Testing cognitive theories with multivariate pattern analysis of neuroimaging data

Peelen, Marius V.; Downing, Paul

Nature Human Behaviour

DOI:
https://doi.org/10.1038/s41562-023-01680-z

E-pub ahead of print: 17/08/2023

Peer reviewed version

Testing cognitive theories using multivariate pattern analysis of neuroimaging data

Marius V. Peelen[1,] †,* and Paul E. Downing[2,] †,*

[1]*Donders Institute for Brain, Cognition and Behaviour, Radboud University, Thomas van Aquinostraat 4, 6525 GD Nijmegen, The Netherlands*

[2]*Cognitive Neuroscience Institute, Department of Psychology, Bangor University, Brigantia Building, Bangor, Gwynedd, LL572AS, United Kingdom*

†*These authors contributed equally*

* *Joint Corresponding Authors*

Correspondence:
marius.peelen@donders.ru.nl (M.V. Peelen)
p.downing@bangor.ac.uk (P. Downing)

Abstract
Multivariate pattern analysis (MVPA) has emerged as a powerful method for the analysis of functional MRI, EEG and MEG data. The new approaches to experimental design and hypothesis testing afforded by MVPA have made it possible to address theories that describe cognition at the functional level. Here, we review a selection of studies that used MVPA to test cognitive theories from a range of domains, including perception, attention, memory, navigation, emotion, social cognition, and motor control. This broad view reveals properties of MVPA, such as the ability to test predictions expressed at the item or event level, that make it suitable for understanding the “how” of human cognition, as well as limitations, and points to future directions.

Technological developments in human neuroimaging over the past few decades have led to an explosion of investigations into the full range of human cognitive abilities, including perception, attention, memory, navigation, emotion, social cognition, motor control, and more. In parallel, researchers concerned with understanding the mind from a functional point of view – what are the cognitive representations and processes that support human behaviour? – have regularly asked whether neuroimaging offers any useful answers to theoretical debates at that level of understanding[1–6].

In the past two decades, and at an increasing pace, researchers have turned to multivariate pattern analysis (MVPA) approaches to the design and analysis of human neuroimaging studies. MVPA capitalizes on the latent information found in patterns of brain activity that are distributed across voxels in an fMRI experiment or across channels in an MEG or EEG experiment (**Figure 1**). Researchers have claimed that these approaches would offer new ways to test mechanistic accounts of cognition[7–10]. The purpose of this review is to take stock of that claim by reviewing a wide sample of recent studies that have tested cognitive theories by developing hypotheses about the patterns of brain activity that emerge while participants perform tasks from many different domains.

Numerous recent reviews have examined MVPA studies from other perspectives, focusing on methodological aspects[8,11–13], philosophical considerations and in-principle limitations of the approach[14–16], applications in brain-computer interfaces[17], historical perspectives[18], "mind-reading"[19,20], and integrating MVPA studies with computational models[21] including deep neural networks[22–24]. In contrast, here we take a pragmatic approach to understanding whether and how MVPA has been used to shed light on theories about the "how" of human behaviour. Unlike previous theoretical and methodological perspectives, the aim of this review is to provide specific examples of studies that have successfully used MVPA to test cognitive theories. While these examples primarily highlight the strengths of MVPA, like any approach it also has limitations, which we discuss after the examples. Furthermore, by presenting these examples, we do not imply that MVPA is the only way to test cognitive theories with neuroimaging data. Finally, we briefly identify some future directions for research that should build on the findings and principles identified here.

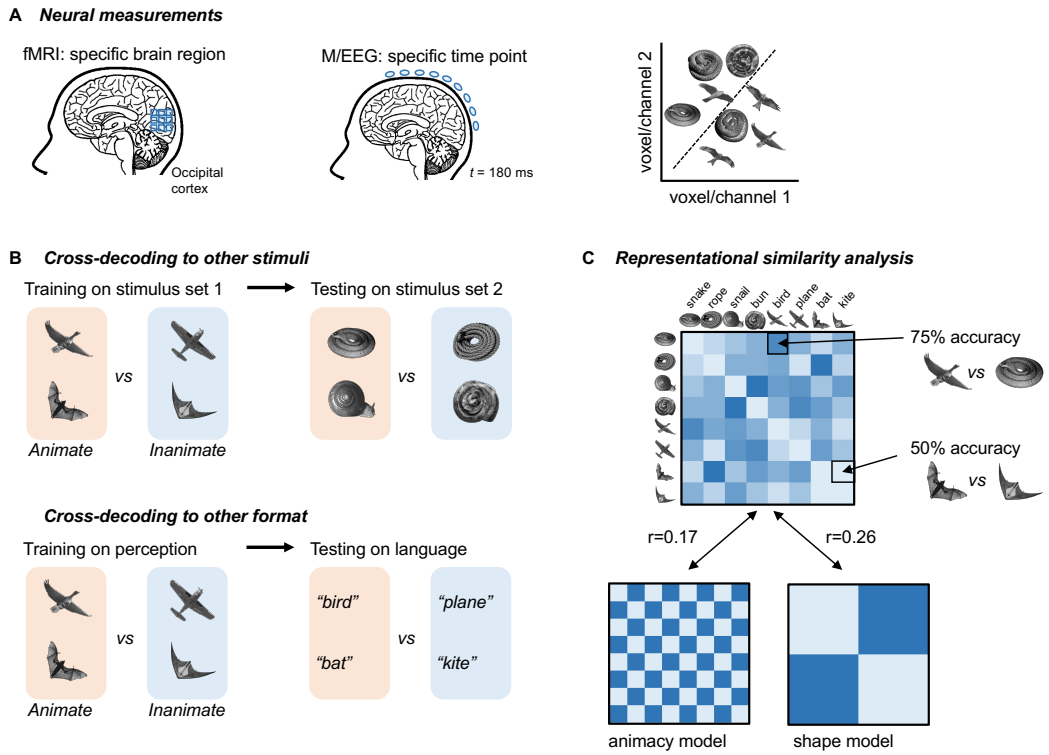

Figure 1. *Overview of key steps in multivariate pattern analysis of neuroimaging data, illustrated with the design of previous fMRI and M/EEG studies investigating the representation of object category and shape[48,116]. **A**) Neural activity is measured indirectly using fMRI or directly using M/EEG. fMRI provides high spatial resolution, with activity patterns measured across a set of voxels (each typically 2x2x2 mm) in specific brain regions (e.g., the occipital cortex). MEG and EEG provide high temporal resolution, with activity patterns measured across channels at specific points in time (e.g., 180 ms after image onset). Activity patterns across voxels (fMRI) or channels (M/EEG) are used to train a classifier to distinguish between two or more classes (e.g., snake vs bird). Typically, the classifier is trained on part of the data and tested on held-out data with the same conditions occurring in the training and testing sets (cross-validation). **B**) Alternatively, the classifier can be tested on different but related conditions, to test for generalization (cross-decoding), for example across different stimulus sets or across format. **C**) Pairwise decoding accuracy can be used as a distance measure in representational similarity analysis[45]: object pairs that are accurately classified are representationally dissimilar (illustrated by darker cells). Other distance measures can also be used[132]. The resulting representational dissimilarity matrix (RDM) of a particular brain region (fMRI), or at a particular time point (M/EEG), can then be correlated with models derived from cognitive theories that make different predictions about the structure of representations. Here, for example, two models are illustrated that emphasize either the status of an object as (in)animate, or its overall shape.*

What is multivariate pattern analysis?

MVPA studies vary widely, but generally depend (sometimes implicitly) on the assumption that neural activity patterns (the distribution of activity across a set of voxels or channels) index the structure of a mental representation or process. In its most basic form, MVPA can be used to test whether the activity patterns in a given brain region are reliably distinct for two different stimulus classes (**Figure 1**). If this is the case, then that region may be considered to represent some dimension that distinguishes those classes. However, the finding of above-chance

classification alone may not provide much theoretical insight, as there are typically many dimensions on which two conditions differ that could drive the classification[15,25]. Therefore, many studies have used more complex approaches to relate brain activity patterns to measures of behaviour, to judgments of (dis)similarity, or to parameters derived from formal computational models. As we will see in the examples below, two approaches have been particularly fruitful for testing cognitive theories: cross-decoding (**Figure 1b**) and representational similarity analysis (**Figure 1c**). Still other studies adopt the logic that multivariate patterns provide a "signature" that indexes the degree to which one of several possible mental states is more engaged in a given task[26]. Finally, multivariate decoding approaches are complemented by encoding methods[27–29]. Unlike decoding, which aims to predict experimental conditions from neural activity patterns, encoding aims to predict neural activity patterns from experimental conditions. While all these approaches have strengths and weaknesses[12,27,29], the claim evaluated here is that they can provide insight, at a fine-grained level, into the links between neural activity patterns and the mechanisms or representations implied by cognitive theories.

As with any other measure of neural activity, the inferential strength of an MVPA finding depends on other factors. Stronger claims will be supported, for example, by results that are robust across different tasks, stimulus items, and participants; by demonstrating a systematic relationship between patterns and overt behaviour; or by demonstrating specificity to brain regions or time windows. Understanding MVPA results from fMRI or M/EEG at the underlying neurophysiological level has proved challenging, as illustrated by debates about the interpretation of above-chance decoding of visual orientation from activity patterns in primary visual cortex[30–35]. However, as illustrated below, MVPA has been successful even without a complete picture of the neurophysiological basis of the activity patterns that differentiate between conditions.

What is a cognitive theory?

A cognitive theory is one that explains how a behaviour emerges from processes and representations at a level that abstracts away from the specific neural substrate. This corresponds closely to the "algorithmic" level of Marr's influential analysis of the tasks confronted by vision, a level that lies between the properties of neurons and neural networks (the "implementation" level), and a description of the problem the organism faces and the relevant information that is available in its environment (the "computational" level)[36,37]. Cognitive theories adopt an information processing perspective to describe mental representations of internal states and of the external world. Representations are powerful because they can make explicit some latent dimensions, while obscuring others, thereby supporting behaviours that rely on those exposed dimensions. By way of analogy, Arabic representations of number make units of ten explicit in a way that Roman numerals do not, so that decimal operations are trivial to perform in one (8 x 10 = 80) but not the other (VIII x X = LXXX). Cognitive theories also take up the formation, manipulation, retrieval, and use of representations: that is, cognition encompasses active mental processes as well as stable mental representations[38,39]. A key aspect of the cognitive approach is the idea that multiple relatively simple processes and representations can interact in different ways, depending on the actor's goals, to produce a wide range of complex behaviours.

Specifying a cognitive model entails answering several questions, such as: what kinds of inputs must be represented, and what kinds of latent information must be extracted from those inputs? What is the format and durability of stored representations? What is the number and kind of processes that manipulate, store, retrieve, or draw inferences about represented information? What are the capacity limits or “bottlenecks" of these processes, and to what extent do they interact with each other? How does information flow within and between processes or stages?[40]

As we review examples of specific studies in the following section, we will show how neuroimaging studies have used MVPA to address some of the questions and problems that emerge from cognitive theories of behaviour. Outside of our scope, there are theories of information representation and transformation that are fundamentally neuroscientific, aimed at capturing the properties of a given brain region, pathway, or network[41]. Likewise, while we are concerned here with MVPA tests of cognitive theories, this logic is sometimes reversed, as when cognitive theories are invoked to explain patterns of activity in neuroimaging results[2]. Finally, while cognitive theories are also used as a framework to understand individual differences[42], psychiatric conditions[43], or cross-cultural variation[44], our focus here (without denying the importance of those topics) is on mental universals, in the tradition of cognitive psychology.

What makes MVPA suitable for testing cognitive theories?

What are some of the properties of MVPA logic that suit it to testing hypotheses generated by models of cognition? We do not claim that other approaches to the design and analysis of neuroimaging experiments (e.g., repetition suppression or mass univariate studies) cannot achieve similar aims. Rather the emphasis is on features of MVPA that naturally align to testing cognitive models. One feature that stands out is that MVPA is readily used to index representations with fine granularity, by measuring the brain states that are tied to specific items, events, or experiences. This property enables contact between theory and data at a level that is required for testing predictions effectively.

Rich item-level data sets have proved powerful in the representational similarity approach[45,46] to MVPA that characterises neural spaces through dense measurement of inter-item similarity. These descriptions can be related across imaging methods (e.g. fMRI-MEG fusion[47]), and also to overt measures of behaviour such as reaction times in search tasks[48,49] or explicit similarity ratings[46], as well as to similarity spaces derived from computational models[50,51]. Importantly, this approach improves the specificity of predictions about neural activity, and hence the ability to distinguish competing models, by going beyond simple “point” predictions (A>B, or A>0) that are known to provide a weak basis for theorising[52,53]. In turn, more exploratory studies, or those motivated more by neuroscientific aims, may reveal neural representational spaces that were not expected from cognitive theories, but that nonetheless inform them.

Neuroimaging studies have often taken advantage of the ability to covertly index mental states without perturbing the behaviours that generate those states[54–57]. This feature has proven still more powerful in combination with the sensitivity of MVPA, for example in studies of activity related to perceptual events that should be ignored[58,59], to activation states in the delay interval of working memory tasks[60,61], to contextual information that is retrieved from memory

incidentally[62], or to the preparation to generate overt behaviours[63,64]. Finally, MVPA can be particularly informative when used to test for generalization by using a cross-decoding approach that compares activity patterns across domains, modalities, tasks, time, or individuals (**Figure 1b**)[65–68]. For example, this approach allowed researchers to show that perceptual aspects of the encoding context are reinstated during retrieval, that emotions and mental states are shared across tasks and individuals, and that visual object representations are activated by attention cues and modulated by scene context, as reviewed in more detail below.

Applications of MVPA to testing cognitive theory.

In each section below, we introduce a cognitive capacity of interest, and describe one or more influential theories from that domain. We then illustrate how MVPA studies of adult human participants have assessed the predictions of those theories or informed debates between theories. Together, these examples highlight the ways in which multivariate neuroimaging methods, leveraging some of the design and analysis advantages noted above, can speak to core issues that have shaped theorising about human cognition for decades. Space limits mean that many other worthy examples have been omitted.

Spatial navigation

Tolman[69] argued that the sophisticated mammalian capability for navigation is enabled by symbolic representations of the spatial environment that support inducing new paths to previously visited locations (commonly referred to as "cognitive maps"[70]). While neuroimaging methods typically require participants to remain immobile, clever task design has enabled MVPA tests of key predictions of this general theory of the neurocognitive processes supporting human navigation. Specifically, MVPA revealed distinct representations of spatial location (one's position in a mental map) and facing direction (one's orientation at a given location) in the medial temporal lobes[71] while participants viewed images of different familiar campus locations that were photographed from each of the four cardinal directions (**Figure 2**). A related study[72] revealed how we encode experienced episodes. Participants wore photo-capturing devices that recorded time- and location-stamped egocentric images over a month. They later viewed some of these images during fMRI, and tried to recall the depicted episodes. Hippocampal activity patterns were more similar for more proximate than more distant autobiographical events, both in terms of their distance in space as well as in time. These studies illustrate how the cognitive map concept has been elaborated by MVPA approaches to human neuroscience, by showing functional dissociations amongst navigation mechanisms, and relating neural activity patterns to real-life experience using representational similarity analysis.

More generally, the evidence speaks to cognitive theories about long-term knowledge representations, showing that the encoding of time and spatial location are partly shared, and yet distinct from the representation of heading direction.

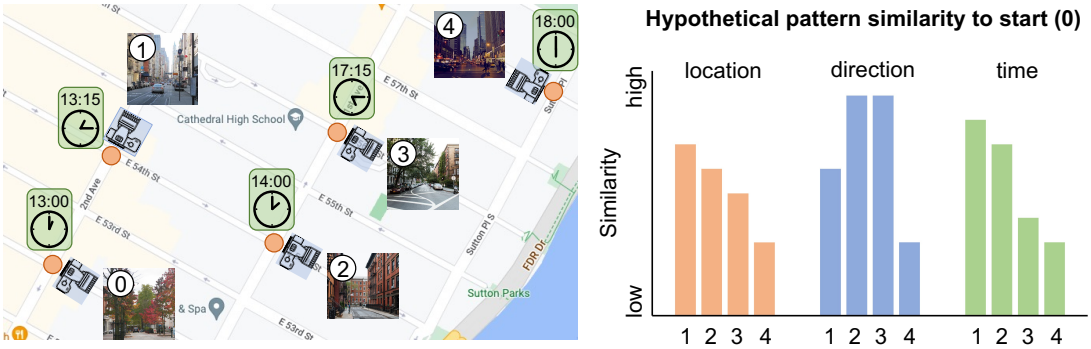

***Figure 2**. As we navigate through space and time, we track our position in familiar locations (orange spots) and the direction we are facing (blue cameras), and we encode memories of specific events at specific times (green clock icons). Studies using MVPA of fMRI data have investigated patterns of brain activity that emerge as participants view locations from familiar environments[71] or recall episodes captured in images from their own daily lives[72]. In medial temporal brain regions, distributed patterns of activity demonstrated some key properties in accordance with map-like cognitive representations. Specifically, in representations of location, patterns were more similar for nearby than distant locations; likewise, in representations of direction, neural patterns were more similar for similar than distinct directions; and finally, events experienced closer in time were captured in more similar activity patterns than more temporally distant events.*

Object perception

Behavioural studies have shown that object recognition is strongly facilitated by scene context[73,74], but there has been debate about the underlying mechanisms. According to interactive accounts, the visual processing of objects and scenes interacts, such that object processing is modulated by expectations derived from scene context[75]. Alternatively, information from objects and scenes may be processed in parallel, with facilitation resulting from post-perceptual evidence integration[76]. Recent studies applied MVPA cross-decoding to fMRI and MEG data to provide evidence for the interactive account[77]. Participants viewed ambiguous objects (e.g., degraded by blurring) within or outside of scene context. Multivariate activity patterns evoked by the ambiguous objects in object-selective visual cortex (measured with fMRI) at 300 ms after stimulus onset (measured with MEG) became more similar to the activity patterns evoked by intact objects (determined in a separate experimental run) when the ambiguous objects were presented in scenes. Because the scenes alone did not evoke object-specific activity patterns, the modulation could not be explained by additive processing of scenes and objects, thus providing evidence for interactive accounts.

In this example, MVPA was used to index the processing of within-scene objects by relating visual cortex response patterns evoked by degraded objects in scenes to canonical (intact) object responses in isolation. Through this cross-decoding approach, differences between conditions could be related to the modulation of visual object processing, which was key to testing the predictions of contextual facilitation models.

Attention

Attentional mechanisms allow cognitive processes to focus on currently relevant information[78]. According to the influential biased competition theory of attention[79], attention biases the

competition between neural representations of multiple simultaneously presented stimuli (e.g., visually presented objects) in favour of the attended stimulus. Neuroimaging studies have used MVPA cross-decoding to provide evidence for a central component of the biased competition theory (and other theories of attention): the attentional template[80]. In visual search tasks, when attention is directed to a particular stimulus attribute, such as an object's shape or colour, neurons in macaque visual cortex representing that attribute show increased activity, even before a search display is presented[81]. This *preparatory activity* constitutes a top-down attentional bias[82], facilitating the processing of subsequently presented target objects. fMRI studies have used MVPA to show that a classifier trained to distinguish between activity patterns evoked by visually presented objects can predict what specific object participants prepared to search for in separate trials, in the absence of visual input[83,84]. Recently, this approach has been used to go beyond what was known from animal work, showing that such attentional templates incorporate contextual expectations during naturalistic visual search[85].

In this example, MVPA was used to address hypothesised intervening stages between the decision to act and the act itself, providing evidence for an attentional template that mediates between a verbal understanding of a task and the active processing of a search array. Here, a key methodological strength was the ability to measure indices of target objects at an item-specific level, as previously achieved in neurophysiological studies.

Memory

Most cognitive theories of memory distinguish between encoding, storage, and retrieval stages. According to the encoding specificity theory[86], memory performance depends on the similarity between encoding and retrieval contexts. Accordingly, remembering a particular fact or episode may include mentally reinstating the encoding context, for example the place where the memory item was encountered. In an MEG study[87], participants learned arbitrary associations between visual images (of scenes or faces) and words (**Figure 3A**). In the retrieval phase, participants were only presented with the words, reporting whether they had seen the words before. Multivariate classifiers were trained on responses evoked by the images presented without the words (in the encoding phase) and tested on responses evoked by the words without the images (in the retrieval phase). Results showed that the category of the word-associated images could be decoded at retrieval, indicating that the encoding context was reinstated. Importantly, by training and testing classifiers at different time points, the analysis showed that the activity pattern at 180 ms after image onset (during encoding) was reinstated around 500 ms after the word retrieval cue (**Figure 3B**), providing evidence that relatively early stages of encoding are reinstated during recollection.

This study thus cleverly used MVPA cross-decoding to provide evidence for contextual reinstatement by associating activity patterns during retrieval with content-specific activity patterns during encoding, at an item-specific level.

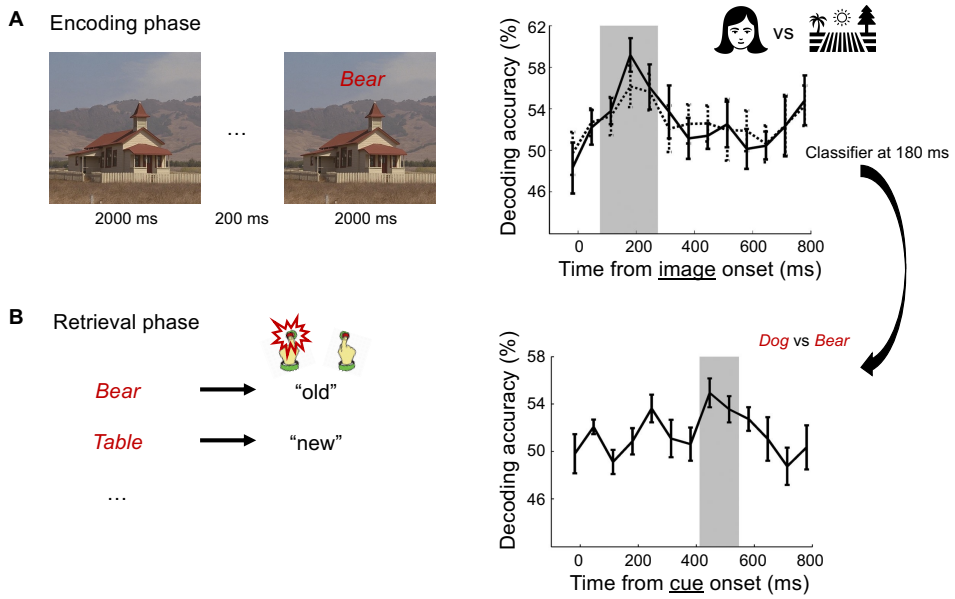

Figure 3. *Schematic summary of an MEG study investigating contextual reinstatement during memory retrieval*[87]*.* ***A****) During the encoding phase, participants viewed pictures of scenes and faces. Each image was first shown in isolation and subsequently together with a semantically unrelated word. Participants were instructed to memorize the picture-word associations. Classifiers were trained on MEG data to distinguish between the visually presented faces and scenes (before words were presented). Decoding peaked at around 180 ms after image onset.* ***B****) During the retrieval phase, participants read words and reported whether or not the word had been presented during the encoding phase. MEG activity patterns carried information about the image category that had been presented together with the words during the encoding phase: a classifier trained to distinguish between faces and scenes at 180 ms after image onset could significantly decode the cue-associated image category at around 500 ms after cue onset. This provides evidence that the word cues activated an early perceptual representation of the encoding context, in line with the contextual reinstatement hypothesis.*

Category Learning

Researchers have used MVPA to test cognitive theories about how new categories are learned. One such study[88] directly compared two formal models, applied to a trial-and-error task requiring participants to assign geometric shapes to one of two categories. Each model posits different internal representations of learned categories: the *prototype* model[89] describes an abstracted encoding of category-defining features, while the *exemplar* model[90] instead emphasizes representations of individual category exemplars. At the behavioural level, each model described categorization decisions equally well. Whole-brain MVPA of fMRI data examined distinct predictions made by the two models about the activity patterns that would be evoked by each learned exemplar. The observed patterns of brain data were significantly more consistent with representational states predicted by the exemplar model, relative to the prototype model. Further analysis demonstrated a close link between patterns of activity in key occipital, parietal, and lateral prefrontal regions, and parameters from the exemplar model that describe attention to category-defining object features.

This study provides a clear example of how the hypothetical representations of two competing models may be translated into item-level predictions about the patterns of neural

activity evoked by those items, revealing mental encoding principles that were not distinguishable at the level of behaviour.

Conceptual knowledge

Conceptual knowledge describes the rich information we can retrieve and manipulate about categories of objects, events, and abstractions such as “chair”, “party”, or “freedom”. What are the key dimensions that describe meaningful, real-world concepts, once they are learned? Researchers have tested the claim from embodied (or grounded) cognition theories that direct sensorimotor experience of the world pervades our mental representations of concepts, even abstract ones[91,92]. MVPA provides a test of this proposal by searching for neural signatures of such first-hand sensorimotor experience during cognition about concepts. A recent fMRI study used representational similarity analysis to compare models of the dimensions that describe knowledge of abstract and concrete concepts, including objects and events, expressed in single words[93]. These different models were based on: 1) taxonomic relationships (e.g., Pippin -> Apple -> Fruit); 2) patterns of local co-occurrence in large text corpora, an index of semantic proximity; and 3) overlap in features of shared “experiential” content that related to sensory, motor, spatial, and affective properties. Across a range of brain regions, the third model was most effective at capturing variance in the activity patterns evoked by the tested concepts. This finding supports the grounded cognition proposal that abstract knowledge about concepts is constituted, at least in part, from a mixture of modality-specific experiences.

This study is an example of how MVPA can index representations with fine granularity. The ability to measure similarities between item-level activity patterns allowed for tests of competing models that describe how semantic knowledge is organised.

Emotion

A long-standing debate in the field of emotion research revolves around the role and importance of basic emotions[94,95]. According to discrete emotion theory[96], a small number of basic emotions (e.g., joy, anger, fear, disgust) are the building blocks of emotional states, each characterized by specific and universal behavioural, physiological, and neural signatures. MVPA has been used to provide evidence for such emotion-specific neural correlates. In one study, participants either viewed short movies depicting the basic emotions or mentally imagined being in a particular emotional state[97]. Classifiers trained to distinguish emotion categories during movie viewing could classify emotion categories during mental imagery, suggesting that basic emotions are supported by discrete neural signatures that generalize across emotion-eliciting conditions. However, evidence against basic emotion theory comes from an fMRI study that investigated the neural similarity of 20 complex emotional states (e.g., jealousy, nostalgia) inferred from brief verbal narratives[98]. For each narrative, participants indicated the degree to which it elicited each of the six basic emotions. These data were used to construct a similarity space, in which complex emotions that elicited basic emotions in similar ways were represented as relatively more similar to each other. Subsequently, MVPA was used to obtain the neural similarity of the activity patterns evoked by the 20 conditions in mentalizing regions (medial prefrontal cortex; temporo-parietal junction) that were previously shown to distinguish basic emotions[99]. Results from this representational similarity analysis showed that a model based on appraisal theory[100,101] significantly outperformed the basic emotion model at explaining the mental “space”

of complex emotions, suggesting that these are high-dimensional and cannot be fully reduced to a combination of basic emotions.

These studies demonstrate how MVPA can provide information about the dimensions that structure our knowledge and experience of emotions, with complex emotions only partly described by reference to basic emotion categories.

Social Cognition

Successful navigation of the social world requires recognizing that the sensations and mental states of other humans can differ from our own. According to simulation theory[102], we understand others' minds through a simulation that is grounded in shared perceptual processes. In contrast, according to the theory-theory[103], social cognition is more like inductive reasoning, in which propositions are assessed to make assessments of others' likely mental states. To provide evidence for the simulation theory, one fMRI study[104] induced aversive emotions (e.g. pain, disgust, and unfair treatment) in participants, who also observed a friend experiencing the same emotions. Cross-decoding analyses (e.g., train on painful vs non-painful stimulation, test on disgusting vs neutral gustatory stimuli) revealed patterns in cingulate and insular regions that generalised across both the aversive stimulus type (pain, gustatory) and, importantly, also over the recipient (self, other). These results support the simulation theory in that they identify shared neural encoding of directly and vicariously experienced sensations. However, MVPA has also provided evidence for the theory-theory[105]. Sighted and congenitally blind individuals listened to stories that described a protagonist learning something of either positive or negative valence, either through the visual modality (e.g., the protagonist sees a break-up note from his partner) or auditory modality (e.g., the protagonist learns her job application was successful via a voicemail). Multivoxel patterns in the right temporal-parietal junction (rTPJ) encoded the information source of the narratives (heard vs seen) and, crucially, did so equally for blind and sighted participants. These results support the theory-theory in that they evidence a concept of "seeing" that is not grounded in first-hand experience (absent in the blind participants), but rather an abstract one that is derived from second-hand experience, such as through language.

Representations make some dimensions explicit while obscuring others. Here we see how some of the representations of emotionally adverse experiences revealed with MVPA make explicit the valence of an emotional experience while obscuring whether it is the self or another person who is having that experience.

Sensori-motor prediction

Influential theories in the domain of motor behaviour state that actions are accompanied by efference copies that serve to predict and suppress the sensory consequences of one's own actions[106,107], reducing self-generated sensations. Research in animals has supported this idea, for example by showing that primary somatosensory cortex (S1) encodes motor-related activity before movement initiation[108]. Two recent fMRI studies used MVPA to provide evidence for similar anticipatory signals in the human brain. In one study[109], participants performed a delayed object manipulation task in the scanner. On each trial, a cue indicated the action to be performed on that trial (e.g., lift object with right hand). Results showed that the effector used in the action (left or right hand) could be decoded from S1 activity patterns during the delay period before the movement, in the absence of sensory input. A related study[64] investigated the

planning of movements at a finer scale (**Figure 4A**), at the level of individual finger presses (thumb, middle finger, little finger). Multivariate activity patterns in S1 carried information about the specific finger that participants planned to move, even on no-go trials, where the action was planned but not executed. Finally, representational similarity analysis revealed that the finger-specific activity patterns during the planning phase resembled the finger-specific activity patterns during the execution phase (**Figure 4C**). These studies provide converging evidence that motor planning activates predicted sensory consequences in S1, in line with classical theories of motor control in which an internal forward model predicts future body states and their sensory correlates.

In this example, the ability of MVPA to distinguish the activity profiles of individual digits shed new light on the way we prepare to produce complex behaviours and anticipate their sensory consequences.

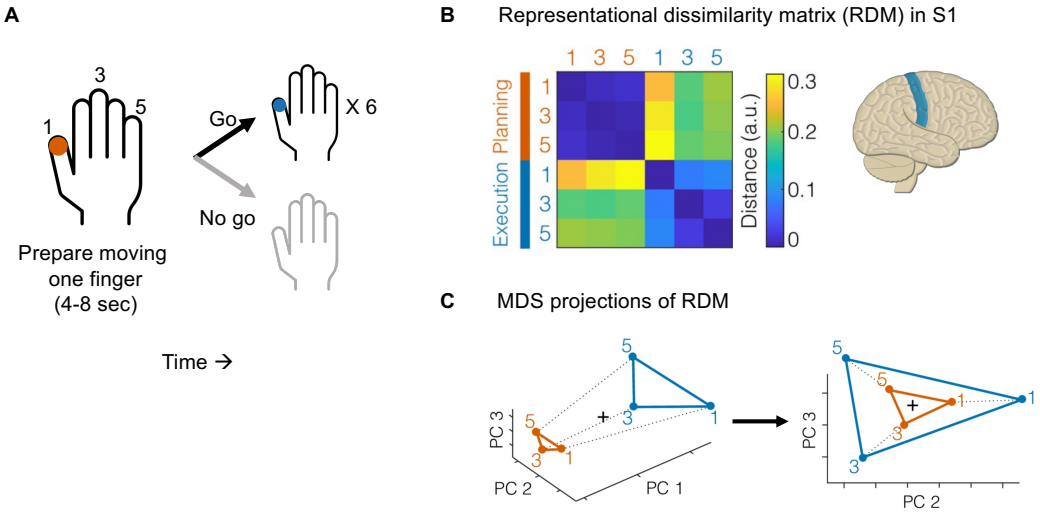

***Figure 4**. Schematic summary of an fMRI study investigating motor prediction[64]. **A**) Single finger movements were cued before an extended motor planning interval of varying duration (left). On "go" trials, participants carried out the movement; it was withheld on "no go" trials, which were further analysed to segregate "pure" planning processes. **B**) Representational dissimilarity matrix in contralateral primary somatosensory cortex (S1) for the planning and execution phases, measured with fMRI. **C**) Multidimensional scaling (MDS) illustration of the correspondence between the relative similarities of activity patterns across planning (orange) and execution (blue) of single finger movements (1=thumb, 3=middle, 5=little). While the main distinction was between planning and execution (principal component 1 (PC1); left panel), the preserved geometry from preparation to execution (PC2-PC3; right panel) supports the proposal that planned behaviour in part entails pre-activation of anticipated sensory outcomes of behaviour.*

Sequential Motor Behaviour

Learning to play even a simple piano tune involves pressing the right keys in the right order at the right times. How does a parallel brain produce such series of behaviours coherently? One family of models posits a chain of associations, such that each behaviour acts, via associative processes, to trigger the next[110]. In contrast, competitive queuing models propose that action sequences are encoded in a parallel scheme[111]. A serial model predicts that preparation of the

first act alone should suffice to elicit the action chain, while a competitive queuing model predicts simultaneous pre-activation of all effectors: the first in the sequence most strongly, and each other in proportion to its position in the temporal sequence. An MEG study supports the competitive queuing model[63]. Participants learned sequences of timed finger keypresses (**Figure 5A**). At test, separate preparatory and performance phases enabled each to be analysed separately. Using a cross-decoding approach, linear classifiers were trained on sensor data acquired during performance and then tested on activity patterns that were evoked during preparation, as an index of the activation state of each finger. Results revealed the predicted ordinal arrangement of preparatory activity (**Figure 5B**), as predicted by competitive queuing models.

This study reveals how multivariate measures of neural activity offer a sensitive and non-intrusive measure of the internal mental states that precede overt behaviour. In this way they are able to test the implications of formal models that make detailed and time-specific predictions about those internal states.

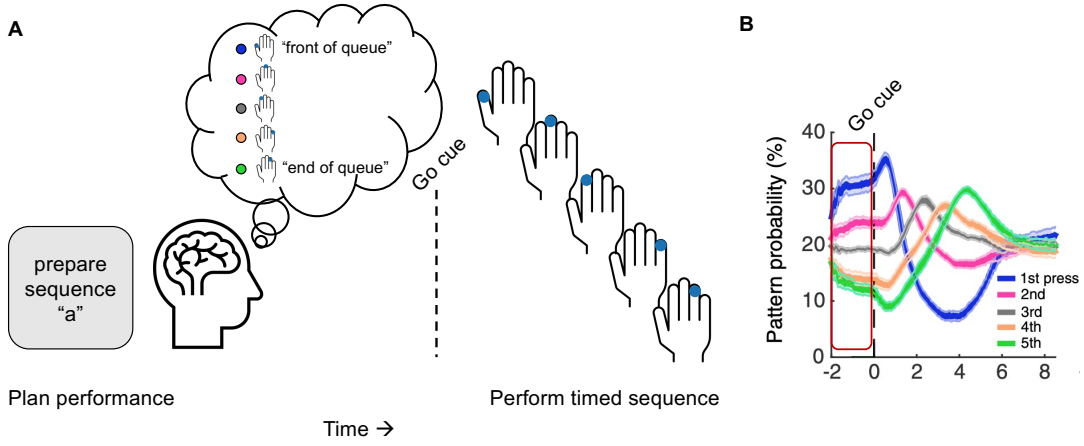

Figure 5*. Schematic summary of an MEG study investigating sequential motor behaviour[63]. **A**) A cue signalled which of several learned finger-movement sequences to prepare, initiating a planning interval. Next a "go" cue instructed the participant to execute the sequence. **B**) MEG recordings captured the strength of activation of patterns representing each finger movement during the planning interval (left, before "go" cue). These were ordered systematically such that the first planned movement dominated the observed MEG patterns, followed by the second, third, and so on, in an orderly queue. This finding is predicted by formal models of "competitive queuing" that postulate a mechanism by which a parallel system can produce serial behaviours.*

Limitations of MVPA for testing cognitive theories

While the aim of our review is to show how MVPA of neuroimaging data can be (and has been) used to test cognitive theories, there are clearly also limitations to this approach. Most of these limitations are shared with other correlational neuroscience methods, some are specific to one or multiple neuroimaging methods (e.g., fMRI, MEG, or EEG), and still others are specifically related to MVPA. Here, we briefly mention some of these limitations, and describe how they apply to several of the example studies reviewed above.

One well-known limitation of correlational methods is that they do not provide evidence that a particular neural measure (e.g., activity of a neuron) causally contributes to the cognitive

process of interest. For example, the studies that used MVPA to reveal object-specific preparatory activity in visual search tasks interpreted this as evidence for an attentional template[83,84]. Similarly, studies that demonstrated the activation of encoding context during memory retrieval assumed this activation to contribute to memory performance[87]. However, an alternative interpretation for these findings is that this such activity reflects epiphenomenal mental imagery that is unrelated to attentional selection or memory. While task design and correlation of neural measures with behavioral measures[112,113] can go some way to alleviating these concerns, ideally MVPA findings are supplemented by causal evidence such as provided by TMS. This approach has been successfully used, for example, to demonstrate that the contextual facilitation in visual cortex observed with MVPA causally contributes to object recognition[114].

MVPA can be highly sensitive to small differences, complicating the interpretation of above-chance decoding. Specifically, the finding that two conditions can be decoded above chance in a particular brain region (in fMRI) or at a particular time point (in M/EEG) is not necessarily informative about the underlying processes driving the decoding. For example, decoding whether a participant experiences aversive or neutral events[104] could be driven by a range of processes, including sensory processing, affective responses, (preparation of) defensive actions, or even artifacts such as small eye or body movements[115]. For this reason, researchers have adopted the cross-decoding approach (**Figure 1b**), in which classifiers are trained on one distinction (e.g., experiencing pain) and tested on another distinction (e.g., viewing someone else in pain) to reveal commonalities and thereby inform interpretation. An advantage of this approach is that demonstrating how classification generalises over a range of materials, tasks, or other contexts reduces the possibility that confounding variables specific to one of those contexts explains decoding performance. Finally, as we have seen, another informative approach is to combine MVPA with RSA to relate neural similarity to the similarity between conditions in competing cognitive models (**Figure 1c**).

While MVPA is typically more sensitive than univariate analyses, it is still limited by the spatial resolution of neuroimaging methods. As such, it is likely that many neural signals are not expressed at the level of voxels (fMRI), electrodes (EEG), or sensors (MEG). Therefore, finding no evidence for a particular neural distinction cannot be taken as evidence that such a distinction does not exist in the brain. For example, the lack of a difference in decoding the source of narratives (heard vs seen) in blind and sighted participants[105] could reflect the insensitivity of MVPA to such information that may be present at other scales. Furthermore, the sensitivity of MVPA differs across brain regions and neuroimaging methods (e.g., fMRI vs MEG[116]), complicating direct comparisons and multimethod integration.

Finally, a general concern that has been raised about neuroimaging research is the “consistency fallacy”. Neuroimaging results that are merely consistent with one cognitive theory cannot be taken as strong evidence for that theory[2,117]. To be most informative, the results should additionally be inconsistent with one or more alternative theories. One example from the studies reviewed here is the finding of above-chance decoding of basic emotion categories[97]. While this finding is certainly consistent with the discrete emotion theory, above-chance decoding is also consistent with alternative theories (e.g., appraisal theory). Indeed, subsequent studies used MVPA in combination with RSA to test more detailed predictions of these theories, showing that the representation of complex emotions cannot be fully captured by a combination

of basic emotions[98]. We suggest that MVPA in combination with cross-decoding or RSA allows for more informative tests of cognitive theories, but these approaches do not in themselves provide a substitute for well-articulated alternative models.

Thus, while neuroimaging and MVPA are not without limitations, many shared with other approaches, we believe that the examples provided in this review show that, when used appropriately, MVPA can be a useful method to test cognitive theories.

Future directions

Further developments in analysis methods promise to make MVPA even more useful for testing cognitive theories. For example, the "hyperalignment" approach[68] aligns representational spaces (as opposed to anatomy) across individuals, making it possible to distinguish aspects of the geometry of those spaces that are shared across individuals from idiosyncratic variation. This approach extends the improvement in representational granularity provided by MVPA from items and events to individual participants[118,119]. Reliable measures at the single-participant level allow tests of hypotheses about links between brain measures and individual differences in behaviour[120], which in turn may address new predictions that arise from cognitive models[121]. Second, multivariate connectivity methods[122,123] measure connectivity in information shared across brain regions, rather than activation *per se,* an approach that may lend itself to cognitive theories that describe transformation of representations in stages over time. Finally, rapid developments in the field of artificial intelligence have generated new computational models and improved decoding algorithms[124], which may in turn provide more sensitive and robust descriptions of the neural patterns that correspond to hypothesised cognitive representations.

Improvements in the quality and resolution of neuroimaging data will also bring new opportunities for testing cognitive theories. For example, high spatial resolution fMRI data, combined with MVPA, allows distinguishing representations within specific cortical layers, which has revealed similarities and differences between representations activated during cognitive tasks of working memory, attention, prediction, and imagery[125]. Further, recent developments in high-density mobile EEG systems now allow for measuring neural activity during natural behaviour[126]. Combined with MVPA, this opens up the possibility to investigate cognitive processes in more ecologically-valid environments and tasks[127], which may provide better ways to test how cognition depends on complex contextual variables than possible in standard controlled laboratory tasks[128].

In sum, these recent analytical and technical developments improve on existing advantages of MVPA for testing specific, granular predictions that cognitive models make about patterns of brain activity and their variability over time, anatomical location, participants, and contexts. Of course, better data and more powerful analysis methods alone are no substitute for continued progress in developing the cognitive theories that we aim to test using these methods. Ideally, such theories are computationally explicit and provide a mechanistic explanation of the cognitive process under investigation[129–131]; among other advantages, those features will tend to support specific predictions about patterns of neural activity in space and time that are measurable with MVPA. Researchers' questions are often shaped, at least in part, by the methods and tools available to them. As applications of MVPA to cognitive theories continue to prove their value, we believe researchers will see in these methods new ways to

develop, improve, and refute those theories, across the span of the human behavioural repertoire.

Conclusion

Since their advent, human neuroimaging methods have attracted criticism of their ability to go beyond localization of neural events, whether in anatomical regions or in time, and thus whether they would be able to shed any light on functional accounts of human behaviour. This criticism and the ensuing responses have rapidly and productively shaped the way researchers approach cognitive neuroimaging; the large-sample, data-rich, and hypothesis-driven studies of today represent a marked leap forward from the relatively simple and often exploratory studies of only 20 years ago. That is not to criticise those earlier researchers (amongst them the present authors!) but rather it is a sign of progress in common with that seen in other life sciences methodologies. In that spirit, we argue that human neuroimaging will continue to deliver on its early promise as one tool in the toolbox for understanding human behaviour, providing neural constraints on cognitive theories[1].

Competing Interests

The authors declare no competing interests.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 725970). The funder had no role in the preparation of the manuscript or the decision to publish. The authors thank Floris de Lange, Roel Willems, and Pieter Medendorp for helpful comments on an earlier version of the manuscript.

References

1. Churchland, P. S. & Sejnowski, T. J. Perspectives on Cognitive Neuroscience. *Science* **242**, 741–745 (1988).
2. Coltheart, M. How Can Functional Neuroimaging Inform Cognitive Theories? *Perspect Psychol Sci* **8**, 98–103 (2013).
3. Downing, P., Liu, J. & Kanwisher, N. Testing cognitive models of visual attention with fMRI and MEG. *Neuropsychologia* **39**, 1329–1342 (2001).
4. Henson, R. What can Functional Neuroimaging Tell the Experimental Psychologist? *The Quarterly Journal of Experimental Psychology Section A* **58**, 193–233 (2005).
5. Mather, M., Cacioppo, J. T. & Kanwisher, N. How fMRI Can Inform Cognitive Theories. *Perspect Psychol Sci* **8**, 108–113 (2013).
6. Page, M. P. A. What Can't Functional Neuroimaging Tell the Cognitive Psychologist? *Cortex* **42**, 428–443 (2006).
7. Norman, K. A., Polyn, S. M., Detre, G. J. & Haxby, J. V. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences* **10**, 424–430 (2006).
8. Haynes, J.-D. A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron* **87**, 257–270 (2015).

9. Poldrack, R. A. & Farah, M. J. Progress and challenges in probing the human brain. *Nature* **526**, 371–379 (2015).
10. Rissman, J. & Wagner, A. D. Distributed Representations in Memory: Insights from Functional Brain Imaging. *Annu. Rev. Psychol.* **63**, 101–128 (2012).
11. Haxby, J. V., Connolly, A. C. & Guntupalli, J. S. Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annu. Rev. Neurosci.* **37**, 435–456 (2014).
12. Hebart, M. N. & Baker, C. I. Deconstructing multivariate decoding for the study of brain function. *NeuroImage* **180**, 4–18 (2018).
13. Grootswagers, T., Wardle, S. G. & Carlson, T. A. Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Journal of Cognitive Neuroscience* **29**, 677–697 (2017).
14. Carlson, T., Goddard, E., Kaplan, D. M., Klein, C. & Ritchie, J. B. Ghosts in machine learning for cognitive neuroscience: Moving from data to theory. *NeuroImage* **180**, 88–100 (2018).
15. de-Wit, L., Alexander, D., Ekroll, V. & Wagemans, J. Is neuroimaging measuring information in the brain? *Psychon Bull Rev* **23**, 1415–1428 (2016).
16. Poeppel, D. The maps problem and the mapping problem: Two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsychology* **29**, 34–55 (2012).
17. Lebedev, M. A. & Nicolelis, M. A. L. Brain-Machine Interfaces: From Basic Science to Neuroprostheses and Neurorehabilitation. *Physiological Reviews* **97**, 767–837 (2017).
18. Haxby, J. V. Multivariate pattern analysis of fMRI: The early beginnings. *NeuroImage* **62**, 852–855 (2012).
19. Haynes, J.-D. & Rees, G. Decoding mental states from brain activity in humans. *Nat Rev Neurosci* **7**, 523–534 (2006).
20. Tong, F. & Pratte, M. S. Decoding Patterns of Human Brain Activity. *Annu. Rev. Psychol.* **63**, 483–509 (2012).
21. Cohen, J. D. *et al.* Computational approaches to fMRI analysis. *Nat Neurosci* **20**, 304–313 (2017).
22. Kriegeskorte, N. & Douglas, P. K. Cognitive computational neuroscience. *Nat Neurosci* **21**, 1148–1160 (2018).
23. Guest, O. & Love, B. C. What the success of brain imaging implies about the neural code. *eLife* **6**, e21397 (2017).
24. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci* **19**, 356–365 (2016).
25. Ritchie, J. B., Kaplan, D. M. & Klein, C. Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. *The British Journal for the Philosophy of Science* **70**, 581–607 (2019).
26. Kragel, P. A., Koban, L., Barrett, L. F. & Wager, T. D. Representation, Pattern Information, and Brain Signatures: From Neurons to Neuroimaging. *Neuron* **99**, 257–273 (2018).
27. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *NeuroImage* **56**, 400–410 (2011).
28. Sprague, T. C. & Serences, J. T. Using human neuroimaging to examine top-down modulation of visual perception. in *An introduction to model-based cognitive neuroscience* 245–274 (Springer, New York, NY, 2015).

29. Kriegeskorte, N. & Douglas, P. K. Interpreting encoding and decoding models. *Current Opinion in Neurobiology* **55**, 167–179 (2019).
30. Op de Beeck, H. P. Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fMRI analyses? *NeuroImage* **49**, 1943–1948 (2010).
31. Swisher, J. D. *et al.* Multiscale Pattern Analysis of Orientation-Selective Activity in the Primary Visual Cortex. *Journal of Neuroscience* **30**, 325–330 (2010).
32. Freeman, J., Brouwer, G. J., Heeger, D. J. & Merriam, E. P. Orientation Decoding Depends on Maps, Not Columns. *Journal of Neuroscience* **31**, 4792–4804 (2011).
33. Alink, A., Krugliak, A., Walther, A. & Kriegeskorte, N. fMRI orientation decoding in V1 does not require global maps or globally coherent orientation stimuli. *Front. Psychol.* **4**, 1–14 (2013).
34. Carlson, T. A. Orientation Decoding in Human Visual Cortex: New Insights from an Unbiased Perspective. *Journal of Neuroscience* **34**, 8373–8383 (2014).
35. Roth, Z. N., Heeger, D. J. & Merriam, E. P. Stimulus vignetting and orientation selectivity in human visual cortex. *eLife* **7**, e37241 (2018).
36. Marr, D. *Vision: A Computational Approach*. (Freeman and Co., 1982).
37. Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. & Poeppel, D. Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron* **93**, 480–490 (2017).
38. Neisser, U. *Cognitive psychology*. (Appleton-Century-Crofts, 1967).
39. Anderson, J. R. *Cognitive psychology and its implications*. (Worth Publishers, 2020).
40. Garner, W. R. *The processing of information and structure*. (Psychology Press, 1974).
41. Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
42. Miyake, A. & Friedman, N. P. The Nature and Organization of Individual Differences in Executive Functions: Four General Conclusions. *Curr Dir Psychol Sci* **21**, 8–14 (2012).
43. Cohen, J. D., Barch, D. M., Carter, C. & Servan-Schreiber, D. Context-processing deficits in schizophrenia: converging evidence from three theoretically motivated cognitive tasks. *Journal of Abnormal Psychology* **108**, 120–133 (1999).
44. Markus, H. R. & Kitayama, S. Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review* **98**, 224–253 (1991).
45. Kriegeskorte, N. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Sys. Neurosci.* **2**, 1–28 (2008).
46. Mur, M., Bandettini, P. A. & Kriegeskorte, N. Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience* **4**, 101–109 (2009).
47. Cichy, R. M., Pantazis, D. & Oliva, A. Resolving human object recognition in space and time. *Nat Neurosci* **17**, 455–462 (2014).
48. Proklova, D., Kaiser, D. & Peelen, M. V. Disentangling Representations of Object Shape and Object Category in Human Visual Cortex: The Animate–Inanimate Distinction. *Journal of Cognitive Neuroscience* **28**, 680–692 (2016).
49. Cohen, M. A., Alvarez, G. A., Nakayama, K. & Konkle, T. Visual search for object categories is predicted by the representational architecture of high-level visual cortex. *Journal of Neurophysiology* **117**, 388–402 (2017).

50. Groen, I. I. *et al.* Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife* **7**, e32962 (2018).
51. Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci Rep* **6**, 27755 (2016).
52. Smith, P. L. & Little, D. R. Small is beautiful: In defense of the small-N design. *Psychon Bull Rev* **25**, 2083–2101 (2018).
53. Meehl, P. E. Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philos. of Sci.* **34**, 103–115 (1967).
54. Kanwisher, N. & Wojciulik, E. Visual attention: Insights from brain imaging. *Nat Rev Neurosci* **1**, 91–100 (2000).
55. Wagner, A. D. *et al.* Building Memories: Remembering and Forgetting of Verbal Experiences as Predicted by Brain Activity. *Science* **281**, 1188–1191 (1998).
56. Curtis, C. E. & D'Esposito, M. Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences* **7**, 415–423 (2003).
57. Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R. & Ungerleider, L. G. Increased Activity in Human Visual Cortex during Directed Attention in the Absence of Visual Stimulation. *Neuron* **22**, 751–761 (1999).
58. Jiang, J., Summerfield, C. & Egner, T. Attention Sharpens the Distinction between Expected and Unexpected Percepts in the Visual Brain. *Journal of Neuroscience* **33**, 18438–18447 (2013).
59. Seidl, K. N., Peelen, M. V. & Kastner, S. Neural Evidence for Distracter Suppression during Visual Search in Real-World Scenes. *Journal of Neuroscience* **32**, 11812–11819 (2012).
60. Serences, J. T., Ester, E. F., Vogel, E. K. & Awh, E. Stimulus-Specific Delay Activity in Human Primary Visual Cortex. *Psychol Sci* **20**, 207–214 (2009).
61. Harrison, S. A. & Tong, F. Decoding reveals the contents of visual working memory in early visual areas. *Nature* **458**, 632–635 (2009).
62. Polyn, S. M., Natu, V. S., Cohen, J. D. & Norman, K. A. Category-Specific Cortical Activity Precedes Retrieval During Memory Search. *Science* **310**, 1963–1966 (2005).
63. Kornysheva, K. *et al.* Neural Competitive Queuing of Ordinal Structure Underlies Skilled Sequential Action. *Neuron* **101**, 1166-1180.e3 (2019).
64. Ariani, G., Pruszynski, J. A. & Diedrichsen, J. Motor planning brings human primary somatosensory cortex into action-specific preparatory states. *eLife* **11**, e69517 (2022).
65. Oosterhof, N. N., Tipper, S. P. & Downing, P. E. Crossmodal and action-specific: neuroimaging the human mirror neuron system. *Trends in Cognitive Sciences* **17**, 311–318 (2013).
66. Kaplan, J. T., Man, K. & Greening, S. G. Multivariate cross-classification: applying machine learning techniques to characterize abstraction in neural representations. *Front. Hum. Neurosci.* **9**, (2015).
67. King, J.-R. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences* **18**, 203–210 (2014).
68. Haxby, J. V., Guntupalli, J. S., Nastase, S. A. & Feilong, M. Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife* **9**, e56601 (2020).
69. Tolman, E. C. Cognitive maps in rats and men. *Psychological Review* **55**, 189–208 (1948).

70. Whittington, J. C. R., McCaffary, D., Bakermans, J. J. W. & Behrens, T. E. J. How to build a cognitive map. *Nat Neurosci* **25**, 1257–1272 (2022).
71. Vass, L. K. & Epstein, R. A. Abstract Representations of Location and Facing Direction in the Human Brain. *Journal of Neuroscience* **33**, 6133–6142 (2013).
72. Nielson, D. M., Smith, T. A., Sreekumar, V., Dennis, S. & Sederberg, P. B. Human hippocampus represents space and time during retrieval of real-world memories. *Proc Natl Acad Sci USA* **112**, 11078–11083 (2015).
73. Biederman, I., Mezzanotte, R. J. & Rabinowitz, J. C. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology* **14**, 143–177 (1982).
74. Davenport, J. L. & Potter, M. C. Scene Consistency in Object and Background Perception. *Psychol Sci* **15**, 559–564 (2004).
75. Bar, M. Visual objects in context. *Nat Rev Neurosci* **5**, 617–629 (2004).
76. Henderson, J. M. & Hollingworth, A. High-level scene perception. *Annu. Rev. Psychol.* **50**, 243–271 (1999).
77. Brandman, T. & Peelen, M. V. Interaction between Scene and Object Processing Revealed by Human fMRI and MEG Decoding. *J. Neurosci.* **37**, 7700–7710 (2017).
78. Chun, M. M., Golomb, J. D. & Turk-Browne, N. B. A Taxonomy of External and Internal Attention. *Annu. Rev. Psychol.* **62**, 73–101 (2011).
79. Desimone, R. & Duncan, J. Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience* **18**, 193–222 (1995).
80. Duncan, J. & Humphreys, G. W. Visual Search and Stimulus Similarity. *Psychological Review* **96**, 433–458 (1989).
81. Chelazzi, L., Miller, E. K., Duncan, J. & Desimone, R. A neural basis for visual search in inferior temporal cortex. *Nature* **363**, 345–347 (1993).
82. Battistoni, E., Stein, T. & Peelen, M. V. Preparatory attention in visual cortex: Preparatory attention in visual cortex. *Ann. N.Y. Acad. Sci.* **1396**, 92–107 (2017).
83. Stokes, M., Thompson, R., Nobre, A. C. & Duncan, J. Shape-specific preparatory activity mediates attention to targets in human visual cortex. *Proceedings of the National Academy of Sciences* **106**, 19569–19574 (2009).
84. Peelen, M. V. & Kastner, S. A neural basis for real-world visual search in human occipitotemporal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 12125–12130 (2011).
85. Gayet, S. & Peelen, M. V. Preparatory attention incorporates contextual expectations. *Current Biology* **32**, 687-692.e6 (2022).
86. Tulving, E. & Thomson, D. M. Encoding specificity and retrieval processes in episodic memory. *Psychological Review* **80**, 352–373 (1973).
87. Jafarpour, A., Fuentemilla, L., Horner, A. J., Penny, W. & Duzel, E. Replay of Very Early Encoding Representations during Recollection. *Journal of Neuroscience* **34**, 242–248 (2014).
88. Mack, M. L., Preston, A. R. & Love, B. C. Decoding the Brain's Algorithm for Categorization from Its Neural Implementation. *Current Biology* **23**, 2023–2027 (2013).
89. Posner, M. I. & Keele, S. W. On the genesis of abstract ideas. *Journal of Experimental Psychology* **77**, 353–363 (1968).
90. Medin, D. L. & Schaffer, M. M. Context Theory of Classification Learning. *Psychological Review* **85**, 207–238 (1978).

91. Barsalou, L. W. Perceptual symbol systems. *Behav Brain Sci* **22**, 577–660 (1999).
92. Glenberg, A. M. Few believe the world is flat: How embodiment is changing the scientific understanding of cognition. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* **69**, 165–171 (2015).
93. Fernandino, L., Tong, J.-Q., Conant, L. L., Humphries, C. J. & Binder, J. R. Decoding the information structure underlying the neural representation of concepts. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2108091119 (2022).
94. Levenson, R. W. Basic Emotion Questions. *Emotion Review* **3**, 379–386 (2011).
95. Gendron, M. & Feldman Barrett, L. Reconstructing the Past: A Century of Ideas About Emotion in Psychology. *Emotion Review* **1**, 316–339 (2009).
96. Ekman, P. Universal and cultural differences in facial expressions of emotions. in *Nebraska symposium on motivation, 1971* 207–283 (University of Nebraska Press, 1972).
97. Saarimäki, H. *et al.* Discrete Neural Signatures of Basic Emotions. *Cereb. Cortex* **26**, 2563–2573 (2016).
98. Skerry, A. E. & Saxe, R. Neural Representations of Emotion Are Organized around Abstract Event Features. *Current Biology* **25**, 1945–1954 (2015).
99. Peelen, M. V., Atkinson, A. P. & Vuilleumier, P. Supramodal Representations of Perceived Emotions in the Human Brain. *Journal of Neuroscience* **30**, 10127–10134 (2010).
100. Ellsworth, P. C. Appraisal Theory: Old and New Questions. *Emotion Review* **5**, 125–131 (2013).
101. Scherer, K. R. The Nature and Dynamics of Relevance and Valence Appraisals: Theoretical Advances and Recent Evidence. *Emotion Review* **5**, 150–162 (2013).
102. Gallese, V. & Goldman, A. Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences* **2**, 493–501 (1998).
103. Saxe, R. Against simulation: the argument from error. *Trends in Cognitive Sciences* **9**, 174–179 (2005).
104. Corradi-Dell'Acqua, C., Tusche, A., Vuilleumier, P. & Singer, T. Cross-modal representations of first-hand and vicarious pain, disgust and fairness in insular and cingulate cortex. *Nat Commun* **7**, 10904 (2016).
105. Koster-Hale, J., Bedny, M. & Saxe, R. Thinking about seeing: Perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults. *Cognition* **133**, 65–78 (2014).
106. von Holst, E. & Mittelstaedt, H. The reafference principle: interaction between the central nervous system and the periphery. *Die Naturwissenschaften* **37**, 464–476 (1950).
107. Wolpert, D. M. & Flanagan, J. R. Motor prediction. *Current Biology* **11**, R729–R732 (2001).
108. Umeda, T., Isa, T. & Nishimura, Y. The somatosensory cortex receives information about motor output. *Sci. Adv.* **5**, 1–14 (2019).
109. Gale, D. J., Flanagan, J. R. & Gallivan, J. P. Human Somatosensory Cortex Is Modulated during Motor Planning. *J. Neurosci.* **41**, 5909–5922 (2021).
110. Terrace, H. S. The simultaneous chain: a new approach to serial learning. *Trends in Cognitive Sciences* **9**, 202–210 (2005).
111. Houghton, G. & Hartley, T. Parallel Models of Serial Behaviour: Lashley Revisited. **2**, 1–25 (1995).

112. Williams, M. A., Dang, S. & Kanwisher, N. G. Only some spatial patterns of fMRI response are read out in task performance. *Nat Neurosci* **10**, 685–686 (2007).
113. Ritchie, J. B. & Carlson, T. A. Neural Decoding and "Inner" Psychophysics: A Distance-to-Bound Approach for Linking Mind, Brain, and Behavior. *Front. Neurosci.* **10**, (2016).
114. Wischnewski, M. & Peelen, M. V. Causal neural mechanisms of context-based object recognition. *eLife* **10**, e69736 (2021).
115. Thielen, J., Bosch, S. E., van Leeuwen, T. M., van Gerven, M. A. J. & van Lier, R. Evidence for confounding eye movements under attempted fixation and active viewing in cognitive neuroscience. *Sci Rep* **9**, 17456 (2019).
116. Proklova, D., Kaiser, D. & Peelen, M. V. MEG sensor patterns reflect perceptual but not categorical similarity of animate and inanimate objects. *NeuroImage* **193**, 167–177 (2019).
117. Mole, C. & Klein, C. Confirmation, refutation and the evidence of fMRI. in *S.H. Hanson & M. Bunzl (Eds.), Foundational issues of human brain mapping* 99–112 (MIT Press, 2010).
118. Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D. & Kriegeskorte, N. Unique semantic space in the brain of each beholder predicts perceived similarity. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 14565–14570 (2014).
119. Anderson, A. J. *et al.* Decoding individual identity from brain activity elicited in imagining common experiences. *Nat Commun* **11**, 5916 (2020).
120. Feilong, M., Guntupalli, J. S. & Haxby, J. V. The neural basis of intelligence in fine-grained cortical topographies. *eLife* **10**, e64058 (2021).
121. Braunlich, K. & Love, B. C. Occipitotemporal representations reflect individual differences in conceptual knowledge. *Journal of Experimental Psychology: General* **148**, 1192–1203 (2019).
122. Anzellotti, S. & Coutanche, M. N. Beyond Functional Connectivity: Investigating Networks of Multivariate Representations. *Trends in Cognitive Sciences* **22**, 258–269 (2018).
123. Ju, H. & Bassett, D. S. Dynamic representations in networked neural systems. *Nat Neurosci* **23**, 908–917 (2020).
124. van Gerven, M. A. J., Seeliger, K., Güçlü, U. & Güçlütürk, Y. Current Advances in Neural Decoding. in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (eds. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Müller, K.-R.) vol. 11700 379–394 (Springer International Publishing, 2019).
125. Lawrence, S. J. D., Formisano, E., Muckli, L. & de Lange, F. P. Laminar fMRI: Applications for cognitive neuroscience. *NeuroImage* **197**, 785–791 (2019).
126. De Vos, M. & Debener, S. Mobile EEG: Towards brain activity monitoring during natural action and cognition. *International Journal of Psychophysiology* **91**, 1–2 (2014).
127. Snow, J. C. & Culham, J. C. The Treachery of Images: How Realism Influences Brain and Behavior. *Trends in Cognitive Sciences* **25**, 506–519 (2021).
128. Willems, R. M. & Peelen, M. V. How context changes the neural basis of perception and language. *iScience* **24**, 102392 (2021).
129. van Rooij, I. & Baggio, G. Theory Before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science. *Perspectives on Psychological Science* **16**, 682–697 (2021).
130. Guest, O. & Martin, A. E. How Computational Modeling Can Force Theory Building in Psychological Science. *Perspectives on Psychological Science* **16**, 789–802 (2021).

131. Muthukrishna, M. & Henrich, J. A problem in theory. *Nat Hum Behav* **3**, 221–229 (2019).
132. Walther, A. *et al.* Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage* **137**, 188–200 (2016).