1

What We Have Learned about Memory from Neuroimaging

Andrea Greve and Richard Henson

Introduction

Functional neuroimaging techniques, such as functional magnetic resonance imaging (fMRI) and electro/magnetoencephalography (EEG/MEG), have had a major impact on the study of human memory over the last two decades. This impact includes not only new evidence about the parts of the brain that are important for memory ("functional localization" or "brain mapping"), which extends what was previously known from patients with brain damage, but also arguably informs our theoretical understanding of how memory works (e.g., Henson, 2005; Poldrack, 2006; though such claims have been questioned, e.g., Coltheart, 2006; Uttal, 2001). In this chapter, we illustrate ways in which functional neuroimaging has influenced our understanding of memory, going beyond research that was previously based primarily on behavioral techniques. We focus in particular on how memory processes might be implemented in the brain in terms of average levels of activity in certain brain areas, patterns of activity within areas, and connectivity between brain areas.

Theoretical Concepts That are Difficult to Measure Behaviorally, e.g., Retrieval States

Tulving (1983) theorized that we adopt a particular mind-set during episodic memory retrieval, a so-called "retrieval mode," which optimizes recovery of information from memory, and allows us to interpret that information as having come from the past (rather than from sensations in the present). Until recently, however, it has been difficult to evaluate theories like this owing to the difficulty of measuring such states behaviorally. Neuroimaging, on the other hand, is able to measure sustained brain activity directly associated with a state. This ability has reinvigorated such theories, leading to new hypothetical states that are assumed to be important for the encoding and retrieval of information, and even prompting new behavioral measures to investigate such theories further (see Chapter 5).

The Wiley Handbook on the Cognitive Neuroscience of Memory, First Edition.
Edited by Donna Rose Addis, Morgan Barense, and Audrey Duarte.

An early example of this use of neuroimaging is the study of Düzel and colleagues (1999), who recorded EEG during sequences of four words. Prior to each sequence, a cue instructed participants to decide whether or not each word was seen in a previous study phase ("episodic task"), or whether each word denoted a living or nonliving entity ("semantic task"). Düzel *et al.* found a sustained positive shift over right frontal electrodes for the episodic task relative to the semantic task. This positive shift emerged shortly after the instruction onset, but prior to the presentation of the first word (i.e., before any retrieval had taken place), and so was interpreted as evidence of a preparatory state for episodic retrieval, i.e., a retrieval mode.

This neuroimaging finding in turn prompted new theoretical proposals. Rugg and Wilding (2000) proposed that there may be different states even within a retrieval mode, in which people are oriented towards retrieving different types of episodic information. They called these "retrieval orientations." For example, Herron and Wilding (2004) reported a more positive-going left frontocentral EEG shift when participants prepared to retrieve the type of encoding task under which an item was studied, compared to when they prepared to retrieve the location in which an item was studied. Another example is the study of Ranganath and Paller (1999), which examined event-related potentials (ERPs) locked to the onset of correctly rejected, new (unstudied) items in a recognition memory test. Because such correct rejections are unlikely to elicit any episodic retrieval, any difference in their associated ERPs as a function of retrieval instructions is likely to be a consequence of a different retrieval orientation. In this case, Ranganath and Paller compared a retrieval task in which participants had to endorse objects that had appeared at study, regardless of their size on the screen ("general task"), with another task in which participants were only to endorse items as studied if they appeared in the same size as at study ("specific task"). A more positive-going ERP waveform to correct rejections was found post-stimulus onset over left frontal electrodes for the specific than for the general task.

It is also possible to measure such state-related brain activity with fMRI, though given its worse temporal resolution relative to EEG, special designs are needed that allow statistical modeling to separate state-related from item-related blood-oxygen-level-dependent (BOLD) responses. For example, Donaldson and colleagues (2001) showed state-related activity associated with blocks of a recognition memory task (relative to blocks of a fixation task) in bilateral frontal opercular areas. Moreover, the same brain areas also showed greater item-related activity for correct recognition (hits) than correct rejections, suggesting that frontal operculum supports both a sustained retrieval mode and transient processes associated with successful retrieval. A subsequent fMRI study by Otten, Henson, and Rugg (2002) provided analogous evidence for dissociable "encoding orientations". These authors found that the mean level of state-related activity during blocks of words varied as a function of the number of words later remembered within each block, independent of item-related activity associated with whether or not individual words were successfully remembered. Furthermore, this relationship between state-related activity and subsequent memory occurred in different brain areas as a function of the study task: occurring in left prefrontal cortex when participants performed a semantic (deep) task, and superior medial parietal cortex when participants performed a phonemic (shallow) task.

Importantly, the neuroimaging studies described above have not only led to new theoretical development (e.g., the concepts of retrieval and encoding orientations), but also prompted new behavioral experiments to further test these concepts. Building on the ERP studies such as that of Ranganath and Paller (1999) described above, Jacoby *et al.* (2005) conducted behavioral investigations of retrieval orientation. They used a second memory test to probe the fate of correctly rejected new items (foils) in a first recognition test, as a function of the retrieval orientation that was adopted during that first memory test. Participants studied one list of items under a semantic (deep) task, and another list of items under a phonemic (shallow) task. In the first recognition test, participants were expected to be oriented towards semantic information when distinguishing foils from deeply encoded targets, but oriented towards phonemic information when distinguishing foils from shallowly encoded targets. If so, the foils in the semantic condition should be processed more deeply than the foils in the phonemic condition, and hence themselves be remembered better on the final recognition test. This is exactly what the authors found. Thus, this (indirect) behavioral assay supported the theories of retrieval orientations that originated from neuroimaging research. Furthermore, this assay has been used to examine how retrieval orientations become less precise as people get older.

Supplementing Behavioral Dissociations with Neuroimaging Dissociations, e.g., Dual-Process Theories

Another situation in which neuroimaging data can complement behavioral data arises when seeking functional dissociations between hypothetical memory processes. For example, there has been a long-standing debate about whether behavioral data from recognition memory tasks are best explained by single- versus dual-process models. Single-process models claim that a single memory-strength variable is sufficient to explain recognition performance, normally couched in terms of signal detection theory (Donaldson, 1996; Dunn, 2004, 2008; Wixted, 2007; Wixted and Mickes, 2010). Dual-process models, however, assume that recognition involves at least two different processes, such as recollection, associated with retrieval of contextual information, and familiarity, providing a generic sense of a previous encounter, but without contextual retrieval (Aggleton and Brown, 1999; Diana *et al.*, 2006; Rotello and Macmillan, 2006; Yonelinas, 2002; see also Chapter 9). It is not clear that behavioral data have yet resolved this debate (though the main protagonists may disagree!). One possible solution is to examine neuroimaging data from the same task: if conditions assumed to entail recollection produce qualitatively, rather than just quantitatively, different patterns of activity across the brain compared to conditions assumed to entail familiarity, then this would appear to support dual-process models (see Henson, 2005, 2006, for further elaboration and assumptions of this type of "forward inference").

A methodological question then becomes how to define a "qualitatively" different pattern of brain activity. With classical statistics, it is not sufficient, for example, to find a significant difference in one brain area for a contrast of a recollection-condition against a baseline condition, and in a different brain area for the contrast of a familiarity-condition with that baseline. This is simply because the failure to find significant activation for each condition in the other brain area could be a null result.

However, even finding a significant interaction between two brain areas and two such contrasts is not sufficient, because we do not know the "neurometric" mapping between fMRI/EEG/MEG signal and the hypothetical processes of interest. This mapping may not be linear (i.e., a doubling in memory strength may not necessarily mean a doubling in BOLD signal or ERP amplitude). Moreover, the neurometric mapping may differ across different brain areas. Indeed, there may be a positive relationship between the neuroimaging signal and a memory process in one area (e.g., increasing BOLD signal associated with increasing memory strength in hippocampus), but a negative relationship between the neuroimaging signal and the same memory process in another area (e.g., decreasing BOLD signal associated with increasing memory strength in perirhinal cortex; Henson, 2006; Squire, Wixted, and Clark, 2007). These considerations mean that even a significant crossover interaction between two areas and two conditions does not refute single-process theories.

Fortunately, there is a method to solve this problem of unknown neurometric mappings, which assumes only that these mappings are monotonic (in other words, the neuroimaging signal must always increase, or always decrease, whenever engagement of the hypothetical process increases, even if it does not increase or decrease in equal steps). This method is called "state-trace analysis," and it was developed in the psychological literature by Bamber (1979). The "reversed association" pattern described by Dunn and Kirsner (1988), and by Henson (2005), is a special case of state-trace analysis. This method requires at least two dependent variables, e.g., neuroimaging signal in two brain areas, and at least three levels of the independent variable, e.g., three memory conditions. When plotting the data from each condition in a space whose axes are defined by the two independent variables, if the resulting "state-trace" is neither monotonically increasing nor monotonically decreasing, then one can refute the hypothesis that there is a single underlying process (for further elaboration, see Newell and Dunn 2008).

This analysis has been recently applied to neuroimaging data, for the first time, by Staresina *et al.* (2013b). These authors examined the amplitude of the initial evoked component (peaking around 400 ms) in ERPs recorded directly from human hippocampus and perirhinal cortex during a recognition memory task. The task enabled definition of three trial types: (1) trials in which an unstudied item was correctly rejected, (2) trials in which a studied item was recognized but its study context was not identified, and (3) trials in which a studied item was recognized and its study context was identified. According to single-process models, conditions 1–3 should be ordered along an increasing continuum of memory strength. However, Staresina *et al.* were able to reject this hypothesis by demonstrating a non-monotonic state-trace, concluding that at least two different processes were occurring in these two brain areas.

While this finding overturns previous claims that a single dimension of memory strength can explain neuroimaging data in the medial temporal lobe during recognition memory tasks (Squire, Wixted, and Clark, 2007; Wixted, 2007), it is important to note that it does not necessarily support specific dual-process memory theories. State-trace analysis only imputes the dimensionality of the underlying causes (assuming a monotonic mapping from those causes to each measurement); it does not constrain what those dimensions are. Thus further theorizing, concerning the precise nature of the experimental conditions, is necessary to infer the nature of the two or more processes that differed across the three conditions in the study by Staresina *et al.*

(2013b). For example, one process may have related to memory strength, while the other could have reflected differences in some other non-mnemonic process that happened to also differ across the three conditions. Note also that, even if there are multiple memory signals in the brain, they may still be mapped onto a single dimension of "evidence of oldness" in order to make a typical old/new recognition decision, i.e., conform to single-process theory in terms of behavioral data.

The use of state-trace analysis for "forward inference" of course resembles the classical "dissociation logic" commonly used in cognitive psychology and neuropsychology (Henson, 2005; Shallice, 2003). In the extreme case, such inferences do not care where in the brain (or when in time) qualitative differences in brain activity are found (cf. "reverse inference," considered in the next section). Indeed, even when brain location may be of interest – such as hippocampus versus perirhinal cortex in the above example of Staresina *et al.* (2013b) – there are limitations to the specificity of such localization. As argued by Henson (2011), for example, as soon as one allows for nonlinear and recurrent transformations of a stimulus (experimental input) by other brain areas, the finding of a non-monotonic state-trace across two measured areas does not necessitate that the processes of interest occur in those areas: the dissociable neuroimaging signals in those areas might instead be due to differing inputs from other (non-measured) areas.

Inferring Memory Processes Directly from Local Brain Activity (Reverse Inference)

In contrast to the dissociation logic above, one of the most common types of psychological inference from neuroimaging data is based on association: namely, that a memory process occured within an experimental condition because a certain brain area was active. The assumptions and limitations of this type of "reverse inference" have been discussed at length (Poldrack, 2006, 2008). In the extreme case, this inference is only valid under a strict form of functional localization: i.e., when there exists a one-to-one mapping between a specific brain area and a specific cognitive process (Henson, 2005). We return to these limitations later, but first give some examples of this type of inference.

One example of a recent MEG study to use reverse inference was reported by Evans and Wilding (2012). This study tested a particular type of the dual-process theories of recognition memory described above: the independent-dual-process model of Yonelinas and colleagues (Diana *et al.*, 2006; Yonelinas, 2002). According to this model, recollection is a probabilistic event whose occurrence is independent of familiarity. This independence assumption has been questioned by others, however (Berry *et al.*, 2012; Pratte and Rouder, 2012; Wixted and Mickes, 2010), and is difficult to test with behavioral data alone, since the independence assumption is normally necessary in order to score the data.

Evans and Wilding (2012) combined MEG with Tulving's (1985) remember/know procedure, which instructs participants to make a *remember* (R) judgment when they can retrieve any contextual information associated with prior study of an item, a *know* (K) judgment if the item seems familiar to them, but they cannot remember any context, or a *new* (N) judgment if the item does not seem familiar. The basis of Evans and Wilding's reverse inference was an extensive EEG literature in which familiarity is believed to occur from 300 to 500 ms post-stimulus, while recollection is believed to occur later, from 500 to 800 ms (Bridson *et al.*, 2009; Donaldson, Wilding, and Allan, 2003; Greve, van Rossum, and Donaldson, 2007; Mecklinger, 2000; Rugg and Curran, 2007; Tendolkar *et al.*, 2000). They therefore measured the amplitude of the event-related fields (ERFs) in these two time-windows for R, K, and N judgments to studied items (i.e., R hits, K hits, and N misses).

According to Yonelinas's model (and in common with signal-detection theories), for a K judgment to be given, the strength of a familiarity signal needs to exceed some criterion (otherwise an N judgment is given instead). This means that, if R judgments are given only when recollection occurs, and the probability of this recollection is independent of the level of familiarity, then the mean level of familiarity for R judgments will be less than that for K judgments (since the occurrence of recollection means that familiarity does not also need to exceed some criterion in order to make an R judgment). Single-process theories, on the other hand, which assume R and K judgments are quantitatively rather than qualitatively different, always predict that memory strength will be highest for R judgments. Thus the rank order of the ERF from 300 to 500 ms should be N–R–K according to the independent dual-process model, but N–K–R according to single-process theories. Evans and Wilding (2012) found support for the first pattern, with ERF amplitude between 300 and 500 ms for R judgments falling in between that for N and K judgments. For the later time-window of 500–800 ms, on the other hand, the order was N = K < R, consistent with a separate, later recollection effect. This finding therefore supports dual-process models in which recollection and familiarity are independent.

Another recent example of a reverse inference in the context of dual-process models of recognition memory comes from the fMRI study of Taylor, Buratto, and Henson (2013). This study combined R/K judgments with brief, masked primes that occurred immediately prior to each item during a recognition memory test. These primes were masked so effectively that participants were rarely able to identify them. Under such conditions, Jacoby and Whitehouse (1989) found that participants are more likely to endorse test items (targets) as previously studied when the preceding prime was the same item (primed condition), relative to when the preceding prime was a different item (unprimed condition). This memory illusion occurs even for new test items that are not in fact studied, and subsequent studies showed that this increased bias to respond "old" is associated with K judgments, not R judgments (Kinoshita, 1997; Rajaram, 1993). This bias is naturally explained within a dual-process framework by assuming that matching primes increase the familiarity of test items, and this increased familiarity is attributed to the study phase (erroneously in the case of new items).

Taylor, Buratto, and Henson (2013) compared the effects of masked "repetition" primes, of the type discussed above, with the effects of masked "conceptual" primes, which were different but semantically related to the target item (though not associatively related; cf. Rajaram and Geraci, 2000). These conceptual primes increased R but not K judgments, thus showing the opposite effect to repetition primes. This finding is difficult to explain along the conventional dual-process lines described above, i.e, in terms of increased fluency being attributed to familiarity (though see Taylor, Buratto, and Henson 2013, for some suggestions). However, one trivial explanation is that the crossover interaction between repetition versus conceptual primes and R versus K judgments was an artefact of the mutually exclusive nature of the R/K procedure. That is, if the repetition and conceptual primes produced different types of fluency (perceptual versus semantic, for example), participants might feel obliged to indicate this by using K judgments for one type of fluency and R judgments for the other. Indeed, this mutually exclusive responding has been claimed to be a weakness of the standard R/K procedure; when participants are asked to give continuous and parallel ratings of both "remembering" and "knowing" for each item, many experimental manipulations are found to affect both R and K ratings (see Brown and Bodner, 2011; Kurilla and Westerman, 2008).

Taylor, Buratto, and Henson (2013) therefore combined their masked priming paradigm with fMRI, and leveraged on previous fMRI studies that have implicated inferior parietal activation in recollection. The authors replicated the increased BOLD signal in these parietal areas for R versus K judgments, but importantly also found that masked conceptual primes, but not masked repetition primes, increased this parietal activation further (relative to the unprimed case). This observation suggests that the conceptual primes were genuinely increasing recollection. This is therefore an example of where a reverse inference from neuroimaging data can be used to rule out an alternative theoretical account: here, that the interaction between R/K judgments and repetition/conceptual primes was a methodological artefact of the mutually exclusive R/K procedure.

Assuming the reverse inferences used by Evans and Wilding (2012) and Taylor, Buratto, and Henson (2013) are valid, both of these neuroimaging studies not only provide additional constraints on theories of recognition memory; they also offer methodological guidance for analysis of behavioral data, such as whether R and K judgments can be assumed to be independent (rather than redundant or exclusive; Knowlton and Squire, 1995; Mayes, Montaldi, and Migo, 2007). However, as mentioned earlier, the assumption of reverse inference, in its most extreme form, requires that the 300–500 ms ERF amplitude (in the Evans and Wilding example) reflects differences in, and only in, familiarity, and that the inferior parietal BOLD amplitude (in the Taylor and colleagues example) reflects differences in, and only in, recollection. If instead the 300–500 ms ERF or parietal BOLD amplitude reflect differences between R, K, and N categories other than their mean familiarity or recollection respectively (e.g., differences in some confounding variable), then the theoretical (reverse) inferences do not follow. For example, electrophysiological signals from 300–500 ms in recognition tasks have been argued not to reflect familiarity per se, but rather forms of implicit conceptual fluency (Paller, Voss, and Boehm, 2007; see also Chapter 3). Likewise, the BOLD signal in parietal cortex might not reflect recollection per se, but rather differences in endogenous or exogenous attention, or perhaps even differences related to motor preparation (given that "old" decisions associated with R judgments tend be made faster on average).

The nature of the mapping between brain measure and cognitive process is of course at the heart of cognitive neuroscience. The extreme form of functional localization assumes that each distinct brain area supports one unique hypothetical function (Figure 1.1a). To avoid making this one-to-one mapping between neuroimaging measure and cognitive process (which may not be provable in the strict sense: Henson, 2005), Poldrack (2006) suggested reverse inferences as probabilistic, according to a Bayesian framework. According to the Bayes' theorem, the probability that a cognitive function F1 was engaged when activity in a certain brain area A1 is observed depends on how likely it is that this brain area is active when function F1 is known to have

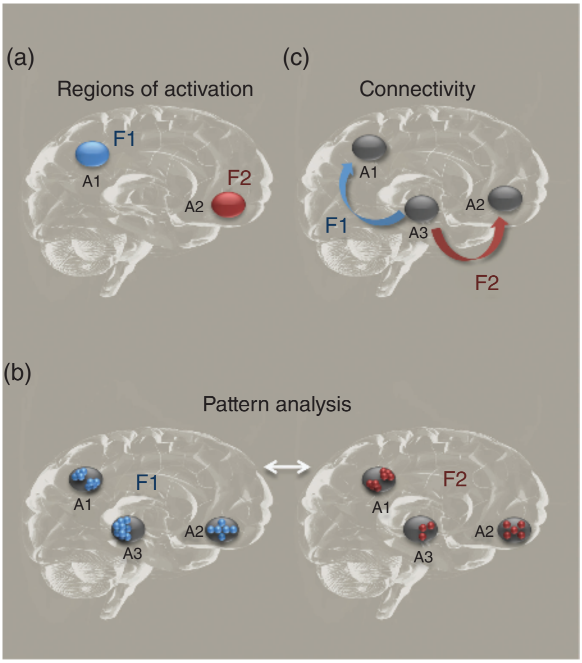

Figure 1.1 Schematic drawings of the human brain that illustrate different potential mappings of distinct memory functions (F1, F2) onto neural activity within and across distinct brain areas (A1, A2, A3). Changes in cognitive function can give rise to modulations in: (a) average levels of activity in different brain areas, (b) the pattern of activity within and across different regions, and (c) the nature of connectivity between multiple brain areas.

occurred, multiplied by the prior probability that function F1 generally occurs, and divided by the baseline probability that brain area A1 is generally active. While the likelihood of A1 being activated, whether or not F1 is assumed to have occurred, can be estimated from databases or meta-analyses, estimating the prior probability of function F1 occurring is problematic (though see Poldrack, 2006, for a possible solution).

In general terms, the implication of this Bayesian formulation is that, even if activity in a certain brain area is very likely to occur with a specific function – for example, a cognitive process reliably activates that area – this is not particularly informative if the same area is also activated in many other situations where that function is not involved. This has led many to criticize the weakness of reverse inferences. More recently, Hutzler (2014) argued that, if one further conditionalizes the probability of a brain area being activated on a subset of tasks (e.g., just those experiments that examined activity during a recognition memory task), then reverse inferences become stronger. In other words, if a brain area has consistently been activated in association with a specific memory process *in the context of recognition memory tasks* (ignoring how often it is activated in other types of tasks), then its activation in a new recognition memory experiment can provide strong evidence that this process has occurred. Thus, if the 300–500 ms ERF and parietal BOLD effects in the Evans and Wilding (2012) and Taylor, Buratto, and Henson (2013) studies have been consistently associated with familiarity and recollection respectively by prior neuroimaging experiments of recognition memory (regardless of whether they occur in other contexts), then this would bolster the reverse inferences described above.

The problem with Hutzler's (2014) argument is that it requires a definition of the subset of tasks over which to estimate the prior probability of activation (e.g., recognition memory tasks just with visual stimuli, or with any type of stimulus?). This debate then returns to the persistent question of cognitive theory, that is, the ontology of basic cognitive processes and their engagement in specific tasks. If this ontology can be established on purely independent grounds (e.g., from behavioral data alone), then reverse inference might become valid, but ironically neuroimaging is then no longer necessary for informing the ontology. An alternative pragmatic approach, suggested by Henson (2005), is that reverse inferences may start as being weak, but can still be used to inform and/or revise the cognitive ontology, leading to new experiments and iterated inferences until there is (hopefully) a convergence of brain mapping and cognitive ontology, such that a one-to-one mapping between brain area and cognitive process is established; at which point, reverse inference then becomes valid (see also Gonsalves and Cohen, 2010; Poldrack and Wagner, 2004).

Anatomical and Functional Scale, High-Resolution fMRI, and Contact with Animal Models

The above discussion raises the important issue of granularity (Henson, 2005): that is, at what level of specificity to define a cognitive process and at what spatial scale to define a "brain area." In terms of memory processes, for example, it is possible that recollection is not a unitary construct, in that retrieval of spatial context might be a dissociable function from retrieval of temporal context (e.g., Duarte *et al.*, 2010) and likewise, familiarity might encompass fluency of multiple different types of processing, e.g., orthographic, phonological, semantic, etc. In terms of brain areas, on the other hand, it is possible that trying to ascribe a single function to the hippocampus is inappropriate because it in fact contains several distinct subfields that each serve a different function, e.g., dentate gyrus (DG), CA1, CA3, and subiculum (Deguchi *et al.*, 2011; Lee *et al.*, 2004; Leutgeb *et al.*, 2004; Schmidt, Marrone, and Markus, 2012; Vazdarjanova and Guzowski, 2004; see also Chapter 6). In this case, averaging activity over all voxels within the hippocampus will obscure such functional differences. Analogously, had Evans and Wilding (2012) averaged over all time samples between 300 ms and 800 ms in their MEG study, then no difference between familiarity and recollection might have been observed.

It is possible that the appropriate level of anatomical granularity will only be found when the spatial resolution of neuroimaging techniques such as fMRI is increased. Indeed, in the extreme, we would like to be able to measure activity in individual neurons (or even individual synapses). This is of course possible in animals, but rarely in humans. Nonetheless, there are many computational models of the hippocampus (and other brain areas) that are based on such single-cell data from animals (Hasselmo and Howard, 2005; Lisman and Otmakhova, 2001; Treves and Rolls, 1994), some of which have hypothesized specialized functions for hippocampal subfields. The advent of high-resolution fMRI means that some of these subfields can now be imaged in humans, which in turn allows a bridge between human and animal data and models. For example, two concepts popularized in computational (neural network) models of the hippocampus are *pattern separation* and *pattern completion* (for more discussion, see Chapter 6). Pattern separation refers to the ability to orthogonalize similar input patterns (e.g., to separate two episodes that occurred in similar contexts), whereas pattern completion refers to the ability to group together different input patterns (e.g., to complete the details of an episodic memory given only a partial cue).

Several recent models attribute pattern separation to the DG. Inputs from cortical areas are assumed to reflect distributed patterns of activity, which are transformed into unique hippocampal representations via the DG and its subsequent sparse projections to the CA3 field. The recurrent connectivity within CA3, on the other hand, is thought to support pattern completion, via conjunctive representations of co-occurring elements. When a noisy or partial cue is presented, these conjunctive codes and recurrent connections enable completion of associated information (which is then projected back into the cortex via other subfields such as CA1 and subiculum). Most fMRI studies to date (which typically have a resolution of 3 mm isotropic) have been unable to resolve these hippocampal subfields, and so it has been difficult to test theories about pattern separation and completion, given that these processes co-occur.

Bakker *et al.* (2008), however, used the higher resolution (1.5 mm isotropic) afforded by recent advances in fMRI to separate BOLD signal across hippocampal subfields. They presented participants with a series of images, in which some images were either the same as previous images in the series, or were similar but not identical. If participants noticed this slight change, the DG showed a novelty response that was also observed for new items, but was absent when exact replicas of previously studied images were shown (though it was not possible to distinguish DG and CA3 even at this resolution). Bakker *et al.* interpreted this pattern as supporting a role of human DG in pattern separation. Neural populations in CA1 and subiculum areas, on the other hand, did not show a novelty response for the similar items and did not differentiate between the similar and identical items, and were interpreted as contributing to pattern completion (see also Johnson, Muftuler, and Rugg, 2008).

Clearly today's high-resolution fMRI is unlikely to be sufficient to reveal all functional subdivisions in our brains, and this may remain the case even if we reach theoretical limits on fMRI resolution, for example, in terms of vascular coverage. Nonetheless, the finer level of spatial granularity offered by higher-resolution fMRI is still likely to furnish insights beyond those afforded by our current resolutions, and thereby further reduce the gap between human and animal models. High-resolution fMRI is also likely to increase the amount of information extracted by multivariate pattern analyses, as discussed next.

Multivariate Pattern Analysis: Processes Versus Representations?

Multivariate pattern analysis (MVPA) is a relatively recent method that uses powerful pattern classification algorithms to determine whether different types of stimuli or cognitive processes can be classified on the basis of patterns of activity over voxels (in fMRI, e.g., Haxby *et al.,* 2001; Norman *et al.,* 2006; Polyn *et al.,* 2005), or over sensors/time-points/frequencies (in MEG/EEG, e.g., Jafarpour *et al.*, 2013). Thus in an fMRI experiment, for example, rather than comparing two conditions in terms of the average BOLD signal across voxels within a region of interest (ROI), the traditional "univariate" approach, MVPA compares them in terms of patterns of signals across voxels within that ROI, which may not necessarily differ in the average signal (Figure 1.1b). These methods have been shown to offer the remarkable ability to "decode" brain activity, for example, to determine what stimulus a person is looking at from the brain activity alone (see Norman *et al.*, 2006, for review; for more discussion of MVPA, see Chapters 2 and 6).

Some caution should be exercised concerning the recent excitement around MVPA, however, which again has to do with the questions of granularity and functional–anatomical mapping. If one were to include all voxels within the brain, then it would not be particularly surprising if MVPA could distinguish two stimuli that were perceivably different. Analyses of this kind are more theoretically interesting when participants have no reportable access to the processes of interest, or when patterns are restricted to various ROIs: for example, to discover that episodic memories can be classified above chance within one ROI, e.g., hippocampus (Chadwick *et al.*, 2010), but not within another, e.g., cerebellum. Yet it should be noted that this latter use of MVPA to "decode" brain activity within ROIs describes another form of functional localization, albeit one that may be more sensitive than traditional analyses that only consider the mean activity within an ROI[1]. Indeed, this use of MVPA is analogous to the issue of spatial resolution discussed in the previous section: a standard-resolution voxel can be viewed as an ROI that averages over what might be quite distinct patterns of activity had a higher resolution been used (i.e., one scanner's voxel is another scanner's ROI!).

Nonetheless, there has been a more important shift in perspective triggered by MVPA, in terms of characterizing the nature of neural representations. One example of this is the development of methods to test whether neural representations are sparse or distributed (e.g., Morcom and Friston, 2012). Another example, which is likely to have a significant effect on the field, is representational similarity analysis (RSA), in which the activity patterns for a large number of different stimuli are compared in terms of their similarity (see Chapter 6). The emergence of structure within the resulting stimulus-by-stimulus "similarity matrix" then gives clues to what an ROI is representing (e.g., animate versus inanimate visual objects; Kriegeskorte *et al.*, 2008). Thus the focus is not so much on whether or not patterns can be classified according to two or more experimentally defined categories, but on letting the data reveal the nature of the categories represented by an ROI (its "representational geometry"). Moreover, the similarity spaces observed in neuroimaging data can then be compared to those predicted by competing computational models (by applying RSA to model outputs, when the models are "presented with" the same stimuli). This approach offers an interesting potential way to test the computational models of the MTL described in the previous section (e.g., in terms of pattern separation and completion).

Moreover, the greater sensitivity of MVPA classification methods over traditional univariate methods should not be dismissed, because it has allowed researchers to track the presence of neural activity patterns (representations) over time in continuous – and hence noisy – brain activity. This has been particularly influential in memory research, where reactivation of memories can be examined by training a classifier on stimuli presented during the study phase, and then testing that classifier's ability to detect the same patterns during retrieval (when the stimuli are no longer present, e.g., cued by a different stimulus). One of the first examples to use this approach was the fMRI study of Polyn *et al.* (2005). These authors wanted to test the contextual reinstatement hypothesis (Bartlett, 1932; Tulving and Thomson, 1973), which states that people retrieve specific episodic details by first activating information about the general properties of such episodes. Polyn *et al.* did this by asking participants to study famous faces, famous locations, and common objects. An MVPA classifier was trained to distinguish these three categories from fMRI data acquired during the study phase. Then, using the fMRI data acquired when participants later freely recalled the names of the studied stimuli, the classifier predicted the category that participants were thinking about, on a moment-by-moment basis. Consistent with the contextual reinstatement hypothesis, high classification about a given category emerged several seconds before specific examples of that category were recalled.

MVPA has also been used in MEG, at least in the context of maintenance in short-term memory. Fuentemilla *et al.* (2010) trained an MVPA classifier to distinguish indoor or outdoor scenes, and then looked for above-chance classification (across sensors) at various times and frequencies during a 5-second retention interval. Interestingly, reactivations of above-chance classification were common in the theta frequency range (around 6 Hz) and correlated with memory performance, although only for blocks in which configural information needed to be retained during that interval. The authors argued that these data support animal models in which theta-coupled replay supports maintenance of information in working memory. Evidence for reactivation during a longer-term retention interval has also recently been found with fMRI. Staresina and colleagues (2013a) tracked the fMRI activity patterns occurring during a retention interval in which participants performed an odd/even distractor task, comparing their similarity to patterns evoked by individual stimuli during the study phase. Greater similarity was found for stimuli that were recalled in the subsequent test phase than for stimuli that were not, which supports the hypothesis that long-term memories are retained and/or consolidated by offline reactivation.

These examples thus illustrate a more subtle effect of the advent of MVPA, namely the theoretical shift in interpreting neuroimaging data in terms of processes versus representations. Results from univariate tests within an ROI are normally interpreted in terms of processes, i.e, the degree to which recollection or familiarity occurred, whereas MVPA results are normally interpreted in terms of representations. In reality, of course, it is impossible to define processes in the absence of representations (and vice versa), and defining both is often only possible in the context of formal models. Greve, Donaldson, and van Rossum (2010), for instance, described a neural network model that simulates two kinds of retrieval processes that operate on the same memory representation. This model simulated both familiarity-based and recollection-based discrimination of old and new items, which paralleled the characteristics reported in the empirical literature. Simulations like this demonstrate how the psychological processes of recollection and familiarity may reflect qualitatively distinct retrieval (read-out) operations that act on the same representations within a single brain area. More generally, explicit neural models like those described by Greve and colleagues, coupled with a subtle shift in perspective between characterizing processes and characterizing representations, may alter the way neuroimaging data are used to inform memory theories.

Functional and Effective Connectivity in Memory, e.g., within MTL

A further logical possibility is that some memory processes/representations are most visible in changes in the connectivity between brain areas, rather than in average activity or activity patterns within each area (Figure 1.1c). Given that memories are likely to be stored in terms of changes in synaptic strengths, and that those occur between as well as within brain areas, it would seem likely that those synaptic changes would alter the functional connectivity between areas. Recollection, for example, might correspond not simply to high activity levels within hippocampus, but rather to high levels of connectivity between hippocampus and other cortical areas, which represent the content of recollected memories (see also Chapter 13). Indeed, it is also possible that the same set of brain areas could enable different memory functions depending on changes in the effective connectivity between them; that is, the same anatomical network could "re-wire" into different functional networks according to different memory processes.

Some of the first fMRI studies to investigate memory-related changes in functional connectivity were performed by Maguire, Mummery, and Büchel (2000). These authors used structural equation modeling (SEM), a technique that tests competing models against each other, to evaluate explicit network models defined over a small number of ROIs. Assuming a model provides a satisfactory fit to the time-series data in each ROI, SEM coefficients for individual connections can then be interpreted in terms of "effective connectivity" between ROIs. Effective connectivity in this context goes beyond functional connectivity (e.g., in Figure 1.1a, simple pairwise correlation between activity in two areas A1 and A2) in that it allows for indirect connections (e.g., in Figure 1.1c, testing whether the correlation between A1 and A2 is actually due solely to a common input from a third area A3, assuming that all the areas that modulate activity within the network have been included in the model). Maguire *et al.* used SEM to address a theoretical debate about the distinction between semantic and episodic memory. The multiple-memory systems view (Tulving 1987) holds that separate memory systems are specialized for processing episodic and semantic information, supported by functionally independent networks. The alternative unitary system view proposes a single declarative memory system (McIntosh, 1999; Rajah and McIntosh, 2005; Roediger, 1984), in which memories can vary along a contextual continuum. Maguire and colleagues tested these theories by acquiring fMRI data while participants judged the accuracy of sentences about four different types of information: autobiographical events, public events, autobiographical facts, and general knowledge. They then defined a memory retrieval network by comparing activity common to all four of these types of sentence against a scrambled sentence baseline condition. This network included medial frontal cortex, left temporal pole, left hippocampus, left anterolateral middle temporal gyrus, parahippocampal cortex, posterior cingulate, retrosplenial cortex, and temporoparietal junction.

SEM then revealed several differences in effective connectivity between areas within this retrieval network as a function of the type of information retrieved. For example, connectivity from temporal pole to parahippocampal gyrus increased during retrieval of autobiographical relative to public events. Connectivity from temporal pole to lateral temporal cortex, on the other hand, increased during retrieval of public relative to autobiographical events. The authors argued that this pattern of results is more consistent with the view that episodic and semantic memories originate from separate systems that differ in the way information is processed, than with the view that semantic and episodic memories emerge from a continuum of representations that differ in contextual detail. Furthermore, the data suggest that brain areas can have multiple functions during memory retrieval, depending on their connectivity with other brain areas.

Gagnepain *et al.* (2010) provided another example of the different perspectives offered by local activity versus effective connectivity. These authors used dynamic causal modeling (DCM) of fMRI data, which can be thought of as an extension of SEM that includes a more sophisticated model of the dynamics of neural interactions and their expression via the haemodynamic (BOLD) response. DCM was applied to fMRI data from a study phase in which participants performed an incidental task on auditory words, and memory was tested 24 hours later using a remember/know procedure. Of primary interest was how neural activity that predicted subsequent R versus K judgments varied as a function of whether or not words at study had been primed via prestudy exposure. Unprimed words showed the usual pattern of greater hippocampal activity for words later attracting R judgments than for words later receiving K judgments. For primed words, however, this pattern was reversed, with decreased activity for words that attracted R than K judgments. This suggests that local hippocampal activity alone is not sufficient to predict subsequent memory. Instead, DCM analysis showed that subsequent R judgments were associated with increased effective connectivity to the hippocampus from the superior temporal gyrus – an area that showed the usual reduction in activity for primed relative to unprimed words. This was explained in terms of priming improving the transmission of sensory information to hippocampus, resulting in stronger associations between that information and its spatiotemporal context. Regardless of whether this explanation is correct, the more important issue for present purposes is that some causes of successful memory encoding may be found in the functional coupling between areas, rather than in local activity within those areas.

Given that much communication between brain areas during memory encoding and retrieval is likely to occur on the scale of tenths of a second, methods for testing effective connectivity are likely to be more theoretically illuminating when applied to MEG/EEG data than fMRI data, because changes in connectivity over such rapid timescales will be invisible to fMRI. Intracranial EEG data acquired directly from the medial temporal lobes of patients about to undergo surgery, for example, have shown transient increases in coupling between hippocampus and perirhinal cortex in the gamma frequency band (around 40 Hz) associated with successful memory encoding (Fell *et al.*, 2001). Recent methods that use DCM to compare different network models of extracranial MEG and EEG data may also prove a useful approach when intracranial data are not available (Kiebel *et al.*, 2008).

Closing the Loop: Inferring Causality from Neuroimaging Data

It is often stated that neuroimaging data are only correlational, and therefore brain activity may be incidental to a memory process of interest, rather than causing that process. This is sometimes then taken to mean that neuroimaging data are somehow inferior to behavioral data. The latter claim, however, would be mistaken, since both measures of brain activity and measures of behavior (for example, accuracy or speed) are measurements of the same neural/cognitive system. Indeed, the behavioral responses only reflect the final output, with less information about the intermediate stages between stimulus and response. In most cognitive neuroscientific (hypothetical-deductive) frameworks, neither type of measurement can directly "cause" a cognitive process; this would only make sense if one measurement were used as a surrogate for a process of interest, according to some theory (for further discussion of this issue, see Henson, 2005). Thus, claims that neuroimaging differences are confounded by concurrent behavioral differences are usually invalid: behavioral differences cannot cause activity differences; rather, brain activity and behavioral responses are normally both considered as the consequence of some hypothetical process. In the context of more mechanistic models of information flow, sensory input can be said to cause activity in one brain area, which can then be said to cause activity in another area, ultimately causing motor output (i.e., a behavioral response).

Of course, what is normally meant by the statement that neuroimaging data are only correlational is that they cannot tell us about the causal role of a brain area in a cognitive process in the same way that lesion data do. This issue would appear to be undeniable, and of course raises the question about how to define causality (Henson, 2005; Weber and Thompson-Schill, 2010). Without getting into philosophical debate, one recent step towards inferring causality from neuroimaging data was made by Yoo *et al.* (2012). Normally, a stimulus or task is manipulated experimentally, and brain and behavioral data are measured in response. Yoo *et al.*, on the other hand, used brain data to control when a stimulus was presented, and measured the consequence for subsequent behavior (i.e., the brain data were used to define the independent variable, rather than being the dependent variable). More precisely, they used real-time fMRI to measure online activity in the parahippocampal place area (PPA), and then presented visual scenes when PPA activity corresponded to either a "good" or "bad" state, where those states were defined by a prior experiment in which PPA activity was related to subsequent memory for scenes. Later testing outside the scanner then showed that recognition memory for the scenes presented during the "good" brain state was superior to that for scenes presented during the "bad" state. This finding thus bolsters the claim for a causal role in PPA activity during memory encoding. This approach still does not correspond to experimental manipulations that directly affect neural activity in a brain area (e.g., transcranial magnetic stimulation, TMS) – in that it relies on spontaneous rather than controlled changes in PPA state – but it is another interesting example of how neuroimaging data can be used to inform neuroscientific theories about how our brains enable our memories.

Conclusion

We have presented a number of examples of neuroimaging studies that we believe have enriched our understanding of human memory. For example, we have illustrated cases where neuroimaging has been informative in investigating memory processes that are difficult to access behaviorally. In other cases, neuroimaging provides additional sources of constraints (e.g., dissociations) that can be used to distinguish competing memory theories. Moreover, neuroimaging has not only offered additional ways to test existing theories, but has also facilitated the development of new experimental paradigms for behavioral studies, and provided the ability to address assumptions underlying some behavioral analysis methods.

We have emphasized that the value of neuroimaging hinges on the types of analysis and inference employed. While most neuroimaging studies have focused on the average activity within brain areas (or within time/frequency windows) and have been portrayed solely in terms of localizing a presumed memory process (in space or time), some neuroimaging studies have tried to reverse this inference, using neuroimaging data to determine whether a memory process occurred in a certain context. Furthermore, recent analysis techniques have started to utilize patterns of activity over voxels or times/frequencies, rather than just averaging that activity, and to consider what these patterns might represent. Other analyses have focused on memory-related changes in the communication between brain regions in terms of effective connectivity. These new analyses in turn force memory researchers to think carefully about how memory processes might be implemented in terms of neural representations and synaptic changes between neural populations. Such thoughts are best formalized in computational models of neuronal networks, which can then be tested in more detail with animal experiments.

Having said this, there are still deep philosophical issues that need to be considered when interpreting neuroimaging data. Issues related to the granularity of cognitive processes and resolvable brain areas, for example, must be considered when interpreting neuroimaging data, for example, for reverse inferences. We also acknowledge that not all neuroimaging studies of memory have made useful contributions to memory theories, and that the neuroimaging field continues to be plagued by tricky statistical issues that may question some published findings. Nonetheless, we do not think these are reasons to "throw the baby out with the bathwater."

Note

[1] This also raises the question of how the ROIs are defined in the first place, which is often based on traditional mass univariate analyses that search through the whole brain, though analogous searchlight methods exist to apply MVPA within a fixed volume, the center of which can be traversed across the entire brain image (Kriegeskorte *et al.* 2008).

References

Aggleton, J.P., and Brown, M.W. (1999). Episodic memory, amnesia, and the hippocampal–anterior thalamic axis. *Behavioral and Brain Sciences*, 22 (3), 425–444.

Bakker, A., Kirwan, C.B., Miller, M., and Stark, C.E.L. (2008). Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science*, 319, 1640–1642. doi: 10.1126/science.1152882.

Bamber, D. (1979). State-trace analysis: a method of testing simple theories of causation. *Journal of Mathematical Psychology*, 19 (2), 137–181. doi: 10.1016/0022-2496(79)90016-6.

Bartlett, F.C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge: Cambridge University Press.

Berry, C.J., Shanks, D.R., Speekenbrink, M., and Henson, R.N. (2012). Models of recognition, repetition priming, and fluency: exploring a new framework. *Psychological Review*, 119 (1), 40–79. doi: 10.1037/a0025464

Bridson, N.C., Muthukumaraswamy, S.D., Singh, K.D., and Wilding, E.L. (2009). Magnetoencephalographic correlates of processes supporting long-term memory judgments. *Brain Research*, 1283, 73–83. doi: 10.1016/j.brainres.2009.05.093.

Brown, A.A., and Bodner, G.E. (2011). Re-examining dissociations between remembering and knowing: binary judgments vs. independent ratings. *Journal of Memory and Language*, 65, 98–108.

Chadwick, M.J., Hassabis, D., Weiskopf, N., and Maguire, E.A. (2010). Decoding individual episodic memory traces in the human hippocampus. *Current Biology*, 20 (6), 544–547. doi: 10.1016/j.cub.2010.01.053.

Coltheart, M. (2006). What has functional neuroimaging told us about the mind (so far)? (Position paper presented to the European Cognitive Neuropsychology Workshop, Bressanone, 2005). *Cortex*, 42, 323–331

Deguchi, Y., Donato, F., Galimberti, I., *et al.* (2011). Temporally matched subpopulations of selectively interconnected principal neurons in the hippocampus. *Nature Neuroscience*, 14 (4), 495–504.

Diana, R.A., Reder, L.M., Arndt, J., and Park, H. (2006). Models of recognition: a review of arguments in favor of a dual-process account. *Psychonomic Bulletin Review*, 13 (1), 1–21.

Donaldson, D.I., Petersen, S.E., Ollinger, J.M., and Buckner, R.L. (2001). Dissociating state and item components of recognition memory using fMRI. *NeuroImage*, 13 (1), 129–142. doi: 10.1006/nimg.2000.0664.

Donaldson, D.I., Wilding, E.L., and Allan, K. (2003). Fractionating retrieval from episodic memory using event-related potentials. In *The Cognitive Neuroscience of Memory: Episodic Encoding and Retrieval* (ed. E.L. Wilding, A.E. Parker, and T.J. Bussey). Hove, UK: Psychology Press, pp. 39–58.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory and Cognition*, 24, 523–533.

Duarte, A., Henson, R.N., Knight, R.T., *et al.* (2010). Orbito-frontal cortex is necessary for temporal context memory. *Journal of Cognitive Neuroscience*, 22 (8), 1819–1831. doi: 10.1162/jocn.2009.21316.

Dunn, J.C. (2004). Remember–know: a matter of confidence. *Psychological Review*, 111, 524–542.

Dunn, J.C. (2008). The dimensionality of the remember–know task: a state-trace analysis. *Psychological Review*, 115 (2), 426–446. doi: 10.1037/0033-295x.115.2.426.

Dunn, J.C., and Kirsner, K. (1988). Discovering functionally independent mental processes: the principle of reversed association. *Psychological Review*, 95 (1), 91–101. doi: 10.1037/0033-295x.95.1.91.

Düzel E., Cabeza, R., Picton, T.W., *et al.* (1999). Task-related and item-related brain processes of memory retrieval. *Proceedings of the National Academy of Sciences of the USA*, 96, 1794–1799.

Evans, L.H., and Wilding, E.L. (2012). Recollection and familiarity make independent contributions to memory judgments. *Journal of Neuroscience*, 32 (21), 7253–7257. doi: 10.1523/jneurosci.6396-11.2012.

Fell, J., Klaver, P., Lehnertz, K., *et al.* (2001). Human memory formation is accompanied by rhinal-hippocampal coupling and decoupling. *Nature Neuroscience*, 4 (12), 1259–1264.

Fuentemilla, L., Penny, W.D., Cashdollar, N., *et al.* (2010). Theta-coupled periodic replay in working memory. *Current Biology*, 20 (7), 606–612. doi: 10.1016/j.cub.2010.01.057.

Gagnepain, P., Henson, R.N., Chételat, G., *et al.* (2010). Is neocortical–hippocampal connectivity a better predictor of subsequent recollection than local increases in hippocampal activity? New insights on the role of priming. *Journal of Cognitive Neuroscience*, 23 (2), 391–403. doi: 10.1162/jocn.2010.21454.

Gonsalves B.D., and Cohen, N.J. (2010). Brain imaging, cognitive processes, and brain networks. *Perspectives on Psychological Science*, 5, 744–752.

Greve, A., Donaldson, D.I., and van Rossum, M.C.W. (2010). A single-trace dual-process model of episodic memory: a novel computational account of familiarity and recollection. *Hippocampus*, 20 (2), 235–251. doi: 10.1002/hipo.20606.

Greve, A., van Rossum, M.C.W., and Donaldson, D.I. (2007). Investigating the functional interaction between semantic and episodic memory: convergent behavioral and electrophysiological evidence for the role of familiarity. *NeuroImage*, 34 (2), 801–814.

Hasselmo, M.E., and Howard, E. (2005). Hippocampal mechanisms for the context-dependent retrieval of episodes. *Neural Networks*, 18 (9), 1172–1190. doi: 10.1016/j.neunet.2005.08.007.

Haxby, J.V., Gobbini, M.I., Furey, *et al.* (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293 (5539), 2425–2430. doi: 10.1126/science.1063736.

Henson, R.N. (2005). What can functional neuroimaging tell the experimental psychologist? *Quarterly Journal of Experimental Psychology, Section A: Human Experimental Psychology*, 58 (2), 193–233. doi: 10.1080/02724980443000502.

Henson, R.N. (2006). Forward inference using functional neuroimaging: dissociations versus associations. *Trends in Cognitive Sciences*, 10 (2), 64–69. doi: 10.1016/j.tics.2005.12.005.

Henson, R.N. (2011). How to discover modules in mind and brain: the curse of nonlinearity, and blessing of neuroimaging. A comment on Sternberg (2011). *Cognitive Neuropsychology* 28 (3–4), 209–223. doi: 10.1080/02643294.2011.561305.

Herron J.E., and Wilding, E.L. (2004). An electrophysiological dissociation of retrieval mode and retrieval orientation. *NeuroImage*, 22, 1554–1562.

Hutzler, F. (2014). Reverse inference is not a fallacy per se: cognitive processes can be inferred from functional imaging data. *NeuroImage*, 84, 1061–1069. doi: 10.1016/j.neuroimage.2012.12.075.

Jacoby, L.L., Shimizu, Y., Daniels, K.A., and Rhodes, M.G. (2005). Modes of cognitive control in recognition and source memory: depth of retrieval. *Psychonomic Bulletin and Review*, 12 (5), 852–857.

Jacoby, L.L., and Whitehouse, K. (1989). An illusion of memory: false recognition influenced by unconscious perception. *Journal of Experimental Psychology: General* 118 (2), 126–135. doi: 10.1037/0096-3445.118.2.126.

Jafarpour, A., Horner, A.J. Fuentemilla, L., *et al.* (2013). Decoding oscillatory representations and mechanisms in memory. *Neuropsychologia*, 51 (4), 772–780. doi: 10.1016/j.neuropsychologia.2012.04.002.

Johnson, J.D., Muftuler, L.T., and Rugg, M.D. (2008). Multiple repetitions reveal functionally and anatomically distinct patterns of hippocampal activity during continuous recognition memory. *Hippocampus*, 18 (10), 975–980. doi: 10.1002/hipo.20456.

Kiebel, S., Garrido, M., Moran, R., and Friston, K. (2008). Dynamic causal modelling for EEG and MEG. *Cognitive Neurodynamics*, 2 (2), 121–136. doi: 10.1007/s11571-008-9038-0.

Kinoshita, S. (1997). Masked target priming effects on feeling-of-knowing and feeling-of-familiarity judgments. *Acta Psychologica*, 97 (2), 183–199.

Knowlton, B.J., and Squire, L.R. (1995). Remembering and knowing: two different expressions of declarative memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21 (3), 699–710.

Kriegeskorte, N., Mur, M., Ruff, D.A., *et al.* (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60 (6), 1126–1141. doi: 10.1016/j.neuron.2008.10.043.

Kurilla B.P., and Westerman, D.L. (2008). Processing fluency affects subjective claims of recollection. *Memory & Cognition*, 36, 82–92.

Lee, I., Yoganarasimha, D., Rao, G., and Knierim, J.J. (2004). Comparison of population coherence of place cells in hippocampal subfields CA1 and CA3. *Nature*, 430 (6998), 456–459.

Leutgeb, S., Leutgeb, J.K., Treves, A., *et al.* (2004). Distinct ensemble codes in hippocampal areas CA3 and CA1. *Science*, 305 (5688), 1295–1298. doi: 10.1126/science.1100265.

Lisman, J.E., and Otmakhova, N.A. (2001). Storage, recall, and novelty detection of sequences by the hippocampus: elaborating on the SOCRATIC model to account for normal and aberrant effects of dopamine. *Hippocampus*, 11 (5), 551–568. doi: 10.1002/hipo.1071.

Maguire, E.A., Mummery, C.J., and Büchel, C. (2000). Patterns of hippocampal–cortical interaction dissociate temporal lobe memory subsystems. *Hippocampus* 10 (4), 475–482. doi: 10.1002/1098-1063(2000)10:4<475::aid-hipo14>3.0.co;2-x.

Mayes, A., Montaldi, D., and Migo, E. (2007). Associative memory and the medial temporal lobes. *Trends in Cognitive Sciences*, 11 (3), 126–135. doi: 10.1016/j.tics.2006.12.003.

McIntosh, A.R. (1999). Mapping cognition to the brain through neural interactions. *Memory*, 7 (5–6), 523–548. doi: 10.1080/096582199387733.

Mecklinger, A. (2000). Interfacing mind and brain: a neurocognitive model of recognition memory. *Psychophysiology*, 37 (5), 565–582. doi: 10.1111/1469-8986.3750565.

Morcom, A.M., and Friston, K.J. (2012). Decoding episodic memory in ageing: a Bayesian analysis of activity patterns predicting memory. *NeuroImage*, 59 (2), 1772–1782. doi: 10.1016/j.neuroimage.2011.08.071.

Newell, B.R., and Dunn, J.C. (2008). Dimensions in data: testing psychological models using state-trace analysis. *Trends in Cognitive Sciences*, 12 (8), 285–290.

Norman, K.A., Polyn, S.M., Detre, G.J., and Haxby, J.V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10 (9), 424–430. doi: 10.1016/j.tics.2006.07.005.

Otten, L.J., Henson, R.N., and Rugg, M.D. (2002). State-related and item-related neural correlates of successful memory encoding. *Nature Neuroscience*, 5 (12), 1339–1344.

Paller, K.A., Voss, J.L., and Boehm, S.G. (2007). Validating neural correlates of familiarity. *Trends in Cognitive Sciences*, 11 (6), 243–250. doi: 10.1016/j.tics.2007.04.002.

Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, 10 (2), 59–63. doi: http://dx.doi.org/10.1016/j.tics.2005.12.004.

Poldrack, R.A. (2008). The role of fMRI in cognitive neuroscience: where do we stand? *Current Opinion in Neurobiology*, 18 (2), 223–227. doi: 10.1016/j.conb.2008.07.006.

Poldrack, R.A., and Wagner, A.D. (2004). What can neuroimaging tell us about the mind? Insights from prefrontal cortex. *Current Directions in Psychological Science*, 13, 177–181.

Polyn, S.M., Natu, V.S., Cohen, J.D., and Norman, K.A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, 310 (5756), 1963–1966.

Pratte, M.S., and Rouder, J.N. (2012). Assessing the dissociability of recollection and familiarity in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38 (6), 1591–1607. doi: 10.1037/a0028144.

Rajah, M.N., and McIntosh, A.R. (2005). Overlap in the functional neural systems involved in semantic and episodic memory retrieval. *Journal of Cognitive Neuroscience*, 17 (3), 470–482. doi: 10.1162/0898929053279478.

Rajaram, S. (1993). Remembering and knowing: two means of access to the personal past. *Memory & Cognition*, 21 (1), 89–102. doi: http://dx.doi.org/10.3758/bf03211168.

Rajaram, S., and Geraci, L. (2000). Conceptual fluency selectively influences knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26 (4), 1070–1074. doi: 10.1037/0278-7393.26.4.1070.

Ranganath, C., and Paller, K.A. (1999). Frontal brain potentials during recognition are modulated by requirements to retrieve perceptual detail. *Neuron*, 22 (3), 605–613.

Roediger, H.L. (1984). Does current evidence from dissociation experiments favor the episodic/semantic distinction? *Behavioral and Brain Sciences*, 7, 252–254.

Rotello, C.M., and Macmillan, N.A. (2006). Remember–know models as decision strategies in two experimental paradigms. *Journal of Memory and Language*, 55 (4), 479–494. doi: 10.1016/j.jml.2006.08.002.

Rugg, M.D., and Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Sciences*, 11 (6), 251–257. doi: 10.1016/j.tics.2007.04.004.

Rugg M.D., Wilding, E.L. (2000). Retrieval processing and episodic memory. *Trends in Cognitive Sciences*, 4, 108–115.

Schmidt, B., Marrone, D.F., and Markus, E.J. (2012). Disambiguating the similar: the dentate gyrus and pattern separation. *Behavioural Brain Research*, 226 (1), 56–65. doi: 10.1016/j.bbr.2011.08.039.

Shallice, T. (2003). Functional imaging and neuropsychology findings: how can they be linked? *NeuroImage*, 20, Supplement 1, S146–S154. doi: 10.1016/j.neuroimage.2003.09.023.

Squire, L.R., Wixted, J.T., and Clark, R.E. (2007). Recognition memory and the medial temporal lobe: a new perspective. *Nature Reviews Neuroscience* 8 (11), 872–883.

Staresina, B.P., Alink, A., Kriegeskorte, N., and Henson, R.N. (2013a). Awake reactivation predicts memory in humans. *Proceedings of the National Academy of Sciences of the USA*, 110 (52), 21159–21164. doi: 10.1073/pnas.1311989110.

Staresina, B.P., Fell, J., Dunn, J.C., *et al.* (2013b). Using state-trace analysis to dissociate the functions of the human hippocampus and perirhinal cortex in recognition memory. *Proceedings of the National Academy of Sciences of the USA*, 110 (8), 3119–3124. doi: 10.1073/pnas.1215710110.

Taylor, J.R., Buratto, L.G., and Henson, R.N. (2013). Behavioral and neural evidence for masked conceptual priming of recollection. *Cortex*, 49, 1511–1525.

Tendolkar, I., Rugg, M., Fell, J., *et al.* (2000). A magnetoencephalographic study of brain activity related to recognition memory in healthy young human subjects. *Neuroscience Letters*, 280 (1), 69–72. doi, 10.1016/S0304-3940(99)01001-0.

Treves, A., and Rolls, E.T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*, 4 (3), 374–391. doi, 10.1002/hipo.450040319.

Tulving, E. (1983). *Elements of Episodic Memory*. New York, NY: Oxford University Press.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne*, 26 (1), 1–12. doi: 10.1037/h0080017.

Tulving, E. (1987). Multiple memory systems and consciousness. *Human Neurobiology*, 6 (2), 67–80.

Tulving, E., and Thomson, D.M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80 (5), 352.

Uttal, W.R. (2001). *The New Phrenology: The Limits of Localizing Cognitive Processes.* Cambridge, MA: MIT Press.

Vazdarjanova, A., and Guzowski, J.F. (2004). Differences in hippocampal neuronal population responses to modifications of an environmental context: evidence for distinct, yet complementary, functions of CA3 and CA1 ensembles. *Journal of Neuroscience*, 24 (29), 6489–6496. doi: 10.1523/jneurosci.0350-04.2004.

Weber, M.J., and Thompson-Schill, S.L. (2010). Functional neuroimaging can support causal claims about brain function. *Journal of Cognitive Neuroscience*, 22 (11), 2415–2416. doi: 10.1162/jocn.2010.21461.

Wixted, J.T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, 114 (1), 152–176. doi: 10.1037/0033-295x.114.1.152.

Wixted, J.T., and Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review*, 117 (4), 1025–1054. doi: 10.1037/a0020874.

Yonelinas, A.P. (2002). The nature of recollection and familiarity: a review of 30 years of research. *Journal of Memory and Language*, 46 (3), 441–517. doi: 10.1006/jmla.2002.2864.

Yoo, J.J., Hinds, O., Ofen, N., *et al.* (2012). When the brain is prepared to learn: enhancing human learning using real-time fMRI. *NeuroImage*, 59 (1), 846–852. doi: 10.1016/j.neuroimage.2011.07.063.