistent insight into the relationship of neural networks. Finally, we computed the averaged sizes of correlation between brain regions as a function of whether it was correlating two regions within the same network or between the two networks, as well as whether data was based on the positive- or negative-valence context. Results showed a significant effect of network membership – correlation size was significantly greater within the same than between networks (Kendall's τ-a: $F_{(1,23)} = 6.36$, $p = 0.01$, $\eta_p^2 = 0.22$), indicating greater representational similarity between regions of the same clan. By contrast, valence had no effect (again indicating a coherent representational structure across valence), nor did the valence × network interaction (both *ps* > 0.55). To ascertain robustness, we repeated all these analyses using a different type of rank-correlation measure (Spearman's ρ) and obtained entirely consistent results (see Figure 7). Together, these results complement our observations of the classification-based analysis and further reveal the subtlety of subdivision within a network.

Finally, we examined three theoretical models, testing their explanatory power to account for the neural data (see Figure 8). The first model was derived from each participant's subjective rating of the pairwise similarity between the personalities of two individuals using a continuous scale (most dissimilar: 0 – most similar: 100). The second model was a hypothetical model based on a binary distinction between self and other – *Present Self* and *Past Self* were collapsed under the 'self' umbrella while *Mother* and *Queen* were under the 'other' umbrella. The third model was a hypothetical model based on gradational changes, with each of the four individuals assumed to be equidistant from one another on the spectrum. We correlated the neural similarity matrix of each brain region with these theoretical models, and statistically examined its reliability (FDR-adjusted for multiple comparisons). We found that the neural pattern of every brain region significantly correlated with the binary model and the graded model (all *p*s < 0.005). As illustrated in Figure 8, for both of the binary and graded model, the neural pattern of the PCC was most correlated with the theoretical models and approached the lower-bound of noise ceiling (indicating the models' near-optimal capability to explain neural data). The correlation sizes between brain regions did not reliably differed, nor did the comparison between the binary model and graded model (all *p*s > 0.05 with FDR-correction). By contrast, the behavioural rating matrix did not reliably correlate with the neural pattern of any brain region. Critically, both the binary and graded models

outstripped the model of behavioural rating in terms of their explanatory power on the neural data (all *p*s < 0.05). While subjective rating best characterised how a participant perceived and compared the extent of resemblance of two persons' character traits, such high-dimensional, multi-faceted understanding (quantified as their rating) was not fully reflected in the local neural patterns of DN and SN regions. Instead, while the binary model was less elaborated (or more impoverished in the dimensions to characterise differences of personality) than behavioural rating, it was better able to capture the 'representational landscape' of neural data. This also corroborates the results of classification decoding and hierarchical clustering that the dimension of 'self *vs.* other' accounted for the lion share of variation, while social distance (imposed onto the superordinate axis of 'self *vs.* other' in the gradational model) explained the residual variance.

Discussion

In the present study, we used a series of multivoxel decoding to unravel the neural representations of self- *vs.* other-referential concept, a pivotal psychological construct that permeates every aspects of social life and has been argued as a core function of the specialised system for social cognition. Across multiple analyses, we found a robust dyadic fractionation between the two subnetworks within the social system – multivoxel patterns reliably dissociated between regions that are canonically affiliated with the DN and those with the SN, evident both in the outcomes of supervised classification and representational similarity. Moreover, representational configuration of the neural response to self- and other-referential thoughts mirrored the psychological continuum of interpersonal distance, from an integral sense of *Present Self* to a distant other (*Queen*), with the broad-stroke distinction of 'self-ness *vs.* otherness' being the principal axis in the representational space, and social distance being the auxiliary axis. Below we discuss the implications of our discoveries for human connectome research and for social neuroscience.

The bipartite split in multivoxel patterns between networks. Decades of research have accumulated a wealth of data about how inter-connected brain regions form large-scale networks, as well as how network architecture is mapped onto cognitive functions (for review, 46). Of particular interest is the functionality of the broad default-mode network. This widely dispersed group of brain regions were originally identified based on their 'deactivation' (relative to task-state) during active engagement in a task and heightened activation-level during wakeful resting periods (26). Owing to its initial 'task-negative' definition, for a long time this network was assumed to play little role in goal-oriented behaviour. Later research has identified its contribution in a panoply of goal-directed cognitive tasks that depend on internally-constructed representations, such as social cognition (e.g., 47), retrieval of episodic/autobiographical memory (e.g., 12), and even short-term

visual memory (e.g., 48). Despite such evidence, the 'misnomer' of default-mode network (which is rooted in its original task-negative definition as the brain's metabolic default) continues to be used extensively. Recent studies have further uncovered a bipartite structure – two functionally separable subnetworks – within the broad default system (e.g., 4-6, 49). The bipartite structure has been discussed under the framework of a macro-scale 'gradient' spanning the entirety of cerebrum (50, 51). This gradient reflects a recursive process of information convergence, occurring in multiple places of the brain, from sensory-motoric representations that encode the 'here-and-now' of external/perceptible entities to multimodal representations that are stored in the DN and encode memories and concepts. Core regions of the DN (e.g., the PCC and vmPFC) sit atop this gradient, while core regions of the SN (e.g., the ATL and IFG) are situated at a tier below the apex position (i.e., the DN). The dissociation between DN and SN might reflect their differential position on this gradient of information convergence. The multivoxel decoding of the present study lends support to this framework – as revealed by both classification- and similarity-based analysis, we found that DN regions contained more information about individual identity, which is an abstract concept distilled from multiple sources of perceptual modalities and accumulated over a long time-window. Thus, compared to SN areas, the more informative DN areas enabled better classification accuracy (as evident in the outcomes of iterative joint-decoding analysis) and exhibited more clear-cut representational segregation between self and other (as evident in the branch of dendrogram trees). These findings are compatible with the interpretation that the DN contains more abstract information for social cognition, despite the fact that the DN and SN both enhanced their activation in response to the demand of social processing (in a univariate 'BOLD amplitude' sense).

Significantly above-chance decoding was found in every ROIs for coarser- and finer-grained classifications of self *vs.* other, indicating that all of them carried sufficient information that allowed the algorithm to delimit a margin that separates two (or four) classes of data-points. Nonetheless, it is noteworthy that the PCC contained most information about personal identities. This is observed across multiple analyses – amongst the ROIs included in the present investigation, decoding accuracy was highest based on the patterns of PCC; when the PCC was added to the combinatorial decoding, it led to greatest boost to the classifier's performance; the neural pattern of the PCC was most correlated with the psychological continuity of self *vs.* other, leading to correlations that approached the noise ceiling (indicating near optimum). These data are consistent with the view that the PCC serves as one of the cores of the default system (32) such that its representational pattern 'echoes' the activities elsewhere in the brain through long-range connections (52) and that its dysfunction is linked with various memory-related ailments (53).

It is important to note that any individual brain region may participate in multiple brain networks and subserve multiple cognitive functions even though a region has a primary network affiliation or certain types of functions that it is frequently associated with (54). The 'network membership' of a brain region can be affected by various factors, such as its position in the network (i.e., 'connector' regions between two networks tend to have more fluid network affiliation, relative to regions within the 'heartland' of a module; see 55, 56), the cognitive state one is under (i.e., contexts can drive a region to fluidly couple with different networks), as well as the methods used to probe the relationship between regions (for discussion, see 57). These factors may particularly affect several 'connector' regions that bridges between two modules, making them exhibit functional features of both modules (such as the dmPFC that bridges core DN nodes with SN nodes; see 56). We speculate that this may offer explanations regarding the functional profiles of dmPFC – previously, using *univariate* analysis, we found that the dmPFC responded preferentially to semantic tasks over episodic tasks, which makes it more akin to the behaviour of SN regions (e.g., the IFG and ATL) than DN regions (e.g., vmPFC; Chiou *et al.*, 2020). However, using *multivariate* analysis, in the present study we found that representational contents of the dmPFC was more similar to those of DN regions relative to SN regions, making it more affiliated to the DN. Together, our data are consistent with the previous literature concerning this region's somewhat inconclusive affiliation with different networks; it also highlights the flexibility of a 'connector' region and the peril of imposing a rigid demarcation on the perimeters between networks.

The representational structure of self- vs. other-concept. Our findings add to the literature that deciphering self- and other-referential thoughts is possible based on neural patterns of DN and SN regions (e.g., 19, 20). Furthermore, the results of our cross-classification analyses have important implications for the psychological theories of self-processing. The ability to successfully cross-classify has been considered as a benchmark of testing whether there is genuine abstraction of neural coding across domains (58). We found that the brain uses robustly generalisable and differentiable neural patterns to encode self-ness and otherness such that this cardinal representational code applies across near and far interpersonal distances and across positive- and negative-valence depictions. On top of the separation between self and other, social distance was another factor that sculpted the representational landscape of the neural codes. As discussed earlier, 'self vs. other' serves as the primary dimension that delineates the most salient difference between classes, while secondary dimensions (e.g., social distance) are couched within the primary segregation. This is reminiscent of representational dimensions of visual perception (59): Animacy (living *vs.* non-living) is the primary dimension that separates animate entities from inanimate items, while object size serves as the secondary dimension that assorts subgroup within the

inanimate class (60). Interestingly, participants' individual rating on pair-wise similarity between individuals did not correlated with the neural pattern of any region. A possible reason behind this result might be that while the local pattern of an ROI is sufficient to represent the binary/broad-stroke distinction of self *vs.* other, as well as the gradational differences along social distance (i.e., the cardinal *plus* the secondary axis), it was insufficient to capture the more sophisticated pattern inherent in the thoughts behind the behavioural rating. As the matrix of behavioural rating illustrates (Figure 8), on average, our participants rated that their own mother was more similar to themselves and drastically different from the *Queen*; this led to a clear cluster that lumped the three personally familiar targets together (*Present Self*, *Past Self*, and *Mother*), separate from the *Queen*. The majority of participants rated that the similarity of *Present Self* and *Mother* was even higher than that of *Present Self* and *Past Self*. This implies elaborated and multifaceted considerations behind the ratings that might require additional neurocomputation beyond the information within a single DN/SN region, whose local pattern cares primarily about self *vs.* other.

Conclusion. In the present study, we reported results of multivoxel decoding that echo with recent proposals of a bipartite fractionation within the macro-scale network of socio-cognitive processing. The multivariate findings extend beyond previous evidence that was based on connectivity by demonstrating that different subnetworks within the social-cognition system dissociated on the amount of information that was loaded on each subnetwork, as well as on the representational geometry that encoded different person-related concepts. These findings inform the burgeoning field of network science and its application in cognitive neuroscience; our results also shed light on the decade-long endeavour of investigating the neural implementation that underlies the thoughts about the distinction of self *vs.* other.

Methods

All of the methodological details regarding experimental paradigm, MRI acquisition parameters, pre-processing of MRI data, definition of ROIs, different approaches of decoding analysis, the stimuli presented in the fMRI experiment, as well as the illustrations of task procedures are reported in *SI Appendix*. Two supplemental analyses are also reported in *SI Appendix* – (*i*) the confusion matrices that accompanied the four-way classification analysis of four individuals and (*ii*) the cross-classification of near vs. far social distance between the 'self' and 'other' domain.

Twenty-four volunteers gave informed consent before the fMRI experiment. The sample consisted of a 10/14 female-to-male ratio, with average age = 33 years-old and SD = 11. All volunteers are right-handed and speak English as their mother tongue. All of them completed the magnetic resonance imaging safety screening questionnaire before the experiment and reported not having any neurological or psychiatric condition. This study was reviewed and approved by the local research ethics committee.

Anonymised analytical codes and fMRI data are publicly available on the MRC Cognition and Brain Sciences Unit Data Repository (https://www.mrc-cbu.cam.ac.uk/bibliography/opendata/).

References

1. N. Steinbeis, The role of self–other distinction in understanding others' mental and emotional states: neurocognitive mechanisms in children and adults. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**, 20150074 (2016).

2. M. D. Lieberman, M. A. Straccia, M. L. Meyer, M. Du, K. M. Tan, Social, self,(situational), and affective processes in medial prefrontal cortex (MPFC): Causal, multivariate, and reverse inference evidence. *Neuroscience & Biobehavioral Reviews* (2019).

3. R. L. Carhart-Harris, K. J. Friston, The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. *Brain* **133**, 1265-1283 (2010).

4. R. M. Braga, K. R. Van Dijk, J. R. Polimeni, M. C. Eldaief, R. L. Buckner, Parallel distributed networks resolved at high resolution reveal close juxtaposition of distinct regions. *Journal of neurophysiology* **121**, 1513-1534 (2019).

5. R. M. Braga, R. L. Buckner, Parallel interdigitated distributed networks within the individual estimated by intrinsic functional connectivity. *Neuron* **95**, 457-471. e455 (2017).

6. R. Chiou, G. F. Humphreys, M. A. Lambon Ralph, Bipartite functional fractionation within the default network supports disparate forms of internally oriented cognition. *Cerebral Cortex* **30**, 5484-5501 (2020).

7. D. Oyserman, K. Elmore, G. Smith, "Self, self-concept, and identity" in Handbook of self and identity (The Guilford Press, New York, NY, US, 2012), pp. 69-104.

8. S. B. Klein, J. Loftus, J. F. Kihlstrom, Memory and temporal experience: The effects of episodic memory loss on an amnesic patient's ability to remember the past and imagine the future. *Social Cognition* **20**, 353-379 (2002).

9. K. Verschueren, S. Doumen, E. Buyse, Relationships with mother, teacher, and peers: Unique and joint effects on young children's self-concept. *Attachment & Human Development* **14**, 233-248 (2012).

10. M. Deković, W. Meeus, Peer relations in adolescence: Effects of parenting and adolescents' self-concept. *Journal of adolescence* **20**, 163-176 (1997).

11. W. M. Kelley *et al.*, Finding the self? An event-related fMRI study. *Journal of cognitive neuroscience* **14**, 785-794 (2002).

12. R. N. Spreng, C. L. Grady, Patterns of brain activity supporting autobiographical memory, prospection, and theory of mind, and their relationship to the default mode network. *Journal of cognitive neuroscience* **22**, 1112-1123 (2010).

13. J. M. Moran, C. N. Macrae, T. F. Heatherton, C. L. Wyland, W. M. Kelley, Neuroanatomical evidence for distinct cognitive and affective components of self. *Journal of cognitive neuroscience* **18**, 1586-1594 (2006).

14. C. L. Philippi, M. C. Duff, N. L. Denburg, D. Tranel, D. Rudrauf, Medial PFC damage abolishes the self-reference effect. *Journal of Cognitive Neuroscience* **24**, 475-481 (2012).

15. F. M. Krienen, P.-C. Tu, R. L. Buckner, Clan mentality: evidence that the medial prefrontal cortex responds to close others. *Journal of Neuroscience* **30**, 13906-13915 (2010).

16. J. P. Mitchell, C. N. Macrae, M. R. Banaji, Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron* **50**, 655-663 (2006).

17. R. S. Chavez, T. F. Heatherton, D. D. Wagner, Neural population decoding reveals the intrinsic positivity of the self. *Cerebral Cortex* **27**, 5222-5229 (2016).

18. A. Yankouskaya *et al.*, An anterior–posterior axis within the ventromedial prefrontal cortex separates self and reward. *Social cognitive and affective neuroscience* **12**, 1859-1868 (2017).

19. D. Hassabis *et al.*, Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex* **24**, 1979-1987 (2013).

20. M. A. Thornton, J. P. Mitchell, Consistent neural activity patterns represent personally familiar people. *Journal of cognitive neuroscience* **29**, 1583-1594 (2017).

21. J. Koster-Hale, R. Saxe, J. Dungan, L. L. Young, Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences* **110**, 5648-5653 (2013).

22. D. D. Wagner, R. S. Chavez, T. W. Broom, Decoding the neural representation of self and person knowledge with multivariate pattern analysis and data-driven approaches. *Wiley Interdisciplinary Reviews: Cognitive Science* **10**, e1482 (2019).

23. B. P. Doré, N. Zerubavel, K. N. Ochsner, Social cognitive neuroscience: A review of core systems. *APA handbook of personality and social psychology* **1**, 693-720 (2014).

24. I. R. Olson, D. McCoy, E. Klobusicky, L. A. Ross, Social cognition and the anterior temporal lobes: a review and theoretical framework. *Social cognitive and affective neuroscience* **8**, 123-133 (2013).

25. R. J. Binney, P. Hoffman, M. A. Lambon Ralph, Mapping the multiple graded contributions of the anterior temporal lobe representational hub to abstract and social concepts: evidence from distortion-corrected fMRI. *Cerebral cortex* **26**, 4227-4241 (2016).

26. M. E. Raichle *et al.*, A default mode of brain function. *Proceedings of the National Academy of Sciences* **98**, 676-682 (2001).

27. Y. Yeshurun, M. Nguyen, U. Hasson, The default mode network: where the idiosyncratic self meets the shared social world. *Nature Reviews Neuroscience*, 1-12 (2021).

28. C. Murphy *et al.*, Hello, is that me you are looking for? A re-examination of the role of the DMN in social and self relevant aspects of off-task thought. *PloS one* **14** (2019).

29. R. N. Spreng, J. R. Andrews-Hanna, The default network and social cognition. *Brain mapping: An encyclopedic reference* **1316**, 165-169 (2015).

30. R. L. Jackson, L. L. Cloutman, M. A. L. Ralph, Exploring distinct default mode and semantic networks using a systematic ICA approach. *Cortex* **113**, 279-297 (2019).

31. G. F. Humphreys, P. Hoffman, M. Visser, R. J. Binney, M. A. Lambon Ralph, Establishing task-and modality-dependent dissociations between the semantic and default mode networks. *Proceedings of the National Academy of Sciences* **112**, 7857-7862 (2015).

32. J. R. Andrews-Hanna, J. S. Reidler, J. Sepulcre, R. Poulin, R. L. Buckner, Functional-anatomic fractionation of the brain's default network. *Neuron* **65**, 550-562 (2010).

33. R. Chiou, G. F. Humphreys, J. Jung, M. A. Lambon Ralph, Controlled semantic cognition relies upon dynamic and flexible interactions between the executive 'semantic control'and hub-and-spoke 'semantic representation'systems. *Cortex* **103**, 100-116 (2018).

34. J. Hiser, M. Koenigs, The multifaceted role of the ventromedial prefrontal cortex in emotion, decision making, social cognition, and psychopathology. *Biological Psychiatry* **83**, 638-647 (2018).

35. R. L. Jackson, C. J. Bajada, M. A. Lambon Ralph, L. L. Cloutman, The graded change in connectivity across the ventromedial prefrontal cortex reveals distinct subregions. *Cerebral Cortex* **30**, 165-180 (2020).

36. S. B. Eickhoff, A. R. Laird, P. T. Fox, D. Bzdok, L. Hensel, Functional segregation of the human dorsomedial prefrontal cortex. *Cerebral cortex* **26**, 304-321 (2016).

37. D. Bzdok *et al.*, Subspecialization in the human posterior medial cortex. *Neuroimage* **106**, 55-71 (2015).

38. D. Bzdok *et al.*, Segregation of the human medial prefrontal cortex in social cognition. *Frontiers in human neuroscience* **7**, 232 (2013).

39. G. Northoff, Personal Identity and Cortical Midline Structure (CMS): Do Temporal Features of CMS Neural Activity Transform Into "Self-Continuity"? *Psychological Inquiry* **28**, 122-131 (2017).

40. F. Pereira, T. Mitchell, M. Botvinick, Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage* **45**, S199-S209 (2009).

41. H. Nili *et al.*, A toolbox for representational similarity analysis. *PLoS computational biology* **10**, e1003553 (2014).

42. Y. Wang *et al.*, Dynamic neural architecture for social knowledge retrieval. *Proceedings of the National Academy of Sciences* **114**, E3305-E3314 (2017).

43. J. A. Clithero, R. M. Carter, S. A. Huettel, Local pattern classification differentiates processes of economic valuation. *Neuroimage* **45**, 1329-1338 (2009).

44. D. V. Smith, J. A. Clithero, C. Rorden, H.-O. Karnath, Decoding the anatomical network of spatial attention. *Proceedings of the National Academy of Sciences* **110**, 1518-1523 (2013).

45. M. A. Lambon Ralph, E. Jefferies, K. Patterson, T. T. Rogers, The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience* **18**, 42-55 (2017).

46. L. Q. Uddin, B. T. Yeo, R. N. Spreng, Towards a universal taxonomy of macro-scale functional human brain networks. *Brain topography*, 1-17 (2019).

47. A. E. Skerry, R. Saxe, Neural representations of emotion are organized around abstract event features. *Current biology* **25**, 1945-1954 (2015).

48. C. Murphy *et al.*, Modes of operation: A topographic neural gradient supporting stimulus dependent and independent cognition. *NeuroImage* **186**, 487-496 (2019).

49. L. M. DiNicola, R. M. Braga, R. L. Buckner, Parallel Distributed Networks Dissociate Episodic and Social Functions Within the Individual. *Journal of Neurophysiology* 10.1152/jn.00529.2019 (2020).

50. D. S. Margulies *et al.*, Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences* **113**, 12574-12579 (2016).

51. J. M. Huntenburg *et al.*, A systematic relationship between functional connectivity and intracortical myelin in the human cerebral cortex. *Cerebral Cortex* **27**, 981-997 (2017).

52. R. Leech, R. Braga, D. J. Sharp, Echoes of the brain within the posterior cingulate cortex. *Journal of Neuroscience* **32**, 215-222 (2012).

53. R. Leech, D. J. Sharp, The role of the posterior cingulate cortex in cognition and disease. *Brain* **137**, 12-32 (2014).

54. L. Pessoa, Understanding brain networks and brain organization. *Physics of life reviews* **11**, 400-435 (2014).

55. R. N. Spreng, W. D. Stevens, J. P. Chamberlain, A. W. Gilmore, D. L. Schacter, Default network activity, coupled with the frontoparietal control network, supports goal-directed cognition. *Neuroimage* **53**, 303-317 (2010).

56. R. N. Spreng, J. Sepulcre, G. R. Turner, W. D. Stevens, D. L. Schacter, Intrinsic architecture underlying the relations among the default, dorsal attention, and frontoparietal control networks of the human brain. *Journal of cognitive neuroscience* **25**, 74-86 (2013).

57. S. E. Petersen, O. Sporns, Brain networks and cognitive architectures. *Neuron* **88**, 207-219 (2015).

58. J. T. Kaplan, K. Man, S. G. Greening, Multivariate cross-classification: applying machine learning techniques to characterize abstraction in neural representations. *Frontiers in human neuroscience* **9**, 151 (2015).

59. T. Konkle, A. Caramazza, Tripartite organization of the ventral stream by animacy and object size. *The Journal of Neuroscience* **33**, 10235-10242 (2013).

60. B. Long, C.-P. Yu, T. Konkle, Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences* **115**, E9015-E9024 (2018).

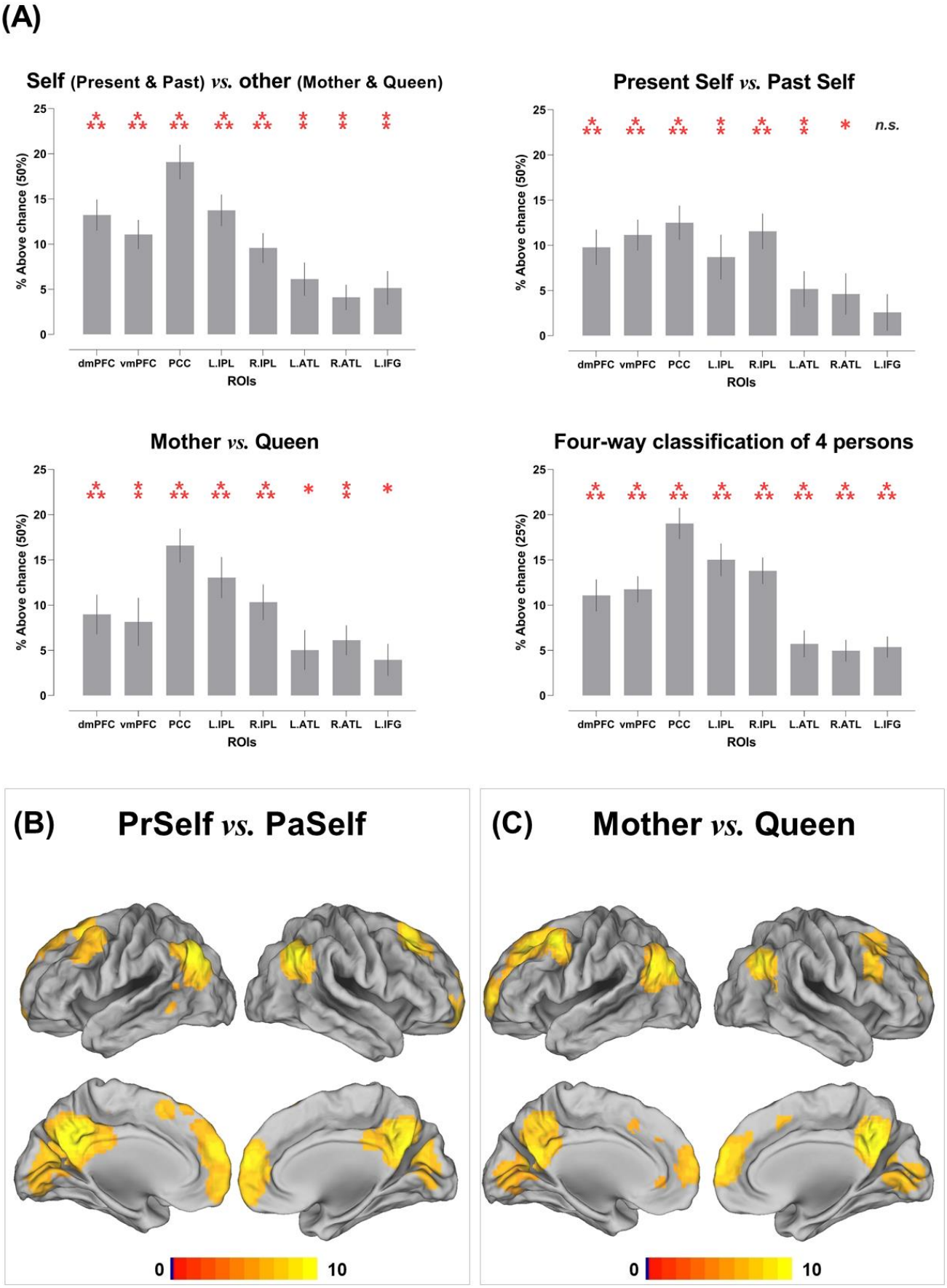
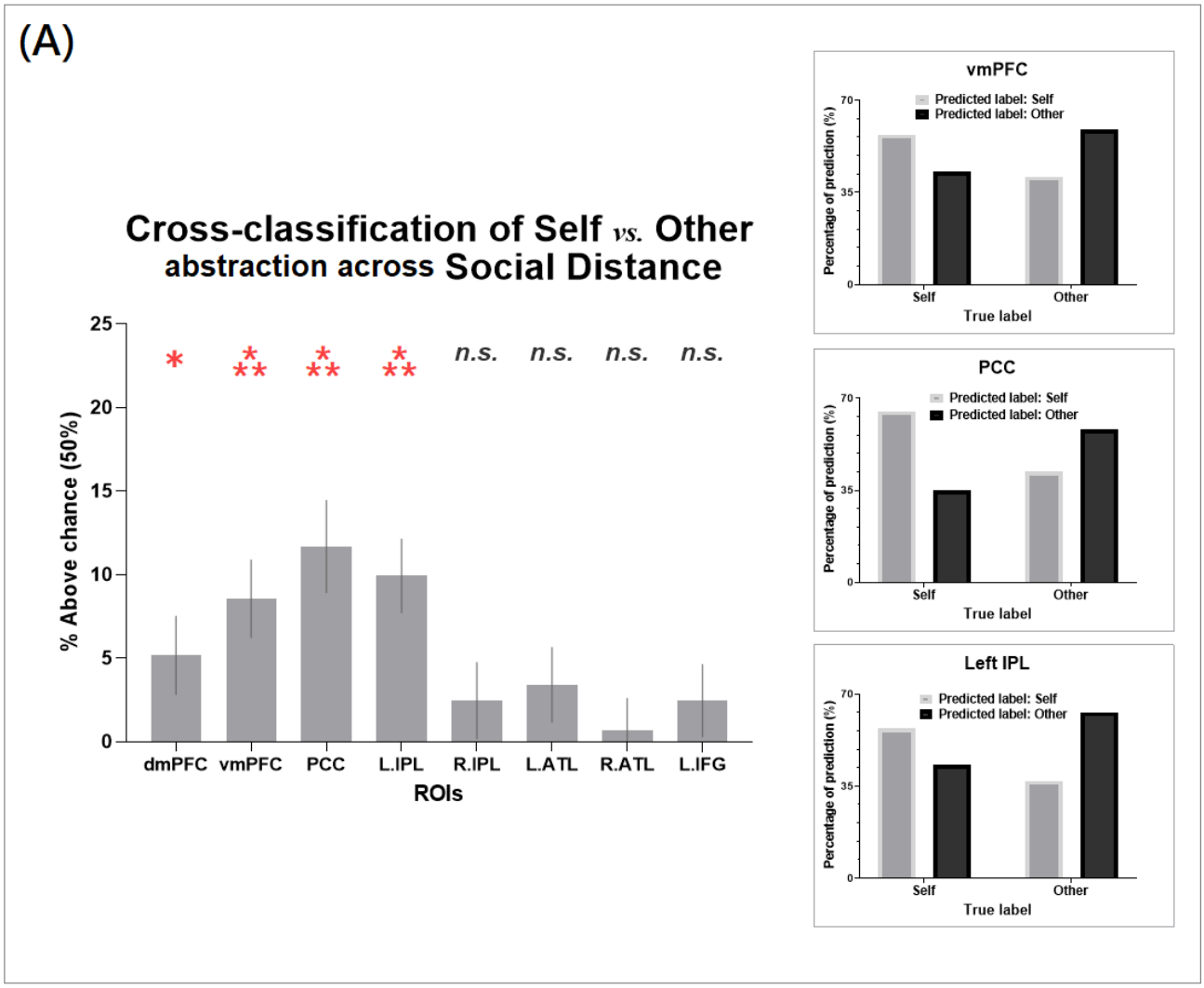

(A)
Self (Present & Past) vs. other (Mother & Queen)
Present Self vs. Past Self
Mother vs. Queen
Four-way classification of 4 persons
% Above chance (50%)
% Above chance (25%)
n.s.
dmPFC vmPFC PCC L.IPL R.IPL L.ATL R.ATL L.IFG
ROIs
(B) PrSelf vs. PaSelf
(C) Mother vs. Queen
0
10

Fig. 1 (A) Classification accuracy of four classification analyses; asterisk indicate significantly above-chance decoding (FDR-corrected for multiple comparisons) ***** $p < 0.01$; ****** $p < 0.005$; ******* $p < 0.001$. Whole-brain searchlight decoding for **(B)** Present Self *vs.* Past Self, as well as **(C)** Mother *vs.* Queen. Correction for multiple comparisons was controlled by constraining voxel-wise FWE under $p < 0.05$


(A)
Cross-classification of Self vs. Other
abstraction across Social Distance
% Above chance (50%)
25
20
15
10
5
0
*
* **
* **
* **
n.s.
n.s.
n.s.
n.s.
dmPFC
vmPFC
PCC
L.IPL
R.IPL
L.ATL
R.ATL
L.IFG
ROIs
vmPFC
Predicted label: Self
Predicted label: Other
Percentage of prediction (%)
70
35
0
Self
Other
True label
PCC
Predicted label: Self
Predicted label: Other
Percentage of prediction (%)
70
35
0
Self
Other
True label
Left IPL
Predicted label: Self
Predicted label: Other
Percentage of prediction (%)
70
35
0
Self
Other
True label

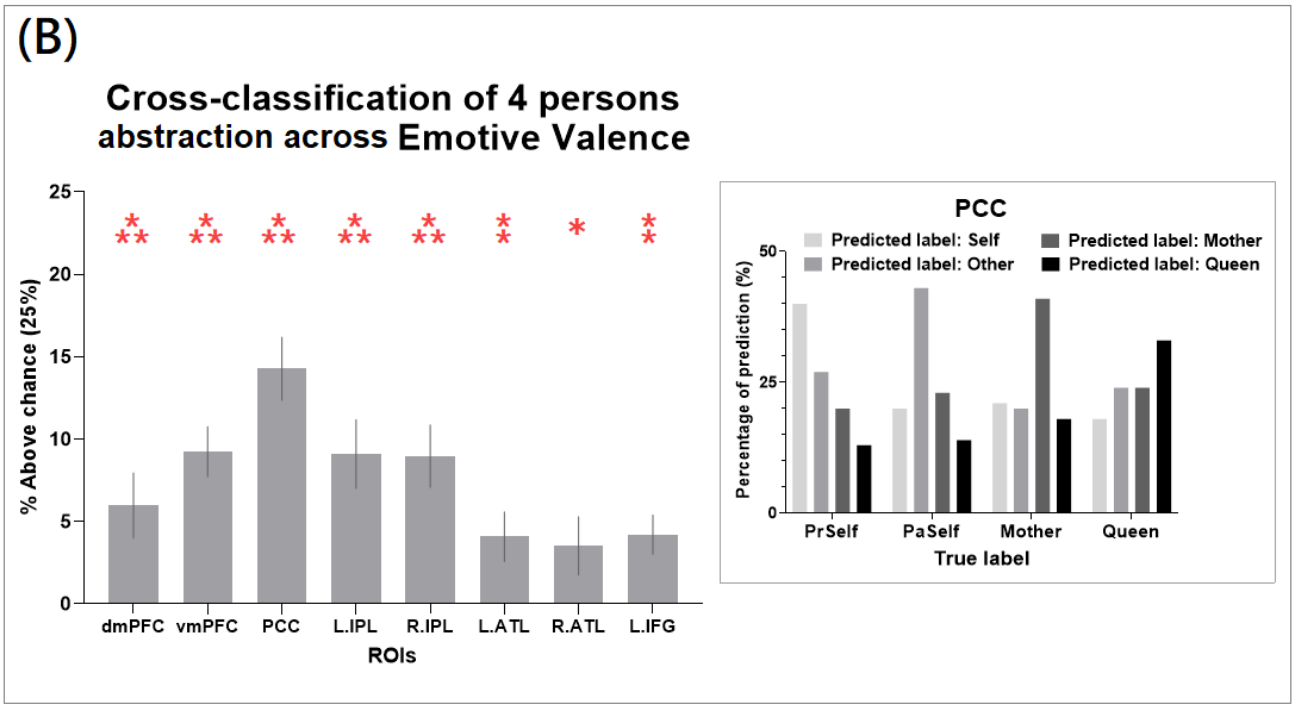
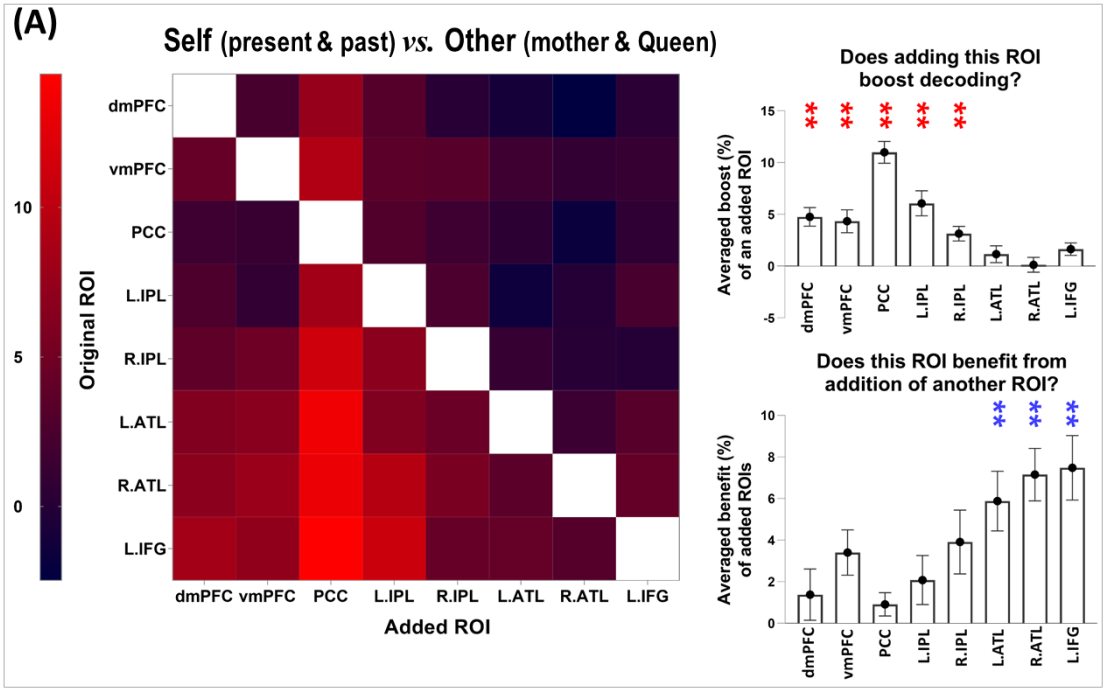
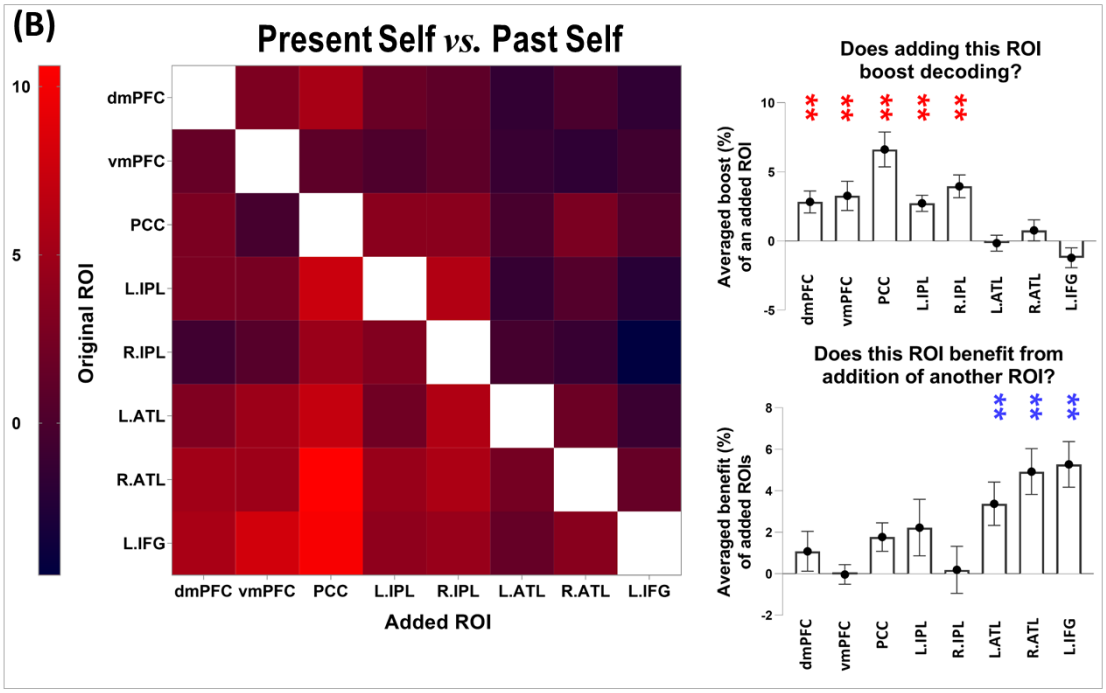

(B)
Cross-classification of 4 persons
abstraction across Emotive Valence
% Above chance (25%)
25
20
15
10
5
0
* **
* **
* **
* **
* **
* *
*
* *
dmPFC
vmPFC
PCC
L.IPL
R.IPL
L.ATL
R.ATL
L.IFG
ROIs
PCC
Predicted label: Self
Predicted label: Other
Predicted label: Mother
Predicted label: Queen
Percentage of prediction (%)
50
25
0
PrSelf
PaSelf
Mother
Queen
True label

Fig. 2 (A) Cross-classification of self *vs.* other between near and far interpersonal distances. **(B)** Cross-classification of four individual identities between positive- and negative-valence stimuli. Inset boxes show the confusion pattern of ROIs with highest cross-classification accuracy. FDR-correction was applied for multiple comparisons. ***** $p < 0.01$; ****** $p < 0.005$; ******* $p < 0.001$


(A)
Self (present & past) vs. Other (mother & Queen)
Original ROI
Added ROI
dmPFC vmPFC PCC L.IPL R.IPL L.ATL R.ATL L.IFG
Does adding this ROI boost decoding?
Averaged boost (%) of an added ROI
Does this ROI benefit from addition of another ROI?
Averaged benefit (%) of added ROIs


(B)
Present Self vs. Past Self
Original ROI
Added ROI
dmPFC vmPFC PCC L.IPL R.IPL L.ATL R.ATL L.IFG
Does adding this ROI boost decoding?
Averaged boost (%) of an added ROI
Does this ROI benefit from addition of another ROI?
Averaged benefit (%) of added ROIs

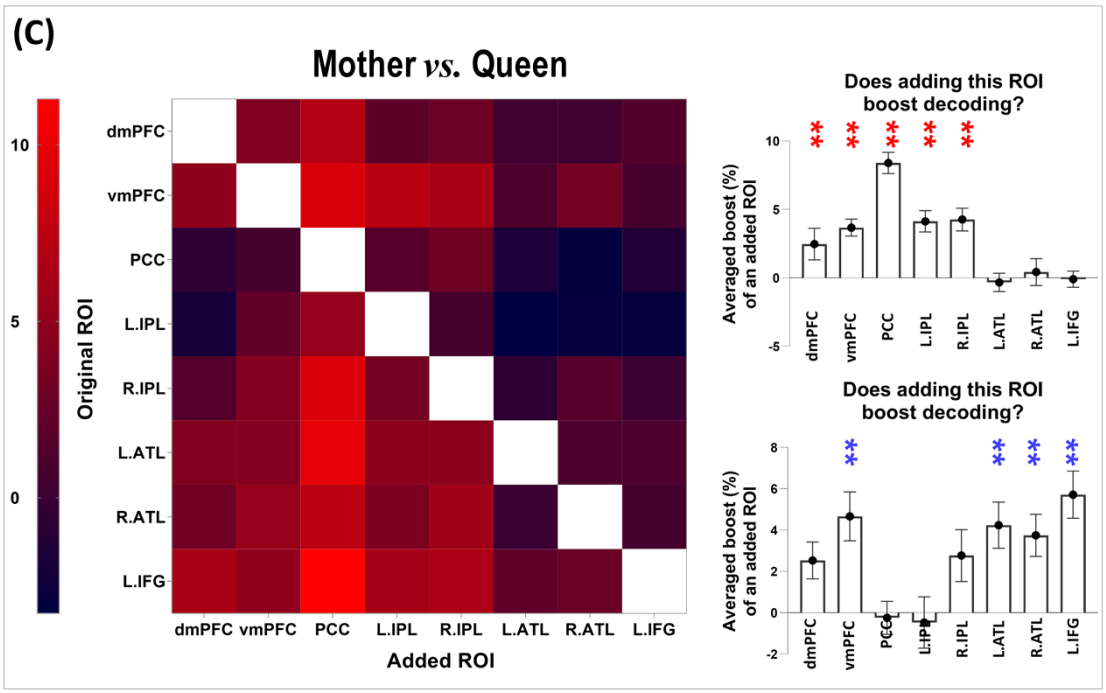

(C)
Mother vs. Queen
Original ROI
Added ROI
dmPFC vmPFC PCC L.IPL R.IPL L.ATL R.ATL L.IFG
Does adding this ROI boost decoding?
Averaged boost (%) of an added ROI
Does adding this ROI boost decoding?
Averaged boost (%) of an added ROI

Figure 3. Combinatorial decoding that was based on an iteratively procedure that examined how decoding performances varied when the neural pattern of an original ROI was joined by the pattern of another ROI. The values on the colour scale indicates changes in decoding accuracy when an ROI was included, compared to the decoding result based on the original alone, from dark purple (below zero, indicating decrement) to bright red (indicating most improvement). Reliable patterns were found across all analyses – three examples here: **(A)** Present Self + Past Self *vs.* Mother + Queen; **(B)** Present Self *vs.* Past Self; **(C)** Mother *vs.* Queen. The bar-graphs indicate that, across analyses and across original ROIs, the addition of an ROI that belongs to the default network led to reliable improvement of decoding results, whereas ROIs of the semantic network reliably benefited most from the addition of another ROI.

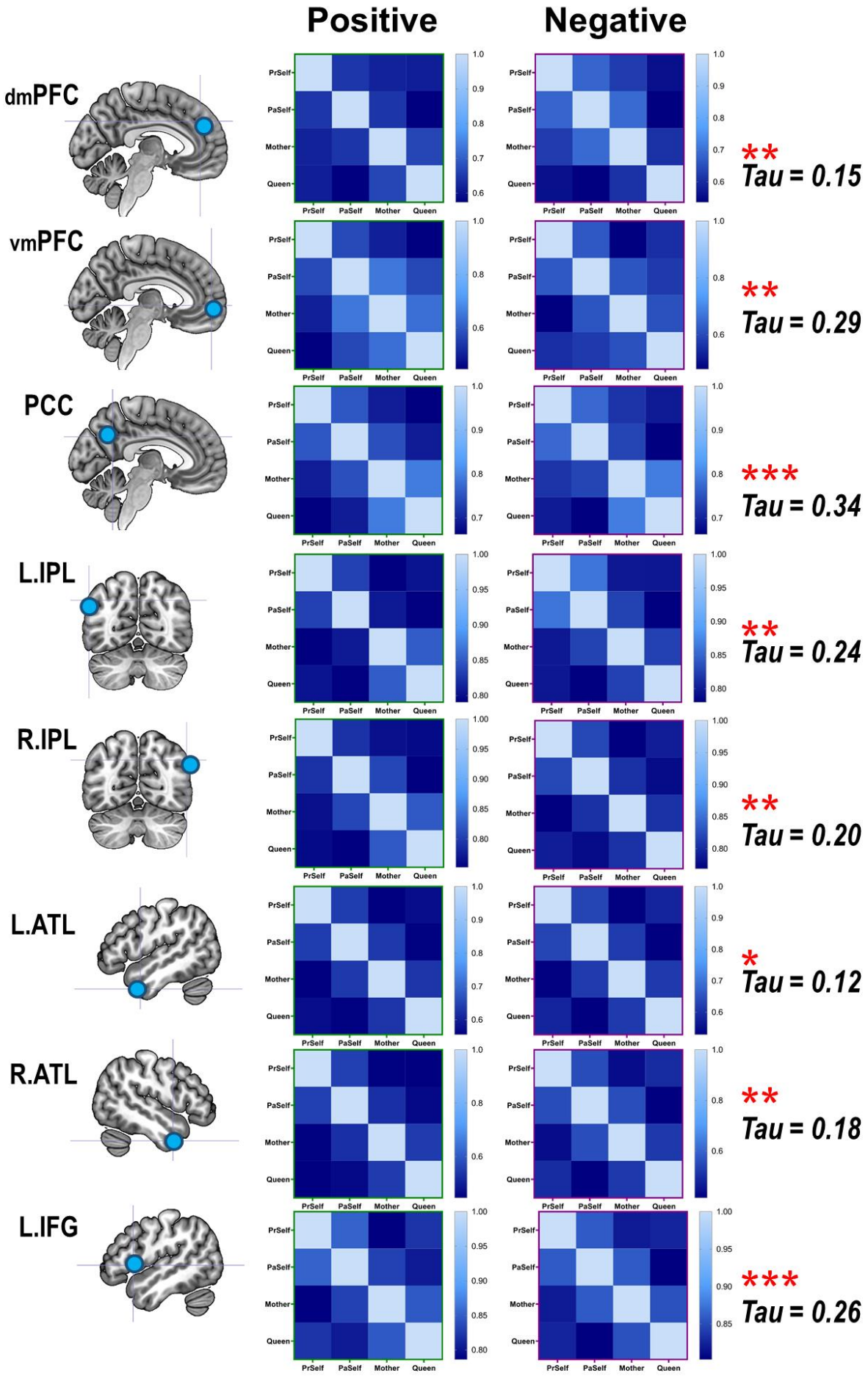

Positive
Negative
dmPFC
vmPFC
PCC
L.IPL
R.IPL
L.ATL
R.ATL
L.IFG
PrSelf
PaSelf
Mother
Queen
**
Tau = 0.15
**
Tau = 0.29
***
Tau = 0.34
**
Tau = 0.24
**
Tau = 0.20
*
Tau = 0.12
**
Tau = 0.18
***
Tau = 0.26

Figure 4. Corroborating the results of cross-classification between positive- and negative-valence stimuli, the results of representational similarity analysis showed that the neural patterns elicited by person-related thoughts are significantly correlated between the contexts of positive- and negative-valence. The correlation was found in every regions of interest.

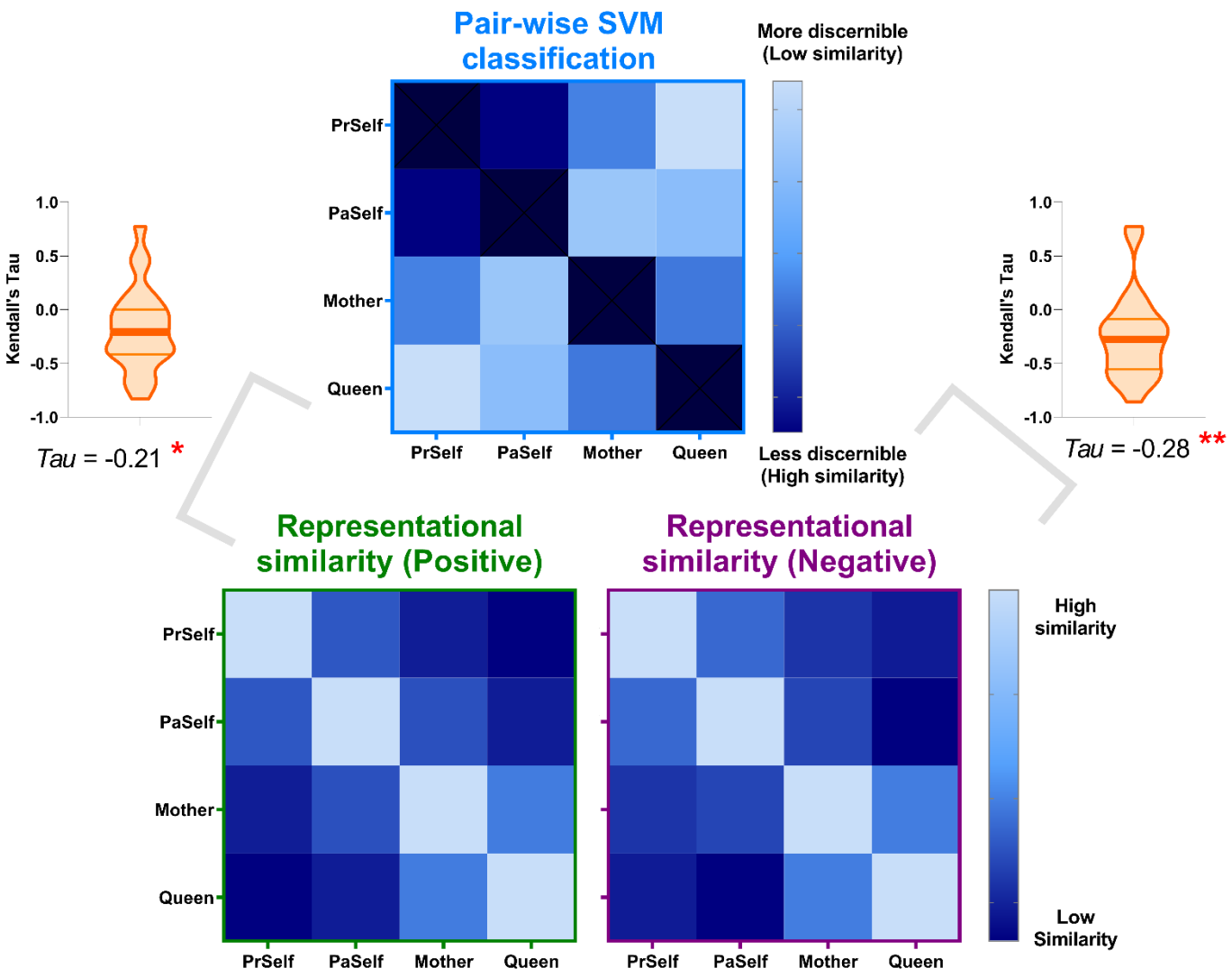


Figure 5. The results of representational similarity and algorithmic classification analyses dovetailed with each other. Illustrated here are example results based on the neural patterns of the posterior cingulate cortex (coherent results were observed for every other region) – when two experimental conditions were representationally *more* similar to each other (hence more confusable), the classifier was less able to predict the correct category label, resulting in significant negative correlations found in both valence conditions.

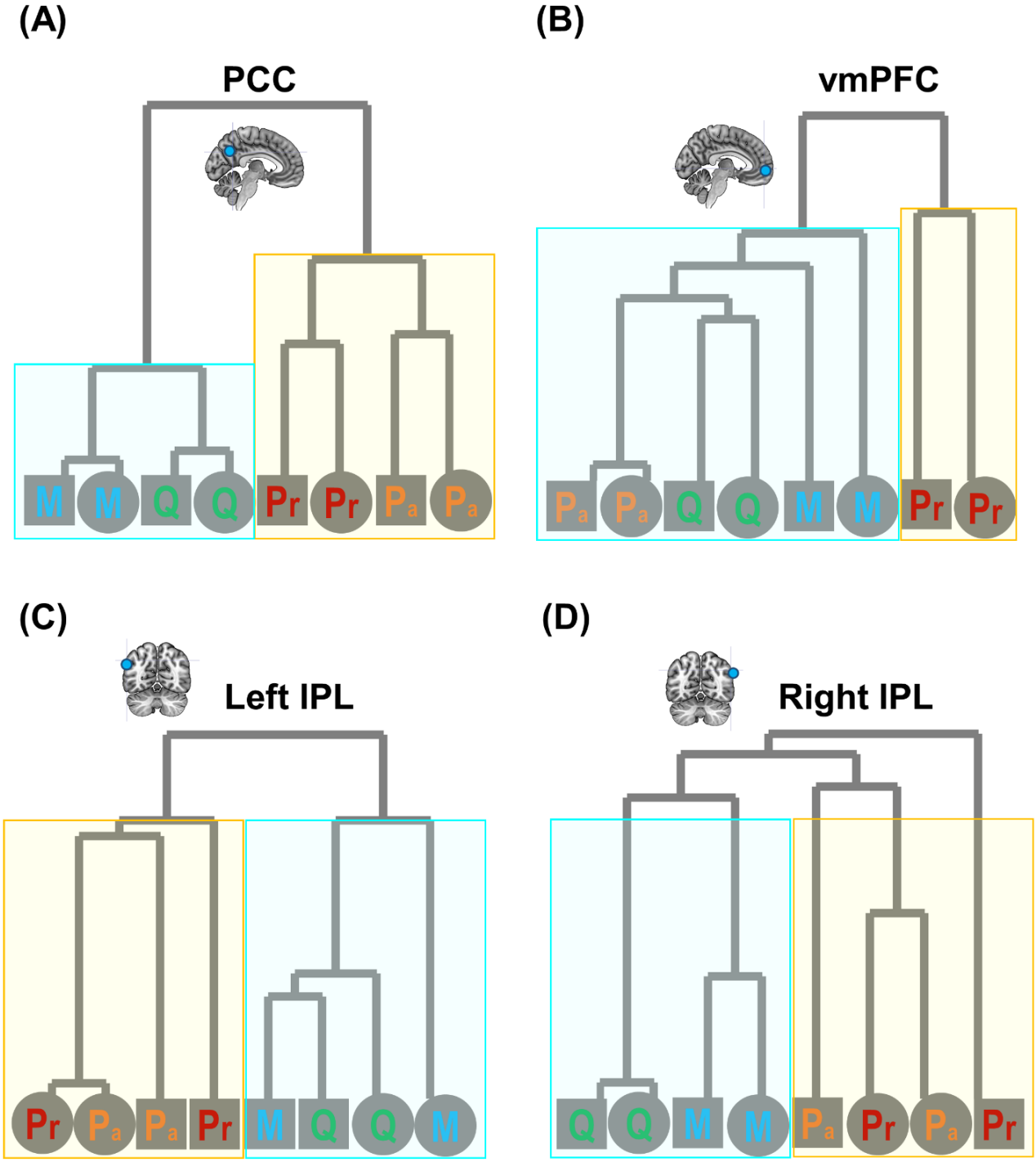


Figure 6. The configurations of each dendrogram tree here show a clear separation between an aggregate class of *self* (P / *Present Self* and Pa / *Past Self*; colour-coded using red and orange, respectively) and aggregate sense of other (*Q / Queen* and *M / Mother*; coded using green and blue, respectively). Corroborating the outcomes of cross-valence classification, here the representational structures show that emotive valence has little impact on the extent of neural similarity and the arrangement of overall pattern. Valence is coded by geometric shapes: circle – positive, square – negative. This reliable segregation between self and other is found in core DN regions of **(A)** the PCC, **(C)** the left IPL, and **(D)** the right IPL. The vmPFC – in Panel **(B)** – is somewhat different that it treats *Present Self* as a unique category different from all other categories.

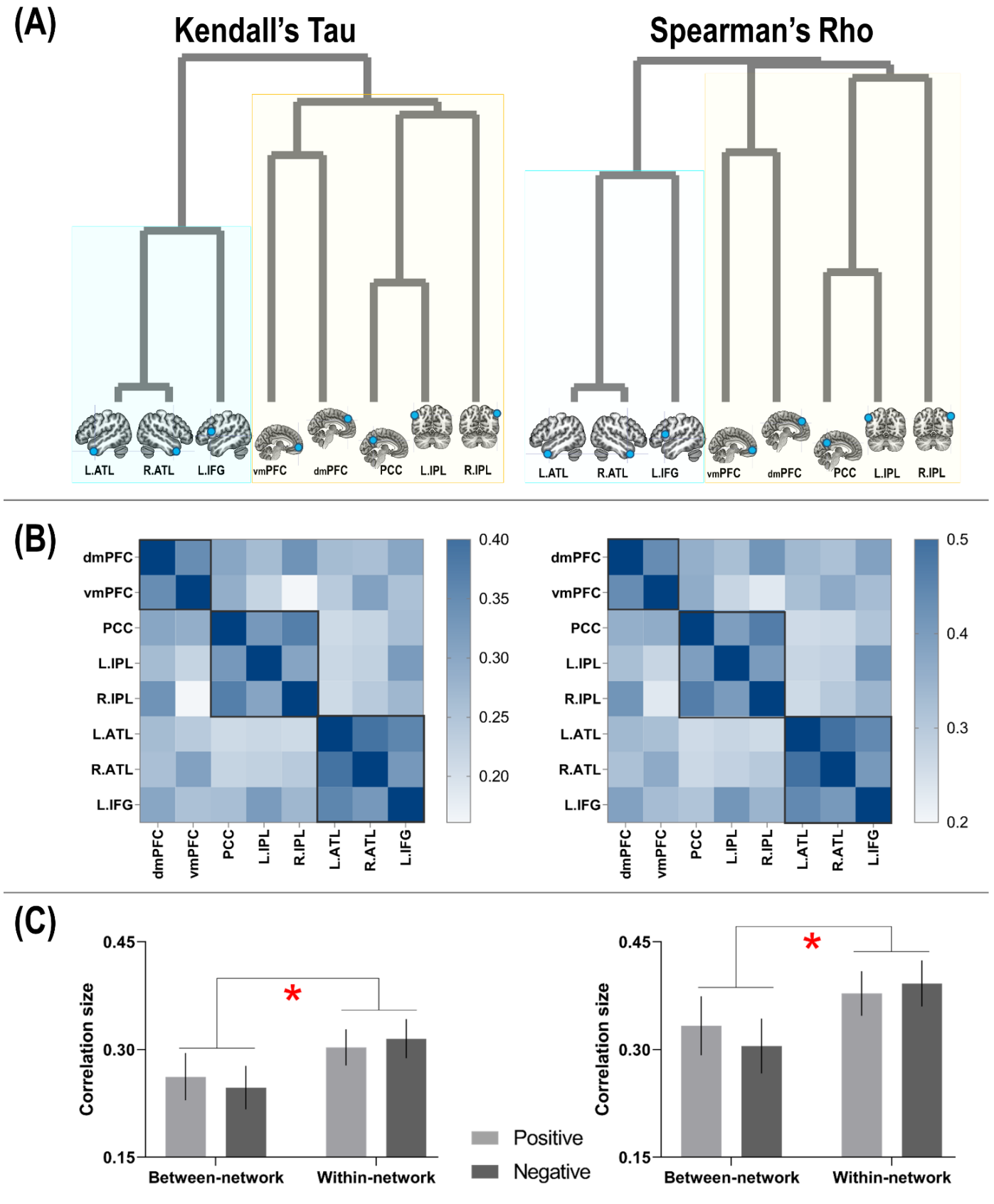

(A)
Kendall's Tau
Spearman's Rho
L.ATL
R.ATL
L.IFG
vmPFC
dmPFC
PCC
L.IPL
R.IPL
(B)
dmPFC
vmPFC
PCC
L.IPL
R.IPL
L.ATL
R.ATL
L.IFG
0.40
0.35
0.30
0.25
0.20
0.5
0.4
0.3
0.2
(C)
Correlation size
0.45
0.30
0.15
*
Between-network
Within-network
Positive
Negative

Figure 7. (A) The configurations of dendrogram show that neural similarity dissociated between the semantic network and default-mode network; this was seen both using Kendall's Tau and Spearman's Rho; **(B)** Here representational similarity is delineated using correlation matrix. A consistent pattern with the dendrogram structure is highlighted using black boxes; **(C)** Within-network representational similarity is significantly higher than between-network similarity, reliably found in both the contexts of positive- and negative-valence.

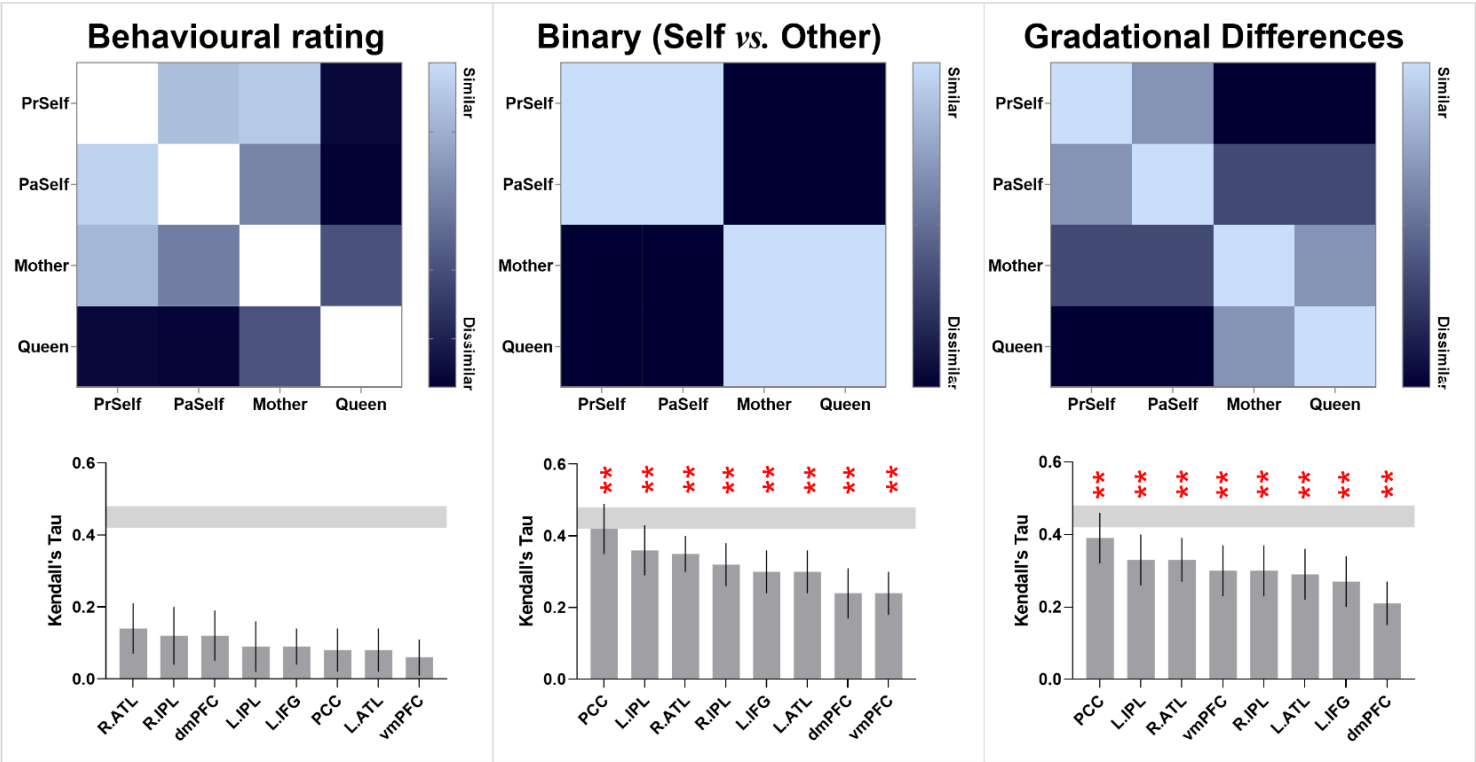


Figure 8. The hypothetical models that assumed a binary difference between self and other (middle) and gradational differences of the four individuals (right) are significantly correlated with the neural pattern of every region. The noise-ceiling is averaged across all regions. Multiple comparisons are controlled (FDR < 0.05) ** $p < 0.005$. By contrast, the outcomes of behavioural rating (left) were not correlated with the neural pattern of any region.