WILEY HANDBOOKS
IN COGNITIVE NEUROSCIENCE

Edited by
Donna Rose Addis, Morgan Barense, and Audrey Duarte

THE WILEY HANDBOOK ON THE

Cognitive Neuroscience of Memory

WILEY Blackwell

The Wiley Handbook on the Cognitive Neuroscience of Memory

The Wiley Handbook on the Cognitive Neuroscience of Memory

Edited by

Donna Rose Addis, Morgan Barense, and Audrey Duarte

WILEY Blackwell

This edition first published 2015

Registered Office
John Wiley & Sons, Ltd., The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Offices
350 Main Street, Malden, MA 02148-5020, USA
9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, for customer services, and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

Library of Congress Cataloging-in-Publication Data

The Wiley handbook on the cognitive neuroscience of memory / edited by Donna Rose Addis, Morgan Barense, Audrey Duarte.
pages cm
Includes index.
ISBN 978-1-118-33259-7 (hardback)
1. Memory. 2. Cognitive neuroscience. 3. Brain–Imaging. I. Addis, Donna Rose, 1977–
II. Barense, Morgan, 1980– III. Duarte, Audrey, 1976–
QP406.W55 2015
612.8′23312–dc23
2015000669

A catalogue record for this book is available from the British Library.

Cover image: Background © happyperson / Shutterstock; profile head © amasterphotographer / Shutterstock

Set in 10/12pt Galliard by SPi Publisher Services, Pondicherry, India

1 2015

Contents

About the Editors

Donna Rose Addis is an Associate Professor and Rutherford Discovery Fellow in the School of Psychology at The University of Auckland. She received her PhD from the University of Toronto in 2005 and completed a postdoctoral fellowship at Harvard University. She has published 60 articles and chapters on autobiographical memory, future thinking, and identity. Dr. Addis has received a number of honors for her work in this area, including the prestigious New Zealand Prime Minister's MacDiarmid Emerging Scientist Prize and the Cognitive Neuroscience Society Young Investigator Award.

Morgan Barense is an Associate Professor in the Department of Psychology at the University of Toronto. She received her PhD from the University of Cambridge in 2006. Dr. Barense has published extensively on how memory functions are organized within the human brain and how memory relates to other cognitive processes, such as perception. She has received many accolades for this work, including a Canada Research Chair in Cognitive Neuroscience, the Early Investigator Award from the Society of Experimental Psychologists, and a Scholar Award from the James S. McDonnell Foundation.

Audrey Duarte is an Associate Professor in the School of Psychology at the Georgia Institute of Technology. She received her PhD from the University of California, at Berkeley in 2004. She has published numerous EEG, fMRI, and neuropsychological studies on age-related changes in episodic memory functioning, and held the Early Career Goizueta Professor Chair at the Georgia Institute of Technology.

About the Contributors

David Badre is an Associate Professor of Cognitive, Linguistic, and Psychological Sciences at Brown University and an affiliate of the Brown Institute for Brain Science. His lab at Brown studies cognitive control of memory and action, with a focus on frontal lobe function and organization. He was named an Alfred P. Sloan Fellow in 2011 and a James S. McDonnell Scholar in 2012, and currently serves on the editorial boards of *Cognitive Neuroscience* and *Psychological Science*.

Caitlin R. Bowman is a PhD student in cognitive psychology at the Pennsylvania State University studying with Dr. Nancy A. Dennis. Her work focuses on the neural basis of age differences in memory processing, particularly false memories. Her recent focus has been investigating the neural resources older adults utilize to avoid false memories. Her long-term goal is to identify causes of age-related memory decline and ways to enhance memory in older adults.

Alexis M. Chambers is a graduate student in the Cognition, Brain, and Behavior program at the University of Notre Dame. Her research spans a variety of domains, including sleep, memory, and emotion. Specifically, she is interested in exploring how sleep promotes selective emotional memory processing.

Christopher R. Cox is a graduate student in the Department of Psychology at the University of Wisconsin–Madison. His research focuses on the application of state-of-the-art methods for fMRI pattern analysis to questions about the neural bases of semantic memory.

Felipe De Brigard is an Assistant Professor of Philosophy at Duke University, and core faculty at the Center for Cognitive Neuroscience and the Duke Institute for Brain Sciences. He did his PhD at the University of North Carolina, Chapel Hill, and then spent two years as postdoctoral fellow in the Department of Psychology at Harvard University. His research centers on the interaction between memory and imagination, as well as the relationship between attention, consciousness, and recollection.

Nancy A. Dennis is an Assistant Professor at Pennsylvania State University, where she directs the Cognitive Aging and Neuroimaging Lab. Dr. Dennis is affiliated with the Center for Healthy Aging, the Social, Life, & Engineering Sciences Imaging Center, the Center for Brain, Behavior and Cognition, and the Huck Institute of Life Sciences.

Her research addresses the cognitive and neural mechanisms underlying age-related differences in memory. Her current work focuses on cognitive control, false memories, and association memories across the lifespan.

Ilana T. Z. Dew is a postdoctoral scholar at the Center for Cognitive Neuroscience at Duke University, where she uses behavioral and functional neuroimaging techniques to study human memory and emotion in young and older adults. She received her PhD in psychology from the University of North Carolina at Chapel Hill.

Danielle Douglas is a PhD student in psychology at the University of Toronto, studying cognitive neuroscience under the joint supervision of Andy Lee and Morgan Barense. Their research focuses on understanding memory processing in the human brain, particularly how memory interacts with perception.

Yana Fandakova is a postdoctoral research fellow at the University of California, at Berkeley. She studied psychology in Berlin and received her doctorate in psychology from the Humboldt Universität zu Berlin in 2012. Her primary research interests are the cognitive and neural mechanisms of developmental change across the lifespan, with a focus on episodic memory and cognitive control development in childhood and aging.

Chris M. Fiacconi is a Post-Doctoral Fellow in the Brain and Mind Institute at the University of Western Ontario. His research interests center broadly on human memory, with a focus on the relation between memory and affect, and the influence of prior experience on current perception and action. His research involves the use of cognitive experiments, psychophysiology, and the assessment of cognitive impairment in patients with various neurological conditions, including dementia and disorders of consciousness.

Kelly S. Giovanello is an Associate Professor of Psychology at the University of North Carolina at Chapel Hill, with appointments in the Biomedical Research Imaging Center and the Institute on Aging. She received her PhD in neuroscience from Boston University. Her research combines behavioral, patient-based, and functional neuroimaging approaches to investigate the cognitive neuroscience of human learning and memory.

Rebecca Gomez is an Associate Professor in the Psychology Department at the University of Arizona in Tucson. She completed a PhD at New Mexico State University before conducting postdoctoral research at the University of Arizona. Her research spans such topics as implicit learning, language acquisition, sleep-dependent memory consolidation, memory reconsolidation, and early learning systems, all with the aim of better understanding learning and memory mechanisms.

Cheryl L. Grady is a Senior Scientist at the Rotman Research Institute at Baycrest Centre in Toronto, Ontario. She is a Professor in the Departments of Psychiatry and Psychology at the University of Toronto, and holds a Tier I Canada Research Chair in Neurocognitive Aging. Her research focuses on age differences in large-scale functional connectivity of brain networks, the influence of lifelong experience (such as bilingualism) on brain structure/function, and variability of brain activity.

Andrea Greve is a Research Scientist at the MRC Cognition and Brain Sciences Unit in Cambridge, UK. Her primary research concerns the cognitive and neural basis of

human memory. One central question guiding her work focuses on how previously acquired knowledge influences the ways in which novel information is learned and retrieved. Her work aims to elucidate the interplay between episodic and semantic memories by combining different methods including behavioral, computational, and neuroimaging techniques.

Angela Gutchess is an Associate Professor of Psychology at Brandeis University, with appointments in Neuroscience and the Volen Center for Complex Systems. She received her PhD in Psychology from the University of Michigan. Her research investigates the influence of age and culture on memory and social cognition, using both behavioral and neuroimaging (fMRI) methods. She is particularly interested in how aging affects memory for self-relevant information and impressions of others.

Richard Henson is an MRC Programme Leader and Director for Neuroimaging at the MRC Cognition and Brain Sciences Unit in Cambridge, UK. His research focuses on the neural bases of memory, including the relationship between recollection, familiarity, and priming. He uses fMRI and EEG/MEG to examine brain activity as healthy volunteers attempt to remember information, and relates these findings, via computational modeling, to the memory problems in aging, amnesia, and dementia.

Michael Hornberger is a Senior Research Associate at the Department of Clinical Neurosciences, University of Cambridge, UK. His research focuses on memory processes in neurodegenerative patients to delineate underlying mechanisms of memory. At the same time, he develops novel memory and neuroimaging biomarkers to improve diagnostics and disease tracking in neurodegenerative disorders.

Almut Hupbach is an Assistant Professor of Psychology at Lehigh University. She received her PhD from the University of Trier, Germany. Her research focuses on the circumstances permitting induction of plasticity in human long-term memory. In particular, she studies the conditions that allow episodic memories to be modified and updated with new information. She is interested in unintentional and intentional processes of memory change, and studies these processes in children and adults.

Muireann Irish is a Research Fellow in the School of Psychology at the University of New South Wales, in Sydney, Australia. She conducts her research at Neuroscience Research Australia, Sydney, and is an Associate Investigator in the Australian Research Council Centre of Excellence in Cognition and its Disorders. Her research focuses on the disruption of episodic memory processes, such as autobiographical memory and future thinking, in neurodegenerative disorders, including Alzheimer's disease and frontotemporal dementia.

Jeffrey D. Johnson is an Assistant Professor in the Department of Psychological Sciences at the University of Missouri. His research focuses on understanding the cognitive and neural processes that contribute to episodic memory encoding and retrieval, through the use of electrophysiology (EEG and ERP), functional neuroimaging (fMRI), and pattern classification techniques.

Elizabeth A. Kensinger received her PhD from the Massachusetts Institute of Technology (MIT) and is Professor of Psychology at Boston College, where she directs the Cognitive and Affective Neuroscience laboratory. Her laboratory researches the

intersection between emotion and memory across the adult lifespan. She has co-authored numerous scientific publications on this topic and is the author of the book *Emotional Memory Across the Adult Lifespan* (Psychology Press, 2009).

Kerstin H. Kipp is a Senior Research Scientist at the Transfer Center for Neurosciences and Learning (ZNL) at Ulm University, Germany. After obtaining a dual degree in psychology and in communication and speech she defended her doctoral degree in cognitive psychology in 2003 at Saarland University in Saarbrücken. In 2011 she habilitated in the field of cognitive neuropsychology about episodic memory, its development and pathologies.

Stefan Köhler is a faculty member in the Department of Psychology and the Brain and Mind Institute at the University of Western Ontario, and is affiliated with the Rotman Research Institute, Toronto, Canada. His research addresses the neural and cognitive mechanisms of human memory, and the interface between memory and perception. His approach focuses on functional neuroimaging, cognitive experiments, and the study of memory impairments associated with various neurological conditions, including epilepsy and Alzheimer's disease.

Andy Lee is an Assistant Professor of Psychology at the University of Toronto and Adjunct Scientist at the Rotman Research Institute at the Baycrest Centre for Geriatric Care, Toronto. His research focuses on the contributions of the medial temporal lobe structures to mnemonic and perceptual processes.

Jackson C. Liang is a PhD candidate in the Institute for Neuroscience at the University of Texas at Austin under the mentorship of Dr. Alison Preston. His research focuses on the unique contributions of medial temporal lobe subregions to remembering category-level perceptual details and understanding how category representations are used to support memory-guided decisions.

Ulman Lindenberger directs the Center for Lifespan Psychology at the Max Planck Institute for Human Development, Berlin, Germany. He is an honorary professor at Freie Universität Berlin, Humboldt Universität zu Berlin, and Saarland University, Saarbrücken, Germany. His research interests are behavioral and neural plasticity across the lifespan, brain–behavior relations across the lifespan, multivariate developmental methodology, and formal models of behavioral change. He received the Gottfried Wilhelm Leibniz Prize from the Deutsche Forschungsgemeinschaft in 2010.

David Maillet is a Postdoctoral fellow at Harvard University. He has received postgraduate scholarships from the Natural Science and Engineering Research Council of Canada throughout his graduate training and is recipient of the FRQ-S Étudiants-chercheurs étoiles Award. He is interested in understanding how aging impacts the neural networks involved in episodic memory.

Chris B. Martin is a doctoral candidate in the Department of Psychology and the Brain and Mind Institute at the University of Western Ontario. His research is generally focused on human memory, with a specific interest in the cognitive and neural mechanisms that support familiarity-based recognition. His recent work has used functional neuroimaging to examine whether the neural correlates of familiarity assessment are organized in a category-specific manner in the medial temporal lobes.

Victoria C. McLelland is a postdoctoral fellow in the Department of Psychology at the University of Toronto, and her research is focused on the neural correlates of episodic memory. She received her PhD from the University of Auckland, where she investigated the role of the hippocampus in imagining and encoding future episodic events.

Axel Mecklinger is Full Professor for Neuropsychology at Saarland University in Saarbrücken and Speaker of the International Research Training Group "Adaptive Minds". Prior to this he worked as a Senior Research Scientist at the Max Planck Institute for Human Cognitive and Brain Sciences in Leipzig. In 1999 he received the Distinguished Scientific Award for Early Career Contributions to Psychophysiology from the Society for Psychophysiological Research. His research interests lie in the cognitive neuroscience of learning, memory, and cognitive control.

Laurie A. Miller is a Clinical Neuropsychologist at Royal Prince Alfred Hospital, Chief Investigator in the Australian Research Council Centre of Excellence in Cognition and Its Disorders as well as Senior Clinical Lecturer at the University of Sydney, Australia. She obtained a BSc degree from Westminster College in Pennsylvania and then Master's and PhD degrees in psychology from McGill University under the supervision of Professor Brenda Milner at the Montreal Neurological Institute. She pursued a two-year postdoctoral position at the University of Auckland and previously worked as a clinician and researcher at University Hospital in London, Ontario. Her interests include the study of memory disorders and their remediation.

Lynn Nadel is Regents' Professor of Psychology and Cognitive Science at the University of Arizona in Tucson. He proposed, with John O'Keefe, the "cognitive map" theory of hippocampal function and, with Morris Moscovitch, the "multiple trace" theory of memory and memory consolidation. He has worked on the role of the hippocampus in human memory, context, stress and anxiety disorders, and Down syndrome, bringing ideas in cognitive map theory to a wide range of research domains.

Erika Nyhus is a postdoctoral Research Associate in Cognitive, Linguistic, and Psychological Sciences at Brown University. She studies the neural processes involved in higher-level cognition, including executive functioning and episodic memory. Her research has addressed these topics through behavioral and neuroimaging (EEG, ERP, and fMRI) methods. This research has shown how frontal cortex, parietal cortex, and hippocampus transiently interact in support of controlled retrieval of episodic memories**.**

Jessica D. Payne is an Assistant Professor and Nancy O'Neill Collegiate Chair in Psychology at the University of Notre Dame. She is the Director of the Sleep Stress and Memory (SAM) Lab. Her research focuses on how sleep and stress independently and interactively influence human memory, emotion, performance, and creativity.

Bradley R. Postle is a Professor of Psychology and Psychiatry at the University of Wisconsin–Madison. He has over 60 peer-reviewed publications in scientific journals, many of these on the neural bases of human working memory and attention, and is associate editor at the *Journal of Cognitive Neuroscience* and *Cortex*.

Alison R. Preston received her PhD from Stanford University and is an Associate Professor of Psychology and Neuroscience at the University of Texas at Austin and a Fellow in the Center for Learning and Memory and the Institute for Neuroscience. Her research combines behavioral and brain imaging techniques to study how we form new memories, how we remember past experiences, and how memory for the past influences what we learn and do in the present.

Kylie A. Radford is a postdoctoral researcher at Neuroscience Research Australia, Clinical Neuropsychologist at Prince of Wales Hospital, and Conjoint Lecturer at the University of New South Wales, Sydney. She completed Doctorate of Clinical Neuropsychology/PhD degrees at the University of Sydney under the supervision of Dr. Laurie A. Miller and Dr. Suncica Lah. Her doctoral project focused on the impact of group-based rehabilitation of memory for patients with neurological disorders at Royal Prince Alfred Hospital.

M. Natasha Rajah is an Associate Professor in the Department of Psychiatry at McGill University, and Director of the Douglas Mental Health University Institute's Brain Imaging Centre. Her research focuses on using neuroimaging methods to understand how healthy and pathological forms of aging impact the neural structures, functions, and network interactions important for successful episodic memory encoding and retrieval.

Timothy T. Rogers is an Associate Professor in the Department of Psychology at the University of Wisconsin–Madison. His research focuses on the cognitive, neural, and computational bases of semantic memory. He is the author, with Jay McClelland, of *Semantic Cognition: A Parallel Distributed Processing Approach* (MIT Press, 2004).

Michael D. Rugg is Distinguished Professor in Behavioral and Brain Sciences and co-director of the Center for Vital Longevity at the University of Texas at Dallas. His research interests are in the cognitive and neural bases of human memory and the effects of aging and neurodegenerative disease on memory function.

Anthony J. Ryals is a postdoctoral fellow in the Department of Medical Social Sciences and the Interdepartmental Neuroscience Program, Northwestern University Feinberg School of Medicine, Chicago. Dr. Ryals studies how explicit and implicit processes operate in episodic memory by using behavioral experimentation as well as electrophysiology and functional neuroimaging. He also studies how subjective metacognitive awareness (or lack of awareness) relates to performance, brain function, and life quality in healthy and memory-impaired populations.

Daniel L. Schacter is William R. Kenan, Jr. Professor of Psychology at Harvard University. He received his PhD from the University of Toronto in 1981, and has since published over 350 research articles and chapters on various aspects of memory and the brain. He is the author of *The Seven Sins of Memory* (Houghton Mifflin, 2002) and several other books, and has received a number of awards for his research, including election to the National Academy of Sciences.

Yee Lee Shing is a research scientist at the Center for Lifespan Psychology at the Max Planck Institute for Human Development in Berlin, Germany. She received her doctorate in psychology from the Humboldt University Berlin in 2008. She is primarily interested in the development and plasticity of cognitive mechanics across the lifespan,

with a focus on brain–behavior relations of episodic memory components. She received the Heinz Maier-Leibnitz Prize from the Deutsche Forschungsgemeinschaft in 2012.

Volker Sprondel is currently a postdoctoral researcher at Saarland University in Saarbrücken. After graduating in psychology from the University of Konstanz, he started working as a doctoral researcher at Saarland University, from which he received his PhD in 2011. His research focuses on human memory and its development during childhood and adolescence from a cognitive neuroscience perspective.

Peggy L. St. Jacques is a Lecturer in the School of Psychology at the University of Sussex, Brighton, UK. She received her PhD from Duke University and completed a postdoctoral fellowship at Harvard University where she was the recipient of the L'Oréal USA for Women in Science Fellowship. Her research investigates autobiographical memory retrieval, aging, and emotion, and how reconstruction processes during retrieval can alter memory.

Indira C. Turney is a PhD student in cognitive psychology at the Pennsylvania State University, studying cognitive neuroscience under the supervision of Dr. Nancy Dennis. Her research examines the influence of age on a variety of memory processes. Her recent work focuses on the age-related neural markers underlying cognitive decline using functional and structural neuroimaging methods. Specifically, she is interested in understanding the cognitive and neural mechanisms mediating both true and false memories.

Melina R. Uncapher is a research scientist in the Stanford Memory Laboratory at Stanford University. Her research program focuses on the cognitive and neurobiological mechanisms that support learning and remembering, with an emphasis on understanding interactions between memory and attention. A prominent focus of her research involves the use of advanced neuroimaging techniques to investigate the functional heterogeneity of parietal cortex, and the role of this region in learning and remembering.

Joel L. Voss is an Assistant Professor based in the Department of Medical Social Sciences and the Interdepartmental Neuroscience Program, Northwestern University Feinberg School of Medicine, Chicago. Dr. Voss directs the Laboratory for Human Neuroscience, which studies brain mechanisms for memory and their disruptions due to neurological injury.

Preface

Experimental psychologists have been working to understand the cognitive processes that contribute to human memory for more than 100 years. Their research has yielded many important models and hypotheses regarding the cognitive and neural mechanisms that underlie short- and long-term memory. Neuroscience methods have advanced our understanding of human memory considerably, but many of these methods, including functional magnetic resonance imaging (fMRI), are relatively new and have been somewhat limited in their ability to resolve many of the controversies and mysteries of human memory functioning. Over the last several years, however, tremendous advances have been made in both the methods themselves and the analysis techniques used to examine the data. For example, due to the increased prevalence of higher field scanners and robust imaging sequences, the spatial resolution of neuroimaging methods has improved dramatically. This advancement has been instrumental in allowing cognitive neuroscientists, including contributors to this handbook, to distinguish the roles of human hippocampal subfields in episodic memory encoding and retrieval. Furthermore, highly sensitive multivariate statistical methods are being applied with increasing frequency, such that it is now possible to determine the extent to which a retrieved memory representation recapitulates that associated with initial encoding. Contributors to this volume use these and other cutting-edge approaches to address a wide variety of topics in the field of human memory, including the effects of healthy aging and dementia on episodic and semantic memory and their underlying neural substrates, the nature of the visual representations maintained in working memory and perception, the potential of therapeutic interventions for memory disorders, among many others.

Given the recent technological developments in neuroscience methods and the number of publications applying these methods, we felt that the timing was apt for this handbook. Although other excellent books on the topic of memory have been published over the last few years, this volume is unique in that it is focused on the recent neuroscience studies that have advanced our understanding of human memory. This book both summarizes this literature and makes predictions about the future studies that will advance our understanding further.

The contributors to this handbook represent many of the best cognitive neuroscientists working in the domain of human memory. We have endeavored to include a wide range of research areas and approaches in this volume. Each author has provided

an excellent review of the work being done in his or her own research area. The chapters are comprehensive enough that graduate students with some experience with the relevant literature can understand them; but also thought-provoking enough that experienced human memory researchers will find them interesting. We imagine that this book might also be useful for professors planning graduate seminars on topics of human memory.

We would like to extend our thanks to the authors who graciously accepted our invitation to contribute to this volume, and to the expert reviewers who provided invaluable feedback that helped perfect it. Needless to say, this handbook would not have been possible without their collective efforts.

Donna Rose Addis, Morgan Barense, and Audrey Duarte

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

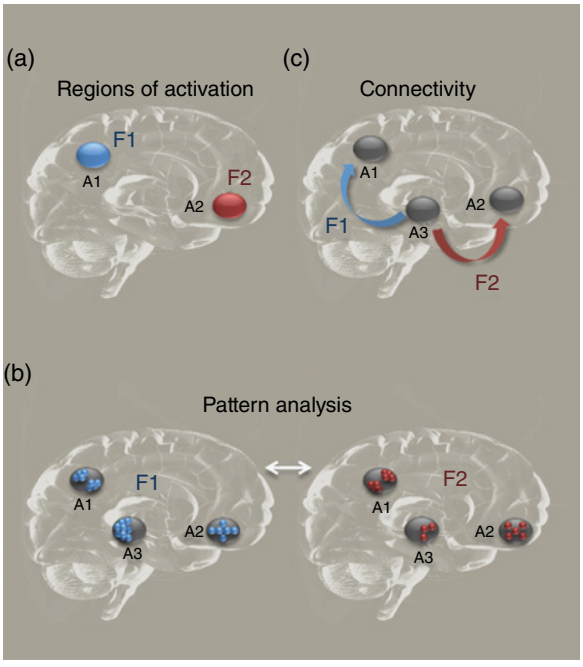

Figure 1.1 Schematic drawings of the human brain that illustrate different potential mappings of distinct memory functions (F1, F2) onto neural activity within and across distinct brain areas (A1, A2, A3). Changes in cognitive function can give rise to modulations in: (a) average levels of activity in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

2

Activation and Information in Working Memory Research

Bradley R. Postle

Introduction

In a 2006 review I wrote that "working memory functions arise through the coordinated recruitment, via attention, of brain systems that have evolved to accomplish sensory-, representation-, and action-related functions" (Postle, 2006a). By and large, ensuing developments in cognitive, computational, and systems neuroscience have been consistent with this perspective. One salient example is a 2011 special issue of *Neuropsychologia* that is devoted to the interrelatedness of the constructs of attention and working memory (Nobre and Stokes, 2011). Interestingly, however, the past several years have witnessed developments in the analysis of high-dimensional datasets, including those generated by functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and multi-unit extracellular electrophysiology, that, in some cases, call for a reconsideration of the interpretation of many of the studies that feature in the Postle (2006a) review and, indeed, in some of those in the more recent, above-mentioned special issue of *Neuropsychologia*. This chapter will consider some of the implications of these recent developments for working memory and attentional research. A second theme that has recently been gaining in prominence in working memory research, although by no means a "new development", is the critical role of network-level oscillatory dynamics in supporting working memory and attentional functions. This second theme has been covered extensively elsewhere, including in a recent chapter by this author (Postle, 2011), and so will not be addressed in detail here.

Activation and Information in the Interpretation of Physiological Signals

The signal-intensity assumption

The idea of sustained activity as a neural basis of the short-term retention (STR) of information (i.e., the "storage" or "maintenance" functions of short-term memory [STM] and working memory) has been a potent one that can be traced back at least as far as the reverberatory trace in Hebb's dual-trace model of long-term memory

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

(LTM) formation: the active reverberation within a circuit being the initial trace that served the function of the STR of the memory until synapses making up the circuit could be strengthened to create the (second) long-lasting trace (Hebb, 1949). Since the 1970s, neurons in the monkey (and other species) that demonstrate sustained activity throughout the delay period of delay tasks have been seen as a neural embodiment of this trace. First observed in prefrontal cortex (PFC) and mediodorsal thalamus by Fuster and Alexander (1971), sustained delay-period activity has since been observed in many brain areas, including not only "high level" regions of parietal and temporal cortex (e.g., Gnadt and Andersen, 1988; Nakamura and Kubota, 1995; Suzuki, Miller, and Desimone, 1997), but also, in a modality-dependent manner, in primary sensory cortex (e.g., Super, Spekreijse, and Lamme, 2001; Zhou and Fuster, 1996). In the human, delay-period neuroimaging signal with intensity that is elevated above baseline has long been considered a correlate of the STR of information (e.g., Courtney *et al.*, 1997; Jonides *et al.*, 1993; Zarahn, Aguirre, and D'Esposito, 1997), and the strength of this elevated signal, in comparison with other conditions, used to support models of the neural organization of working memory function. Thus, for example, statistically greater delay-period activity for, say, object information versus spatial information has been taken as evidence for the neural segregation of the STR of these two types of information (e.g., Courtney *et al.*, 1996; Owen *et al.*, 1998). The gold standard of evidence that a signal represents the STR of information has been evidence of monotonic increases in signal intensity with increasing memory load ("load sensitivity"; e.g., Jha and McCarthy, 2000; Leung, Seelig, and Gore, 2004; Postle, Berger, and D'Esposito, 1999; Todd and Marois, 2004; Xu and Chun, 2006).

Last to be reviewed here is the use of functional localizers to identify putatively category-selective regions of the brain. The classic example is that of the "fusiform face area" (FFA), a region in mid-fusiform gyrus that is typically found to respond with stronger signal intensity to the visual presentation of faces than of objects from other categories, such as houses (Kanwisher, McDermott, and Chun, 1997). In working memory and attention research, a commonly used strategy has been to identify "category-specific" regions of cortex with functional localizer scans (e.g., alternating blocks of faces versus houses), then to see how activity in these regions of interest (ROIs) varies during cognitive tasks that feature stimuli from these same categories. Thus, for example, a neural correlate of object-based attention is inferred when signal intensity in the FFA and in an analogous region of "house-selective" cortex is positively correlated with endogenous attentional cues, despite the fact that face and house stimuli are always present in a superimposed, translucent display (O'Craven, Downing, and Kanwisher, 1999). Similarly, neural correlates of the STR of face versus scene information are inferred from the fact that delay-period activity in an FFA ROI is greater for face memory than for house memory, and the converse is true for a "parahippocampal place area" (PPA) ROI (Ranganath, DeGutis, and D'Esposito, 2004).

Each of the types of experimental strategy reviewed in the preceding paragraph draws on a common underlying assumption, which is that one can infer the active representation of a particular kind of information from the signal intensity in a local area of the brain (see also Chapter 1). For expository expediency, I will refer to this as the *signal-intensity assumption*. In recent years, however, it has become increasingly clear that the signal-intensity assumption is subject to important limitations. Empirically, this has been seen in an increasing number of studies in which it fails to

account for working memory performance. And, as we shall see, an increasing appreciation for the multivariate nature of neuroimaging datasets (and, indeed, of brain function) provides a perspective from which the limitations of the signal-intensity assumption become clearer.

First, a brief review of empirical demonstrations reveals that the seemingly straightforward interpretation of elevated delay-period activity as serving a mnemonic function can be problematic. These can be organized into two categories: failures of specificity, and failures of sensitivity. The former refers to instances in which elevated delay-period activity can be shown to serve a function other than the STR of information; the latter, to instances in which behavioral performance makes clear that the subject is successfully remembering information, yet no evidence of elevated delay-period activity can be found.

Examples of failures of specificity include the finding that neurons with elevated delay-period activity in a memory task exhibit similarly sustained activity during the "delay" period of a visually guided saccade task, when no memory is required (Tsujimoto and Sawaguchi, 2004). Moreover, neurons that in a "standard" paradigm seem to encode a sensory representation of the to-be-remembered sample stimulus can be shown in a rotation condition to dynamically change during a single delay period from retrospectively representing the location of the sample to prospectively representing the target of the impending saccade (Takeda and Funahashi, 2002, 2004, 2007). Another example is a study designed to dissociate the focus of spatial attention from the focus of spatial memory that finds the majority of delay-active neurons to track the former (Lebedev *et al.*, 2004). (Limits of the specificity assumption will also factor importantly in the consideration of "reverse inference" in neuroimaging, which appears further along, in the section on *Implications of MVPA for ROI-based analyses.*)

Examples of failures of sensitivity include the finding that, in the monkey, STM for the direction of moving dots in a sample display can be excellent, despite the failure to find directionally tuned neurons, in either area MT or the PFC, that sustain elevated activity across the delay period (Bisley *et al.*, 2004; Zaksas and Pasternak, 2006). In a human fMRI study in which subjects maintained one of two different memory loads across a 24-second delay period, sustained, elevated delay-period activity was observed in several frontal and parietal sites during the delay, but none showed load sensitivity, leaving uncertain whether these regions were actually involved in storage (Jha and McCarthy, 2000).

Against this backdrop, there has been an increased appreciation for limitations of the univariate analytic framework within which hypotheses about differences in signal intensity are most commonly tested (see Chapter 1). With neuroimaging data, the most familiar approach is to solve the general linear model (GLM) in a mass univariate manner (e.g., Friston *et al.*, 1995). That is, the GLM is solved effectively independently at each of the (typically) thousands of data elements in a dataset. Typically, this approach leads to the identification of elevated or decreased signal intensity in voxels occupying a several-cubic-millimeter (or larger) volume of tissue, and the pooling across these voxels to extract a spatially averaged time-course. Using this univariate approach to implement the signal-intensity assumption often engages a second assumption that can also be problematic, that of homogeneity of function. That is, by pooling across "activated" (or "deactivated") voxels, one is assuming that all pooled voxels are "doing the same thing." Finally, the interpretation of the activity from this cluster

of voxels often entails a third, often implicit, assumption, which is that this locally homogeneous activity can be construed as supporting a mental function independent of other parts of the brain (i.e., modularity).[1] Each of these assumptions is difficult to reconcile with the increasingly common recognition that neural representations are high-dimensional, and supported by anatomically distributed, dynamic computations (e.g., Bullmore and Sporns, 2009; Buzsaki, 2006; Cohen, 2011; Kriegeskorte, Goebel, and Bandettini, 2006; Norman *et al.*, 2006).

Information-based analyses

An important conceptual advance in neuroimaging methods occurred with the publication by Haxby and colleagues (2001) of evidence that meaningful information about neural representations can be obtained from the patterns of activity in unthresholded fMRI data. This breakthrough was soon followed by the application of powerful machine-learning algorithms to fMRI datasets in an approach that has come to be known as multivariate pattern analysis (MVPA) (e.g., Haynes and Rees, 2006; Kriegeskorte, Goebel, and Bandettini, 2006; Norman *et al.*, 2006; Pereira, Mitchell, and Botvinick, 2009; see also Chapters 1 and 6). As the name implies, MVPA differs fundamentally from signal-intensity-based approaches in that it treats neural datasets as single high-dimensional images, rather than as a collection of independent low-dimensional elements. Therefore, it affords the detection and characterization of information that is represented in patterns of activity distributed within and across multiple regions of the brain. A detailed explication of the details underlying MVPA and its implementation to neuroimaging datasets is beyond the scope of this chapter, but what bears highlighting here is that MVPA is not subject to many of the problematic assumptions associated with signal-intensity-based analyses. This includes not only the assumptions of homogeneity of function and of modularity, but also, and most importantly for the topic of this chapter, the very assumption that the STR of information is accomplished via sustained, elevated activity. Indeed, tests of this assumption are the first applications of MVPA to working memory research that I will review here.

The possibility that the STR of information may not depend on sustained activity that is elevated above a baseline (typically, the inter-trial interval) was demonstrated by two MVPA studies of visual STM that focused on primary visual cortex (V1). These studies demonstrated that, although V1 did not show elevated activity during the delay period, it nonetheless contained representations of the to-be-remembered stimuli that spanned the delay period (Harrison and Tong, 2009; Serences *et al.*, 2009). In addition to building on what had been reported from V1 in the monkey (Super, Spekreijse, and Lamme, 2001), these studies clearly demonstrated the increased sensitivity of MVPA relative to signal-intensity-based analyses, in that no studies applying the latter to an fMRI dataset had previously implicated V1 in the STR of visual information.

A clear next step would be a direct test of the assumption that elevated delay-period activity carries trial-specific stimulus information. To implement it, Riggall and Postle (2012) acquired fMRI data during delayed recognition of visual motion, and analyzed the data with both GLM and MVPA. The former identified sustained, elevated delay-period activity in superior and lateral frontal cortex and in intraparietal sulcus (IPS), regions that invariably show such activity in studies of STM and working memory.

When we applied MVPA, however, the pattern classifiers implementing the analysis were unable to recover trial-specific stimulus information from these delay-active regions (Figure 2.1). This was not merely a failure of our MVPA methods, because the same classifiers successfully identified trial-specific stimulus information in posterior regions that had not been identified by the GLM: lateral temporo-occipital cortex, including the MT+ complex, and calcarine and pericalcarine cortex. Nor was it the case that the frontal and parietal regions were somehow "unclassifiable," because pattern classifiers were able to extract trial-specific task instruction-related information from these regions. Specifically, MVPA showed the frontal and parietal regions to encode whether the instructions on a particular trial were to remember the speed or the direction of the moving dots that had been presented as the sample stimulus, a finding consistent with previous reports from the monkey (Freedman and Assad, 2006; Swaminathan and Freedman, 2012). Thus, it is unlikely that the failure to recover stimulus-specific information from the frontal and parietal regions (i.e., that the to-be-remembered direction of motion was 42°, 132°, 222°, or 312°) is because they encode information on a finer spatial scale than the posterior regions for which item-level decoding was successful.[2] Rather, our conclusion is that the elevated delay-period activity that is measured with fMRI may reflect processes other than the storage, per se, of trial-specific stimulus information. Further, and consistent with previous studies (Harrison and Tong, 2009; Serences *et al.*, 2009), it may be that the short-term storage of stimulus information is represented in patterns of (statistically) "sub-threshold" activity distributed across regions of low-level sensory cortex that univariate methods cannot detect.

The finding from Riggall and Postle (2012) has potentially profound implications for our understanding of the neural bases of the STR of information, because it calls into question one of the more enduring assumptions of systems and cognitive neuroscience. We have reason to believe that it will hold up, because other groups are reporting

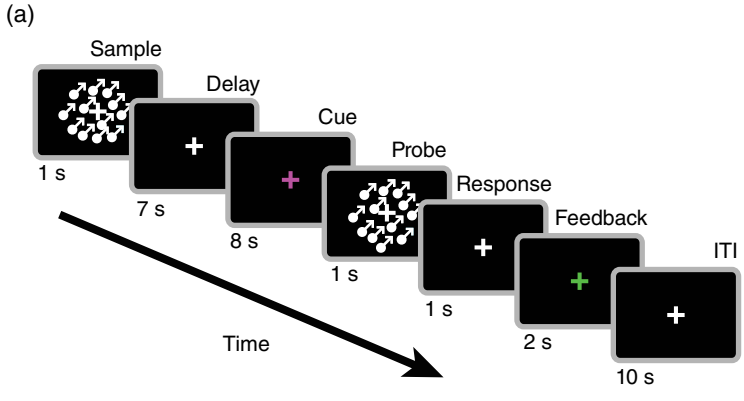

Figure 2.1 (a) Behavioral task from Riggall and Postle (2012). Subjects maintained the direction and speed of a sample motion stimulus over a 15-second delay period. Midway through the delay period, they were cued as to the dimension on which they would be making an upcoming comparison against the remembered sample, either direction or speed. At the end of the delay period, they were presented with a probe motion stimulus and had to indicate with a button press whether it did or did not match the sample stimulus on the cued dimension.

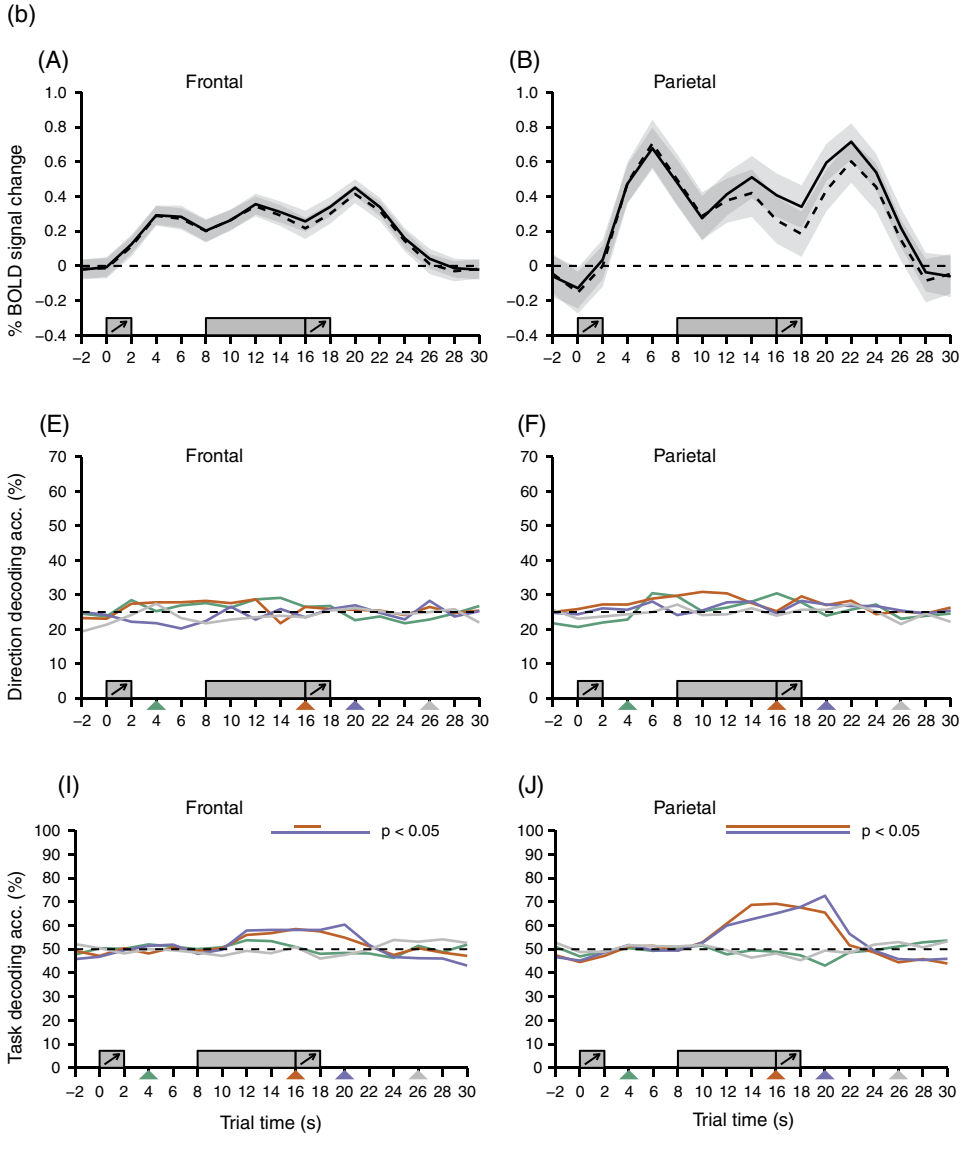

Figure 2.1 (continued) (b) BOLD and MVPA time-courses from four ROIs. Sample presentation occurred at 0 seconds, and at 8 seconds subjects were cued that either direction or speed of sample motion would be tested on that trial. (A–D) Average ROI BOLD activity. Data from direction-cued trials use solid lines and speed-cued trials use dashed-lines, bands cover average standard error across subjects. (E–H) ROI stimulus-direction decoding results and (I–L) ROI trial-dimension decoding results. Each waveform represents the mean direction-decoding accuracy across subjects ($n = 7$) for a classifier trained with data limited to a single time-point in the trial and then tested on all time-points in the hold-out trials (e.g., the green line illustrates the decoding

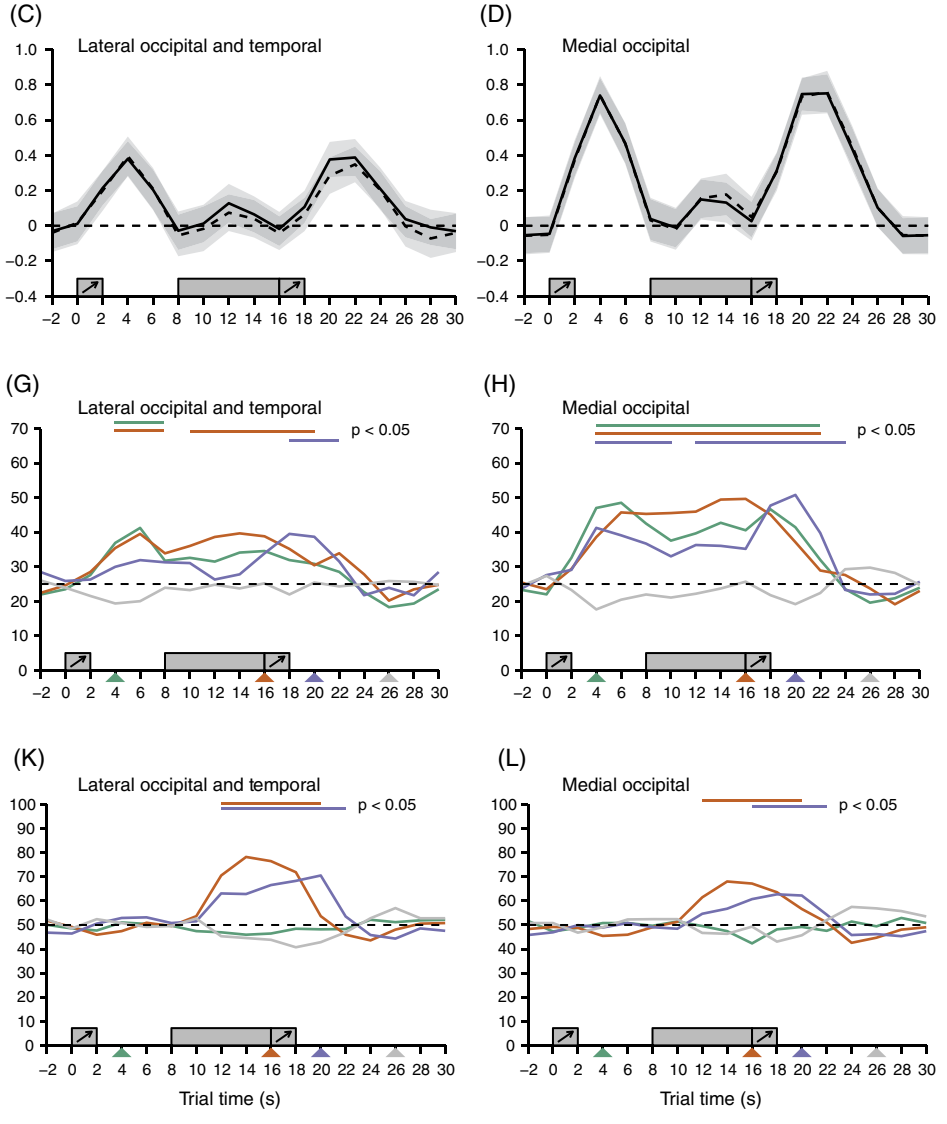

Figure 2.1 (continued) time-course from a classifier trained on only data from time-point 4, indicated by the small green triangle along the *x*-axis.) Horizontal bars along the top indicate points at which the decoding accuracy for the corresponding classifier was significantly above chance ($p < 0.05$, permutation test). Schematic icons of trial events are shown at the appropriate times along the *x*-axis. Data are unshifted in time. From Riggall, A.C., and Postle, B.R. (2012). The relation between working memory storage and elevated activity, as measured with fMRI. *Journal of Neuroscience*, 32, 12990–12998. © 2012. Reproduced with permission of the Society for Neuroscience.

compatible findings. For example, Linden and colleagues (2012) have reported a failure with MVPA to recover delay-period stimulus category information from frontal and parietal cortex, and Christophel, Hebart, and Haynes (2012) reported a failure to recover delay-period information specific to complex artificial visual stimuli from frontal cortex. The distinction between classifying at the item level (e.g., Christophel, Hebart, and Haynes, 2012; Riggall and Postle, 2012) versus at the category level (as by, e.g., Linden *et al.*, 2012) is an important one, in that the former provides the stronger evidence for memory storage, per se.

At this point it is useful to introduce the idea of an *active neural representation*. To illustrate, although Riggall and Postle (2012) contrasted signal intensity, a traditional index of "activation," versus classifiability, it is important to note that successful classification also depended on evaluation of levels of activity within individual voxels. Thus, there is an important distinction to be made between signal-intensity-based *activation*, which can be construed as a first-order physiological property,[3] and a multivariate-pattern-based *neural representation*, a second-order property (not just activity, but the pattern of activity) that is detectable with MVPA but not with univariate approaches. Nonetheless, a MVPA-detectable neural representation is an active representation, in the sense that neural activity must organize itself to create this pattern, and the neural representation is only present (i.e., only active) for the span of time that we assume it to be psychologically active. For example, Riggall and Postle (2012) found that MVPA of stimulus direction was only successful during a trial, when subjects were presumed to be thinking about a stimulus, and not during the inter-trial interval, when it is assumed that they were not.

Another way to illustrate this idea of an active neural representation is to consider information held in LTM. For example, Polyn *et al.* (2005) and Lewis-Peacock and Postle (2008) assumed that all of the US-citizen participants for their studies were familiar with the American actor John Wayne prior to volunteering for the studies. However, it is also assumed that none of them were *actively* thinking about John Wayne prior to being shown his image during the course of the study. Thus, there existed in the brains of these subjects an *inactive* neural representation of John Wayne that was not detectable by MVPA during portions of the experiment when subjects were not thinking about John Wayne. This neural representation became *active* when subjects were viewing an image of the actor, or retrieving this image from memory, and MVPA was sensitive to this change in the state of the LTM representation. The concept of an active neural representation is of central importance to the next studies to be reviewed here.

One of the questions raised by the Riggall and Postle (2012) findings is: What is the function of the sustained, delay-period activity that has been reported in the hundreds (if not more) of published studies on the neural correlates of STM and working memory since the 1970s? Several possible answers to this question (and it is almost certainly true that there are several answers) have been reviewed in the first section of this chapter. Our group has also begun addressing this question from the theoretical perspective that working memory performance may be achieved, in part, via the temporary activation of LTM representations. First, in a study employing MVPA that will not be reviewed in detail here, we established the neural plausibility of this idea (Lewis-Peacock and Postle 2008). In two more recent studies, we have worked from models that posit multiple states of activation, including, variously, a capacity-limited focus of attention, a region of direct access, and a broader pool of temporarily activated

representations, all nested within the immense network of latently stored LTM (Cowan, 1988; McElree, 2001; Oberauer, 2002). Importantly, these models distinguish the STR of information – which can be accomplished in any of the activated states of LTM – from attention to information – which is a capacity-limited resource that can be applied only to a small subset of highly activated representations.

The first of two studies that will be reviewed in this context (Lewis-Peacock *et al.*, 2012) was an fMRI study of a multi-step delayed-recognition task (adopted from Oberauer, 2005) (Figure 2.2). Each trial began with the presentation of two sample stimuli, always selected from two of three categories (lines, words, and pronounceable pseudowords), one in the top half of the screen and one in the bottom half (Figure 2.2b). After offset of the stimulus display and an initial delay period, a retrocue indicated which sample was relevant for the first recognition probe, followed by a second delay, followed by an initial Y/N recognition probe (and response). Critically, during the second delay both items needed to be kept in STM, even though only one was relevant for the first probe. This is because the first probe was followed by a second retrocue that, with equal probability, would indicate that the same item (a "repeat" trial) or the previously uncued item (a "switch" trial) would be tested by the trial-ending second Y/N recognition probe (and response). Thus, the first delay was assumed to require the active retention of two items, whereas the second delay would feature an "attended memory item" (AMI) and an "unattended memory item" (UMI).[4] The third delay would only require the retention of an AMI, because it was certain that memory for the item not cued by the second retrocue would never be tested. This design therefore allowed us to assess the prediction that there are different levels of neural activation corresponding to different hypothesized states of activation of LTM representations (Cowan, 1988; McElree, 2001; Oberauer, 2002).

Prior to performing this task, subjects were first scanned while performing a simple delayed recognition task (Figure 2.2a), and the data from this phase-1 scan were used to train the classifier that was then applied to the data from the multi-step task described in the previous paragraph. For phase 1, subjects were trained to indicate whether the probe stimulus matched the sample according to a category-specific criterion – synonym judgment for words, rhyme judgment for pseudowords, and an orientation judgment for line segments. Our rationale was that by training the classifier (separately for each subject) on data from the delay period of this task, we would be training it on patterns of brain activity related to the STR of just a single representational code: phonological (pseudoword trials), semantic (word trials), or visual (line trials). This, in turn, would provide the most unambiguous decoding of delay periods entailing the STR of two AMIs versus one AMI and one UMI versus one AMI.

In all trials, classifier evidence for both trial-relevant categories rose precipitously at trial onset and remained at the same elevated level until the onset of the first retrocue. This indicated that both items were encoded and sustained in the focus of attention across the initial memory delay, while it was equiprobable that either would be relevant for the first memory response. Following onset of the first retrocue, however, classifier evidence for the two memory items diverged. Post-cue brain activity patterns were classified as highly consistent with the category of the cued item, whereas evidence for the uncued item dropped precipitously, becoming indistinguishable from the classifier's evidence for the stimulus category not presented on that trial (i.e., not different from baseline). If the second cue was a repeat cue, classifier evidence for the already-selected memory item remained elevated and that of the uncued item remained indistinguishable

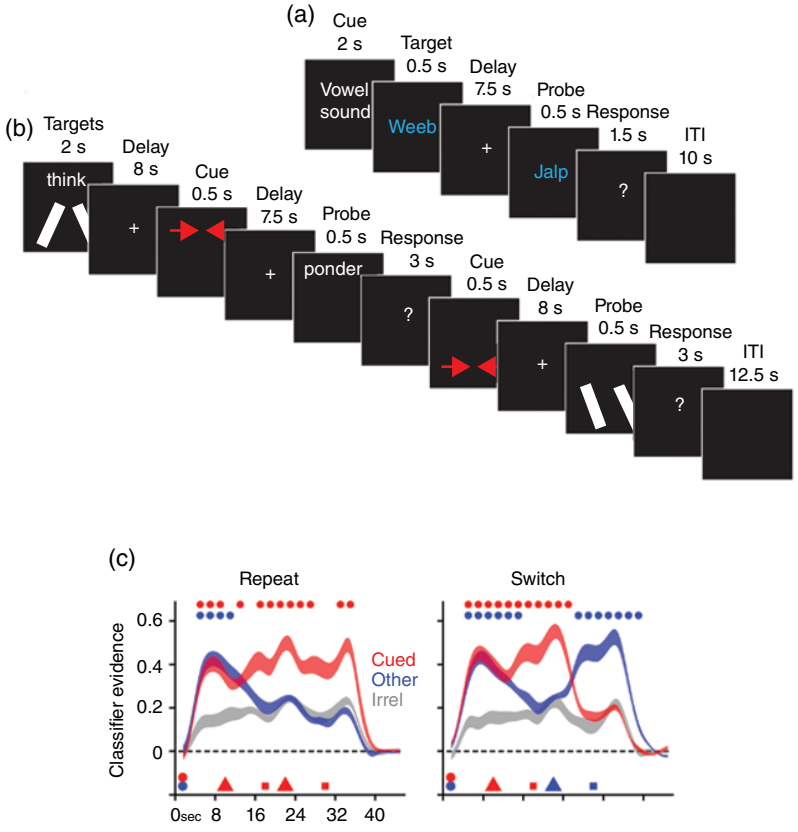

Figure 2.2 (a,b) Behavioral tasks from Experiment 2 of Lewis-Peacock *et al.* (2012). (a) In the first phase, subjects performed short-term recognition of a pseudoword (phonological STM), a word (semantic STM), or two lines (visual STM). (b) In the second phase, during the same scanning session, subjects performed short-term recognition with two stimuli (between-category combinations of pseudowords, words, and lines). On half of the trials, the same memory item was selected as behaviorally relevant by the first and second cues (repeat trials), and on the other half of trials the second cue selected the previously uncued item (switch trials). (c) Classifier decoding from Experiment 2 of Lewis-Peacock *et al.* (2012). Results are shown separately for repeat (left) and switch (right) trials. Classifier evidence values for phonological, semantic, and visual were relabeled and collapsed across all trials into three new categories: *cued* (red, the category of the memory item selected by the first cue), *other* (blue, the category of the other memory item), and *irrel* (gray, the trial-irrelevant category). The colored shapes along the horizontal axis indicate the onset of the targets (red and blue circles, 0 seconds), the first cue (red triangle, 10 seconds), the first recognition probe (red square, 18 seconds), the second cue (red or blue triangle, 22 seconds), and the final recognition probe (red or blue square, 30 seconds). Data for each category are shown as ribbons whose thickness indicate ±1 SEM across subjects, interpolated across the 23 discrete data points in the trial-averaged data. Statistical comparisons of evidence values focused on within-subject differences. For every 2-second interval throughout the trial, color-coded circles along the top of each graph indicate that the classifier's evidence for the *cued* or *other* categories, respectively, was reliably stronger ($p < 0.002$, based on repeated-measures *t*-tests, corrected for multiple comparisons) than the evidence for the *irrel* category. Reproduced with permission from Lewis-Peacock, J.A., Drysdale, A., Oberauer, K., and Postle, B.R. (2012). Neural evidence for a distinction between short-term memory and the focus of attention. *Journal of Cognitive Neuroscience*, 23, 61–79. © 2012 Massachusetts Institute of Technology.