From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Neuroscience

2016

Attention and memory in real time

Debettencourt, Megan T

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01rn301385z

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

Attention and memory in real time

Megan T. deBettencourt

a dissertation
presented to the faculty
of princeton university
in candidacy for the degree
of Doctor of Philosophy

recommended for acceptance
by the Princeton Neuroscience Institute

Advisers: Kenneth A. Norman & Nicholas B. Turk-Browne

September 2016

© Copyright by Megan T. deBettencourt, 2016. All rights reserved.

Abstract

Information streams towards the brain at every moment. The nervous system is responsible for processing this information without substantial delays, in order to attend to important information and store memories for future use. Accordingly, for neuroscientists, I propose that it is not sufficient to be able to deduce what *was* happening in the brain, but we should also understand what *is* happening in the moment. Real-time neural analyses offer the opportunity to do so, by closing the loop between experimenter and participant.

By tracking neural processes in real time, we can return rapid feedback to modify cognition, such as attention and memory. In Chapter 1, I used this approach to track and provide feedback about fluctuations of sustained attention in real time. When asked to sustain our attention to something, attention naturally fluctuates. I developed a paradigm to deliver real-time feedback about fluctuations of attention concurrent to task performance. Then, I observed how and whether this feedback modified cognitive abilities. This novel approach of closed-loop neurofeedback benefited attention performance and increased the neural discriminability of the attended information.

In addition to providing feedback, real-time analyses can also be used to optimize experimental design and influence memory encoding. Information encoded when attention is lapsing is much less likely to be later remembered. In Chapter 2, by monitoring fluctuations of attention, the delivery of information was timed to be at the right or wrong moment, when attention was measured to be high or low. This manipulation affected later memory and contributed to our understanding of the tight link between attention and memory.

Finally, after memories have already been encoded, their retrieval can also be manipulated. The retrieval of a memory is facilitated when it occurs in the same context as was present when the memory was initially encoded. In Chapter 3, I used real-time neurofeedback to guide participants closer or farther away from the original encoding context during retrieval, influencing what and how much they remembered.

The work presented in this dissertation opens up several avenues to explore and perturb attentional and mnemonic processes and the neural mechanisms that support them.

Contents

Abstract iii

0 Introduction 1
- 0.1 Towards a real time understanding of cognition . . . 1
- 0.2 Background on brain-computer interfaces . . . 2
- 0.3 Real-time fMRI . . . 4
- 0.4 Techniques for returning feedback . . . 8
- 0.5 Real-time fMRI neurofeedback to facilitate learning . . . 13
- 0.6 Summary . . . 15

1 Closed-loop training of attention with real-time brain imaging 17
- 1.1 Introduction . . . 18
- 1.2 Real-time neurofeedback . . . 23
- 1.3 Training effects in behavior . . . 26
- 1.4 Training effects in the brain . . . 30
- 1.5 Contribution of specific brain systems . . . 32
- 1.6 Discussion . . . 35
- 1.7 Methods . . . 38

2 Forgetting from lapses of sustained attention 55
- 2.1 Introduction . . . 56
- 2.2 Experiment 1 . . . 58
- 2.3 Experiment 2 . . . 64
- 2.4 General discussion . . . 68

3 Externalizing mental context reinstatement in real time 71
- 3.1 Abstract . . . 71
- 3.2 Introduction . . . 72

3.3 Real-time neurofeedback . . . 77
3.4 Behavioral effects . . . 78
3.5 Replaying context . . . 83
3.6 Whole-brain activity differences during feedback . . . 87
3.7 Discussion . . . 89
3.8 Methods . . . 93

4 Conclusion and future directions 103
4.1 Conclusion . . . 103
4.2 Clinical applications . . . 104
4.3 Future directions . . . 105

References 106

Listing of figures

1.1 Study procedure 21
1.2 Real-time pipeline 23
1.3 Brain-behavior relationship 25
1.4 Classifier-to-stimulus transfer function 26
1.5 Change in behavior 28
1.6 Average RTs surrounding lures 30
1.7 Searchlight analyses 32
1.8 Potential sources of feedback 34
1.9 Histogram of feedback 46
1.10 Removing the influence of RT 50

2.1 Experiment design 60
2.2 Relating sustained attention and memory 63
2.3 Online adaptive experiment design 66
2.4 The fate of triggered trials 67

3.1 Study procedure 75
3.2 Classifier evidence during feedback 78
3.3 Effects of cue validity on memory recall performance 79
3.4 Context reinstatement correlates with behavior 81
3.5 Effects of feedback on memory recall performance 85
3.6 Feedback mediates the link between context reinstatement and memory . . . 86
3.7 Whole-brain univariate activity during context reinstatement 88
3.8 Voxelwise correlations with feedback 89

Acknowledgments

I am immensely grateful to my advisors, Ken Norman and Nick Turk-Browne, for their time and attention. None of the work contained in this thesis would have been possible without them and their prompt replies to all of my incessant emails. Over the course of my PhD, I met weekly with both of my advisors and often also with Jon Cohen, the third member of my thesis committee. Our discussions were invaluable for developing the projects from their beginnings, and their feedback helped shape the projects into what they became. I continue to marvel at the passion and dedication that they have each brought to every one of our meetings.

I benefited enormously from the collaborative neuroscience community and the incredible computational and imaging resources at Princeton. I would like to thank all of my labmates and friends in the department for their support over the years. I am especially indebted to all of my officemates over the years who tolerated me and my loud keyboard. I would also like to thank my cohort, who have been with me since the first day of class here. Beyond Princeton, I am thankful for my collaborators in Austin who let me explore clinical applications of my work as well as my undergraduate advisors and labmates who led me to Princeton.

Finally, I would not be here without the love and encouragement from my parents, my brother, and especially Manu. Thank you so much for all that you have done for me.

0

Introduction

0.1 Towards a real time understanding of cognition

Neuroscientists and psychologists embark on understanding a cognitive phenomenon by designing an experiment. Traditionally, such an experiment can be described by discrete experimental conditions (e.g. face and scene images, lines of different orientations, or directing attention to particular regions of space). These conditions are assumed to elicit different neural representations and/or behavior, which can be analyzed and understood through the data collected during the experiment. After the experiment is complete and all data are collected, analyses are used to understand how the neural responses differed between the conditions. Recent analytical and computational tools have progressed to the

point where this delay between data collection and analysis is no longer necessary. Analyzing the data on the fly can increase the power of our experiments and improve our understanding of the brain.

For psychologists, real-time analysis opens an avenue of real-time neurofeedback. Rather than treating the brain as a black box, information about their mental state can be communicated to participants in real time, in order to engage, perturb, modify, and even improve cognitive function. This requires a fundamental shift in the experimental design, in that the experiments are no longer cleanly segmented, but continuous and interactive: Neurofeedback affects the brain, which affects our measurements, which affects the next moment of feedback.

Neurofeedback can also externalize to participants otherwise opaque internal cognitive processes. Using neurofeedback, participants can be made aware of the fluctuations of their mental processes, which they might not otherwise be aware of until it is too late. Provided with such feedback, participants can use this information to adjust their strategies, and it can be determined whether this creates long-lasting changes in behavior.

0.2 Background on brain-computer interfaces

Real-time neurofeedback is one form of brain-computer interface (BCI). This technique has been developed for both invasive and non-invasive modalities, with goals ranging from answering basic science questions to creating clinical, educational or occupational applications. (BMIs) can be considered a class of BCIs, with the goal of characterizing motor-related signals for potential neuroprostheses (Sadtler et al., 2014; Serruya et al., 2002) However, BCIs can investigate cognitive functions beyond motor control, including how at-

tention is deployed and controlled (Cerf et al., 2010; deBettencourt et al., 2015; Schafer & Moore, 2011).

BCIs involving invasive neuronal recordings were able to reveal the relationship between external and internal attention, and how this influences representations in the medial temporal lobes (Cerf et al., 2010). This was accomplished using single unit activity measured via intracranial electroencephalography. Researchers first measured "concept cells" in the medial temporal lobe, which selectively and specifically respond to a particular concept (e.g., a celebrity). Then, participants were presented with a composite image, overlaying pictures of two different concepts. Participants were given a target concept, and were instructed to increase the neural activity for that image. Critically, feedback was provided by linking the image visibility with the neural activity for the related concept. They demonstrated that, with feedback, participants could upregulate the activity of a target concept, and thus fade in the image. By externalizing internal attention through neurofeedback, it allowed for a more successful and precise targeting of attention towards specific information. However, given the limitations intracranial EEG, it was unknown whether this orienting signal towards the target concept originated within the medial temporal lobe. Additionally, this study did not examine whether there any behavioral consequences for this manipulation. Nonetheless, it demonstrated that providing attention can bias representations in the medial temporal lobe.

A related study demonstrated that volitional control with feedback can have important ramifications for behavior (Schafer & Moore, 2011). In this study, investigators were interested in the attentional control signals from the frontal eye fields, an area of prefrontal cortex. Monkeys were provided auditory feedback based on the spiking activity in frontal

eye fields. First, researchers demonstrated that with feedback, monkeys were able to up- and down-regulate the activity in this region. Moreover, controlling the activity in this area had profound effects on behavior for a visual attention task. Interestingly, these effects were not found by naturally occurring positive or negative deflections in neural activity that were not associated with feedback. They did not compare feedback from frontal eye fields with feedback from other regions or other indices (e.g. behavioral performance). However, this study demonstrated that providing real-time feedback to demonstrate neural control can be used to characterize the role of the frontal eye fields in attention.

Although both of these studies used invasive measures of neural activity, there have also been numerous examples of neurofeedback from noninvasive modalities. Usually such neurofeedback has relied on scalp EEG, given its high temporal resolution (Parra et al., 2003). Although functional magnetic resonance imaging (fMRI) is another popular non-invasive modality, there has been relatively little research into its potential as a BCI. This might be partly due to the fact that fMRI measures blood oxygenation level, and this signal has a relatively low temporal resolution due to its dependence on hemodynamic response (Logothetis et al., 2001). On the other hand, fMRI has relatively high spatial resolution, which can be important for disentangling the contribution and interaction from different brain regions (Turk-Browne, 2013). Relatively recently, scientists have been developing BCIs and providing neurofeedback using fMRI data.

0.3 Real-time fMRI

One of the earliest descriptions of real-time fMRI, the technique of collecting fMRI data, analyzing it, and providing outcomes to the participant in the scanner, was over 20 years

ago (Cox et al., 1995; Goddard et al., 1997). The technical hurdles are considerable given the high-dimensionality of the whole-brain fMRI data – approximately 100,000 samples every two seconds. Real-time fMRI is an exciting bridge between engineering and psychology, and there has been growing interest in using this technique for clinical as well as basic science research (Stoeckel et al., 2014; Sulzer et al., 2013).

One of the earlier demonstrations of the potential for real-time fMRI was for pain suppression (deCharms et al., 2005). In this study, participants were presented with a flame on the screen, the size of which corresponded to the univariate activity in the rostral anterior cingulate cortex (rACC). Participants were capable of up- and down-regulating rACC activity with feedback, and this affected their perception of pain. Down-regulating rACC activity was found to be beneficial for patients suffering from chronic pain. This paper highlighted the potential clinical applications of real-time fMRI and inspired additional basic science research into this technique.

A prominent study demonstrated that even implicit training via real-time fMRI could produce robust changes in behavior as well as arbitrate theoretical debates in the field of cognitive neuroscience (Shibata et al., 2011). Shibata and colleagues used real-time fMRI to address whether plasticity in early visual cortex was sufficient to support perceptual learning. A subject-specific pattern classifier was trained to discriminate different orientations in early visual cortex. Then, participants underwent multiple days of training, where they were told to increase the size of a dot on the screen. Critically, they were not told that the dot corresponded to the evidence for a particular target orientation. Although the strategies were varied and unrelated to oriented lines, participants succeeded at upregulating the activity for the target orientation. Following this training, participants were improved sen-

sitivity in a discrimination task only for the target (trained) orientation. This technique of implicit or "incepted" real-time fMRI neurofeedback has since been used in an additional study to examine the role of early visual cortex in forming associations (Amano et al., 2016). By not instructing participants how to best achieve a desired neural state, incepted neurofeedback can demonstrate that activating certain neural patterns alone can cause a certain behavioral result.

Knowledge of how the fMRI signal is fluctuating has also been used to trigger experimental conditions (Hinds et al., 2013; Yoo et al., 2012). If By waiting until the fMRI signal is high or low, trials can be initiated when the signal is high or low. This allows precise targeting of neural activity. One of the earlier attempts first demonstrated that lower parahippocampal cortex activity resulted in better memory for scene images (Yoo et al., 2012). Then, parahippocampal cortex activity was monitored and when it was especially low, a scene image was presented. These images were better remembered than when scene images were triggered on the basis of parahippocampal cortex activity was high. Another study observed that ongoing neural fluctuations affected vigilance performance (Hinds et al., 2013). In this study, it was observed that the activity measured in supplemental motor area (relative to default mode netwok) was correlated with reaction time in a sustained attention task. This correlation was further examined by triggering a target detection task. One potential problem to this approach is that fMRI measures blood oxygenation levels, and this hemodynamic signal is lagged 4-6 seconds (Logothetis et al., 2001). Nonetheless, these studies demonstrate the feasibility of using fMRI for triggering designs. This use of real-time fMRI to initiate trials can allow experiments to more directly identify the behavioral and neural consequences of pretrial neural activity.

If experimenters wish to characterize a relationship between fMRI activity and behavior, real-time manipulations have the potential to be more powerful than standard methods. For example, in triggering designs, real-time fMRI forces the variable of interest, the fMRI activity, to be true for that relationship (e.g. high or low). When an experiment is designed with traditional, randomized condition structure, the fMRI activity is likely sampled from some normal distribution over the range of activities. With real-time analyses, we can restrict the data to be from certain otherwise under-sampled regions, such as the tails of distributions, and better understand the relationship between neural activity and behavior. It is possible also that returning neurofeedback recruits additional mechanisms in the brain to comprehend and respond to feedback (Papageorgiou et al., 2013). From a theoretical perspective, psychologists typically design experiments to 'cause' some behavior, but real-time fMRI opens the possibility for more directly examining the causes and consequences of particular fMRI activity behavior.

Feedback about brain activity can also be provided with the goal of increasing, decreasing, or sustaining a certain level or pattern. Recently, researchers demonstrated the capability of using real-time fMRI as a type of "cognitive neurostimulation" (MacInnes et al., 2016) to learn to increase activity in the ventral tegmental area of the brain for motivational purposes. Relatedly, feedback can be thought of as a type of cognitive patch clamp, by having participants regulate the activity of particular regions (or possibly networks) in the brain. By incentivizing people to maintain activity at a particular level, this can encourage them to increase, decrease or sustain neural activity at a particular level. In this way, neurofeedback can be used to partial out and test hypotheses about the roles of particular brain regions for producing behaviors of interest.

0.4 Techniques for returning feedback*

Many different techniques have been employed in real-time fMRI neurofeedback designs. Traditionally, after processing fMRI data in real time, a summary measure has been presented on the experimental display via a meter or scale of some sort (deCharms et al., 2005). Other experiments have employed triggering designs in which neural activity was used to initiate an experimental sequence, such as a trial onset (Hinds et al., 2013; Yoo et al., 2012). More recently, closed-loop neurofeedback has been developed to deliver feedback concurrent to task peformance and create adaptive designs (deBettencourt et al., 2015; Lorenz et al., 2016). Finally, recent prominent work has used incepted feedback, in which the participants are instructed to regulate a meter on the screen without being told what it corresponds to, or even that it corresponds to neural activity at all.

One of the challenges in neuroscience is that isolating the role of particular brain regions in learning is complicated by the dense interconnectivity of the brain. When stimuli are displayed, this sets off a cascade of neural activity throughout the visual processing stream. Any of these areas, ranging from early visual cortex to hippocampus, could be candidates for forming associations between stimuli, making it challenging to tease apart the role of each of these areas in association formation. Real-time functional Magnetic Resonance Imaging (rtfMRI) neurofeedback may provide a non-invasive means of establishing where learning is taking place. In a recent study, Amano and colleagues employed this technique to demonstrate the role of early visual cortex in associative learning (Amano et al., 2016).

RtfMRI neurofeedback merges fMRI data collection, analysis, and stimulus presentation into a single closed-loop procedure. fMRI data files are processed as they arrive, by

*Some of the contents of this section were published in deBettencourt & Norman (2016)

comparing them against a baseline level of activity (deCharms et al., 2005; MacInnes et al., 2016) or pre-trained multivariate pattern classifier (deBettencourt et al., 2015; Papageorgiou et al., 2013; Shibata et al., 2011). Feedback about the fMRI measurements is then communicated to the participant, often by updating the display (Sulzer et al., 2013).

In most published rtfMRI neurofeedback studies, participants are informed of the nature of the feedback and given instructions about how to maximize it (deBettencourt et al., 2015; deCharms et al., 2005; MacInnes et al., 2016; Papageorgiou et al., 2013). However, this is not strictly necessary: Feedback without instruction can result in lasting neural and behavioral changes (Shibata et al., 2011). One powerful demonstration of this technique addressed the role of early visual cortex in perceptual learning: Participants were asked to maximize a feedback signal that (unbeknownst to them) reflected whether brain activity corresponding to a particular line orientation was present in early visual cortex; after repeatedly activating these orientation-specific patterns, participants demonstrated improved ability to discriminate between that orientation and other orientations. The authors of (Shibata et al., 2011) referred to this as "incepted" neurofeedback since participants were learning about specific perceptual details without awareness of those details, in homage to the science-fiction movie "Inception" where information was inserted into people's dreams.

Amano and colleagues extended this "inception" approach to ask whether early visual cortex can support the formation of associations between an orientation and a color (Amano et al., 2016). First, they trained a multivariate pattern classifier to discriminate red versus green stimuli based on activity patterns in early visual cortex elicited by viewing those colors. Next, participants underwent a training protocol designed to coactivate a particular orientation and color (red). Critically, during this protocol, the red color was never present

onscreen. Participants were repeatedly presented with an achromatic grating of a particular orientation and given feedback about their neural activity. They were told to maximize this feedback, but not given specific instructions about how to do this. Unbeknownst to the participants, this feedback corresponded to how well their brain pattern (in the present) matched the brain pattern (at training) associated with perceiving red. In a post-test, participants showed a significant tendency to perceive the trained orientation in the color targeted with neurofeedback; this bias was specific to the trained orientation. These results suggest that it is possible to induce associations, without awareness or even stimulus presentation, using activity patterns from early visual cortex.

Importantly, Amano and colleagues did not merely want to show that association learning happens in this paradigm; they also made the strong claim that the color-orientation associations were formed within early visual cortex. To substantiate this claim, the authors completed multiple analyses aimed at mapping out the neural evidence for color information in the brain. For these analyses, the authors used pattern classifiers applied to localized brain areas ("searchlights") to decode fine-grained patterns of activity. First, they demonstrated that color was widely decodable in early and higher-level visual areas during the classifier-training run (when color was visible). That is, when color was on the screen, color information robustly spread through a broad swath of cortex. Then, the authors completed searchlight analyses during the feedback runs. During periods of neurofeedback, the areas where color could be decoded were largely restricted to the early visual cortex, which was the locus of the feedback. To the extent that information about the color did not spread beyond early visual cortex, this suggests that early visual cortex was the only place where associative learning about color occurred.

It is possible that follow-up experiments will be able to indicate with even more certainty that the activity is not spreading into higher visual areas. Most critically, it would be useful to compare these results with feedback from another region. Given that early visual cortex represents visual information in a localized fashion, training directed at a limited region of the visual field or one eye should not transfer completely to other regions or the other eye if early visual cortex supplies the feedback. The authors could contrast these effects with a follow-up study in which participants receive feedback based on color-related information from a different region, such as V4. Insofar as higher-level visual areas have broader receptive fields, we would expect that more transfer between locations and between eyes would occur in this case. Any interesting differences between the associations (e.g., in time of acquisition, duration, or specificity) as a function of the neural locus of feedback could illustrate important differences between nearby neural regions in associative learning. By comparing different regions, neurofeedback could be used to characterize how the nature of associations can change throughout the visual processing hierarchy.

The apparently limited spread of color information in the neurofeedback condition highlights a unique benefit of the "inception" approach, in comparison to overtly presenting the color onscreen or having participants deliberately imagine the color. As noted above, when participants are aware of the color, color-related information can spread all the way up the visual hierarchy, making it possible for participants to draw upon general-purpose associative learning mechanisms in the medial temporal lobes. Numerous studies have shown that the hippocampus can rapidly bind arbitrarily-related features that are present in awareness (Eichenbaum & Cohen, 2001), even if the features pertain to very low-level stimulus features like orientation (Bosch et al., 2014) and even if the to-be-associated

features are only present in memory, not in perception (Gershman et al., 2013; Ramirez et al., 2009; Rudy & O'Reilly, 2001). By limiting the spread of color information to early visual cortex, the neurofeedback procedure used by Amano and colleagues (Amano et al., 2016) diminishes the likelihood of engaging higher-level association-formation mechanisms.

In making this point, the study by Amano and colleagues (Amano et al., 2016) does not just shed light on associative learning — it also helps to show more generally how real-time fMRI neurofeedback can be used as a tool to study the neural basis of cognition. Work that will be described in Chapter 1 shows that real-time fMRI neurofeedback can be used to amplify internal cognitive states by "externalizing" them. Here, feedback about fluctuations in sustained attention was provided by making the to-be-attended stimuli more or less visible, with the goal of making participants more aware of these fluctuations and thus better able to control them; this externalization procedure reduced attention lapses and improved behavior. The Amano study (Amano et al., 2016) showed almost the opposite benefit – instead of using real-time fMRI to make participants more aware of a process that is usually hard to control, this procedure makes it possible to take a process (color perception) that is usually conscious, and renders it unconscious. This study makes it clear that this "inception" procedure is not just a parlor trick: The unconscious nature of this percept appears to limit its spread in the brain, thereby making it possible to make sharper inferences about where learning is taking place.

0.5 Real-time fMRI neurofeedback to facilitate learning[†]

In our daily lives, we often receive rich feedback when we've made an error. However, this feedback might be a delayed or distorted representation of the underlying cognitive state that caused the error. In this regard, real-time knowledge about the underlying mental representations might be useful. The use of real-time fMRI to provide feedback can be geared toward facilitating learning that is difficult to achieve, or less efficient, with other methods. This can be contrasted with alternative uses of real-time fMRI, such as triggering task events or optimizing task parameters, where the goal is not necessarily to produce a lasting learning effect. To understand the learning induced by real-time fMRI, three aspects of this process should be considered: (1) How is learning measured? (2) Which mechanisms are responsible? and (3) How meaningful are the effects? We will describe these three aspects in turn, and then briefly discuss a case study.

Learning can be said to occur when experience influences some component of behavior and/or alters brain structure or function. In the case of real-time fMRI, the relevant experience can consist of elements of the task, neurofeedback about brain activation, representations, or connectivity, and cognitive processes that arise due to the task, feedback, or attempts to control feedback with strategies. The consequences of this experience can be assessed using various dependent measures linked to learning in more standard experiments in cognitive psychology and neuroscience.

In behavior, learning can be reflected in improved perception (Fahle & Poggio, 2002), memory recall and recognition (Yonelinas, 2002), anticipation/prediction (Bubic et al., 2010), priming (Tulving & Schacter, 1990), and motor action (Stadler & Frensch, 1998). In

[†]Some of the contents of this section were published within Stoeckel et al. (2014)

the brain (particularly fMRI), learning can be reflected in enhanced (Schwartz et al., 2002) or attenuated (Grill-Spector et al., 2006; Turk-Browne et al., 2008) activation within sensory systems, activation in learning and memory systems (Brewer et al., 1998; Pagnoni et al., 2002; Poldrack et al., 2001; Wagner et al., 1998), changes to the multivariate representational space of brain regions (Folstein et al., 2013; Schapiro et al., 2013), changes to functional connectivity (Büchel et al., 1999), increased gray matter volume (Draganski et al., 2004), and altered white matter (Zatorre et al., 2012). All of these measures are potential targets for real-time fMRI studies seeking to induce learning with neurofeedback.

These changes in behavior and the brain can result from different mechanisms. Neurofeedback from a brain region may produce a type of instrumental conditioning, whereby activation of that region becomes rewarding. An increase in the activity of the region, and possibly additional inputs from the reward system and regions involved in cognitive control, may induce local plasticity. This plasticity could be reflected in alterations of the selectivity and circuitry of neurons in that region (Sur & Rubenstein, 2005) or larger-scale consequences. Learning may occur over a vast range of timescales, from immediately (in the case of priming) to over weeks (in the case of perceptual learning).

Relatedly, effects persist for different durations depending on the type of learning and brain system involved, from milliseconds for adaptation in the visual system (Grill-Spector et al., 2006) to years for episodic memories consolidated in association cortex (McClelland et al., 1995). Also, learning procedures vary in terms of whether the effects generalize to other contexts or not, from being hyperspecific to the training context (Jiang & Song, 2005) to more flexible (Turk-Browne & Scholl, 2009).

These three aspects of learning — dependent measures, neural mechanisms, and tim-

ing/generalization properties — allow real-time fMRI neurofeedback studies to be classified and opportunities for new research to be identified. Take, for example, a study of perceptual learning induced by real-time fMRI (Shibata et al., 2011). The dependent measures in this study consisted of improved accuracy in identifying a trained orientation and improved multivariate classification of this orientation from early visual cortex. The proposed mechanism is that neurofeedback encouraged participants to generate orientation-specific neural activity patterns in this region, which resulted in local plasticity. The properties of learning in this task are that it took 5-10 sessions of training over a month, the effects lasted at least that long, and the benefit of training was specific to one orientation at a particular contrast. Various aspects of this classification could be investigated in future studies, such as what other brain systems are responsible for creating the activity patterns in visual cortex, and whether other forms of perceptual learning (Xiao et al., 2008) could lead to more generalized benefits.

0.6 Summary

Real-time experimentation in general, and especially real-time fMRI, offers new avenues for engaging dynamically with cognitive functions and may provide a nuanced view on the relationship between mental activity and behavior. There are many unanswered practical as well as theoretical questions in this burgeoning field. In practice, real-time fMRI can be complementary to traditional fMRI experiments for understanding the brain. However, despite its great promise, it remains relatively underutilized. Real-time neurofeedback confers the benefit of continuous and customizable feedback about specific mental representations. It makes it possible to distinguish between situations where behavior might be

correct, but the neural representations are wrong, and provides participants with the opportunity to correct this.

1

Closed-loop training of attention with real-time brain imaging*

Here we used closed-loop neurofeedback to improve sustained attention abilities and reduce the frequency of lapses. During a sustained attention task, the focus of attention was monitored in real time with multivariate pattern analysis of whole-brain neuroimaging data. When indicators of an attentional lapse were detected in the brain, we gave human participants feedback by making the task more difficult. Behavioral performance improved after one training session, relative to control participants who received feedback from other

*The contents of this chapter were published in deBettencourt et al. (2015)

participants' brains. This improvement was largest when feedback carried information from a frontoparietal attention network. A neural consequence of training was that the basal ganglia and ventral temporal cortex came to represent attentional states more distinctively. These findings suggest that attentional failures do not reflect an upper limit on cognitive potential and that attention can be trained with appropriate feedback about neural signals.

1.1 Introduction

Our ability to sustain attention over long periods of time is limited, both in the laboratory (Chun et al., 2011; Mackworth, 1948) and in the real world (Czeisler et al., 2005; Redelmeier & Tibshirani, 1997). This has been demonstrated using vigilance tasks in which participants monitor for and detect infrequent stimuli (Dinges & Powell, 1985; Sarter et al., 2001; Wolfe et al., 2005). Behavior in these tasks is predictive of attention disorders (Johnson et al., 2007) and is reliable over time (Robertson et al., 1997). Within the normal population, there is considerable variability in attentional abilities as measured by these tasks, and this variability is related to other perceptual and mnemonic processes (Drew & Vogel, 2008). We hypothesized that lapses in these tasks — and in life — occur because humans do not adequately monitor how well they are attending from moment to moment. Lapses emerge gradually and may be detected too late, after the chain of events that produces behavioral errors has been initiated. Accordingly, one way to train sustained attention might be to provide a more sensitive feedback signal, such that participants can learn to sense upcoming lapses earlier and prevent them from manifesting in behavior.

To pursue this approach, we created a continuous feedback signal customized to each

participant, reflecting moment-to-moment variations in their sustained attention. Participants were presented with a series of composite stimuli containing a mixture of information relevant and irrelevant to the task. Online analysis was used to track their attentional state, operationalized as the amount of task-relevant information active in their brains minus the amount of task-irrelevant information. Finally, this measure was provided to participants as feedback by altering the appearance of the next stimulus. When participants were attending well (that is, more task-relevant information was detected in their brains), we increased the proportion of task-relevant information in the stimulus. Conversely, when they were attending poorly (that is, more task-irrelevant information was detected), we reduced the proportion of task-relevant information in the stimulus. In this way, we amplified the consequences of their attentional state, rewarding them with a stronger stimulus and an easier task for staying focused and punishing them with a degraded stimulus and a more difficult task for lapsing. We hypothesized that this would make attentional lapses more salient and that participants would be able to exploit this feedback to learn to improve their sustained attention.

For online analysis, we combined real-time functional magnetic resonance imaging (rtfMRI) (LaConte, 2011; Sulzer et al., 2013; Weiskopf et al., 2004) with multivariate pattern analysis (MVPA) (Norman et al., 2006). The rtfMRI component of the system involved immediately acquiring measurements of the blood oxygen level-dependent (BOLD) response over the whole brain. This technique has been used previously to display univariate activity for pain regulation (deCharms et al., 2005), to display multivariate activity for inducing perceptual learning (Shibata et al., 2011) and to trigger stimulus presentation based on univariate activity in brain regions associated with memory encoding (Yoo et al., 2012)

and vigilance (Hinds et al., 2013). Our approach was related to these latter triggering designs, in the sense that brain states were used to control stimuli rather than controlling a separate feedback scale or gauge, but differed in that the stimulus triggered by a brain state at one moment in time influenced the brain state at the next moment, which in turn influenced the next stimulus, and so on. In other words, after a stimulus was triggered, the trial did not end and there was no delay imposed before the next stimulus could be triggered (Hinds et al., 2013; Yoo et al., 2012). This approach of continually updating task stimuli as they perturb brain states has been referred to as "closed-loop" (LaConte, 2011). The MVPA component of the system decoded differences in whole-brain BOLD activity patterns reflecting attention to the task-relevant versus task-irrelevant stimuli. The combination of MVPA and rtfMRI is well suited for rapidly decoding distributed cognitive processes such as attention.

There are other, simpler ways of delivering real-time feedback: for example, based on electroencephalography (EEG), eye tracking or manual responses. We used rtfMRI for two reasons. First, in combination with advanced analytical techniques, fMRI may provide more direct access to internal brain states. For instance, we sought to identify which specific kind of information a participant was attending to over time rather than whether they were attentive in general (often called alertness, arousal or mindfulness), as is reflected in pupil size (Yoss et al., 1970) and response time variability (Rosenberg et al., 2013). Second, by using fMRI, we not only gain a sensitive neural measure for feedback but also the ability to characterize the neural mechanisms that support attention training. We take advantage of this opportunity by considering both how training alters the brain and which brain regions provide useful feedback signals for training. We do not claim that this is the only or best

approach for training attention, but simply that it may prove valuable because of its sensitivity, its ability to generate neuroscientific data that can help constrain our interpretation, and its potential to lay the foundation for further advances in the use of other methods.

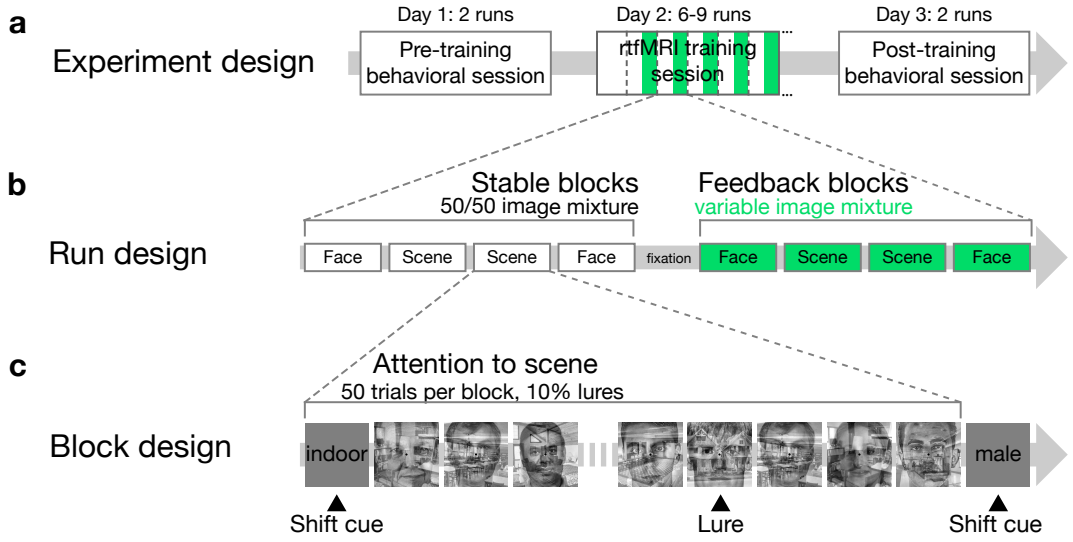

Figure 1.1: Study Procedure. (a) Participants completed three sessions on different days, with different numbers of task runs. (b) Each run contained eight blocks of the sustained attention task. The pre-training, post-training, and stable blocks of the rtfMRI training sessions contained composite stimuli with an equal mixture of faces and scenes. In the feedback blocks of the rtfMRI session, the mixture of images was determined by real-time analysis of brain activity. (c) Each block began with a cue (1 s) that indicated the attended category (e.g., scene) and target subcategory (e.g., indoor). This was followed by a brief fixation period (1 s) and then 50 sequential stimuli (1 s each, no interstimulus interval), of which 90% were targets and 10% were lures. The blocks ended with a fixation period (4—6 s).

This study involved three sessions on different days: behavioral pre-training, rtfMRI training and behavioral post-training (Figure 1.1). Participants performed the same sustained attention task in all sessions, viewing blocks of face/scene composite stimuli. Before each block, they were cued to attend to one task-relevant category (for example, scenes) and were instructed to ignore the other, task-irrelevant category (in this case, faces). Within the

task-relevant category, they responded ('go' trial) if the image was from a specified target subcategory that appeared with high frequency throughout the study (for example, indoor scenes; 90% of trials). They withheld their response ('no-go' trial) for the other, infrequent lure subcategory (in this case, outdoor scenes; 10%) (Robertson et al., 1997; Rosenberg et al., 2013). Sustained attention was assessed behaviorally using signal detection measures. The average false alarm rate from the behavioral pre-training session was 0.31 (s.e.m. = 0.03). In other words, college-aged adults made 30% errors in an ostensibly trivial task, which demonstrates that sustained attention abilities were limited at the start of the study, as expected.

During the rtfMRI training session, each of several training runs contained eight blocks of the sustained attention task in a counterbalanced design. The first four, 'stable' blocks were used for MVPA training and the last four, 'feedback' blocks were used for neurofeedback. During the stable blocks, composite stimuli were presented with a mixture of 50% face and 50% scene. A whole-brain classifier was trained over a moving window of recent stable blocks to discriminate attention to faces versus scenes. This attentional manipulation is known to elicit distinct patterns of neural activity (Al-Aidroos et al., 2012; O'Craven et al., 1999). During feedback blocks, the trained classifier was used to decode in real time which category was being attended. The output was then used to update the mixture of the composite stimulus for the next trial (Figure 1.2).

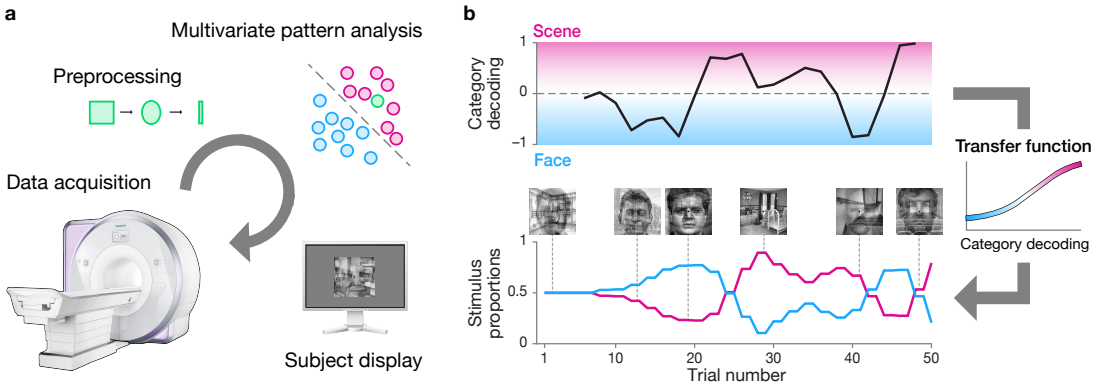

Figure 1.2: Real-time pipeline. (a) During feedback blocks, each brain volume (green) was acquired, preprocessed with masking, smoothing and z-scoring, and analyzed during the next volume with a multivariate classifier trained on volumes from recent stable blocks in which faces (blue) or scenes (pink) were attended. The result was averaged with the results for the two preceding volumes and used to update the stimulus shown to the participant on trials in the subsequent volume. (b) The classifier output indicated how attentive the participant was to the task-relevant versus task-irrelevant categories. This output was converted to a mixture proportion using a sigmoidal transfer function: less attention to the task-relevant category resulted in a decrease in the proportion of that category's image in the composite stimulus on the next trial. These values were updated throughout the block as attention fluctuated over time.

1.2 Real-time neurofeedback

The classifier's output would be useful for training attention only if it provided an accurate measurement of attentional state (that is, attention to face versus scene). To assess the validity of this measure, we performed n-fold cross-validation on the stable blocks. Note that bottom-up stimulation in these blocks was identical at the category level regardless of whether participants were instructed to attend to faces or to scenes. The average decoding accuracy was 0.78 (s.e.m. = 0.02), which was highly reliable relative to chance (0.50) across participants ($p < 0.00001$, bootstrap resampling). This robust decoding validated our measure of top-down attentional state.

As a further preliminary step, we sought to verify that the classifier's output was mean-

ingfully related to participants' behavior (Figure 1.3). Across participants, there was a strong positive correlation between decoding accuracy in the stable blocks and performance in the earlier behavioral pre-training session ($r = 0.70, p = 0.000008$, Spearman rank correlation). This relationship was also evident within participants (Leber et al., 2008; Weissman et al., 2006): behavioral accuracy on no-go trials — that is, whether participants correctly withheld their response or responded incorrectly — could be predicted using the classifier evidence for the task-relevant category from the brain volumes immediately preceding trial onset (correct rejection: mean evidence = 0.78, s.e.m. = 0.02; false alarm: mean evidence = 0.74, s.e.m. = 0.02; $p < 0.00001$). This effect remained robust after controlling for response time (RT) differences (Figures 1.3 and 1.5). This further confirmed that the classifier provided a predictive and behaviorally relevant measure of attention.

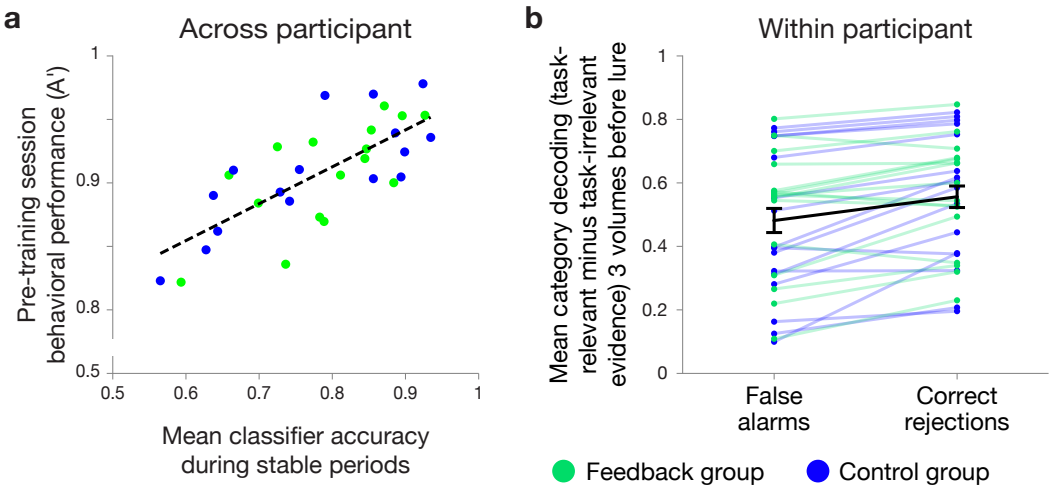

Figure 1.3: Brain-behavior relationship. To verify that the classifier could provide useful feedback, we examined how predictive it was of behavior. (a) Across participants, average decoding accuracy from the stable blocks of the rtfMRI session (determined by offline MVPA with n-fold cross-validation) was highly correlated with behavioral performance in the pre-training session. (b) Within participants, there was greater classifier output for the task-relevant category than for the task-irrelevant category before correctly rejecting than before false alarming to a lure trial. Error bars represent ±1 s.e.m.

The feedback blocks used real-time classifier output to modulate the proportion of task-relevant versus task-irrelevant information in the composite stimuli. As outlined above, the proportion of the task-relevant stimulus on the next trial was increased when there was greater neural evidence of the task-relevant category in the preceding trial, whereas it was decreased when there was greater neural evidence of the task-irrelevant category. The motivation for weakening the task-relevant image when measures of attention waned was to amplify and externalize the consequences of the participant's attentional state, providing them with an error signal, with the goal of increasing their self-monitoring ability. The opposite — strengthening the task-relevant image when attention lapsed — might have stabilized performance at that moment, but it may also have incentivized such lapses by making the

task easier, thus undermining learning. The precise mapping between classifier output and mixture proportion was controlled by a sigmoidal transfer function (Figure 1.4).

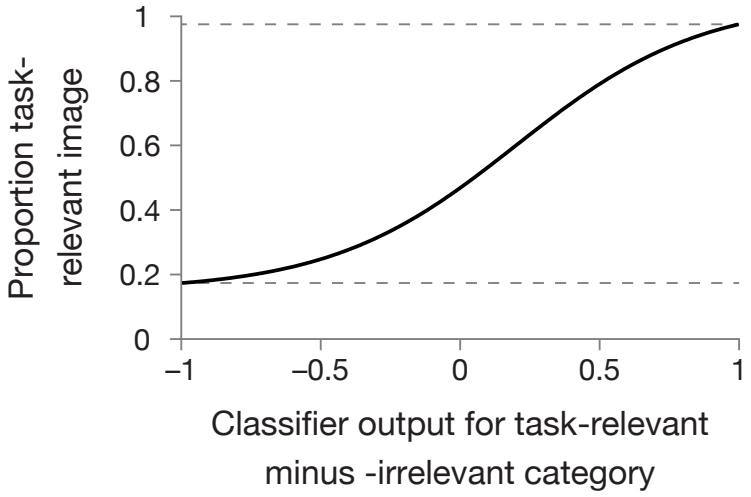

Figure 1.4: Classifier-to-stimulus transfer function. The volume-by-volume classifier output for the task-relevant minus task-irrelevant category was mapped to the proportion of the image from the task-relevant category using a sigmoidal function. The inflection point on the classifier axis was set to 0.60, based on the average decoding accuracy in a pilot study. Given the nonlinearities in the function, this helped calibrate the feedback to a more sensitive range of classifier values. Image proportion ranged from 0.17 to 0.98, preventing the task-relevant image from ever disappearing completely, and providing a foothold for recovery from a serious lapse.

1.3 Training effects in behavior

The rtfMRI neurofeedback produced a significant training effect: behavioral sensitivity improved from the pre-training session to the post-training session ($p = 0.01$; Figure 1.5). This improvement was quantitatively related to what happened during the training session, as it could be predicted by the extent to which a participant's neurofeedback became more positive over time ($r = 0.78, p = 0.002$). To further verify that improved sensitivity was the result of accurate neurofeedback, we collected data from 16 control participants

who were each uniquely matched in age, gender and handedness to one of the 16 feedback participants. During the pre-training session, there was no difference between the groups in false alarm rate ($p = 0.72$) or sensitivity ($p = 0.90$). Control participants were given identical instructions and underwent the same procedure, except that, during the feedback blocks, their feedback was yoked to their matched participant in the experimental group, rather than to measures of their own attentional state. This yoking ensured that control participants were exposed to the same overall stimulus statistics and variations in task difficulty. Their sensitivity did not reliably increase from pre- to post-training ($p = 0.26$), and the change was weaker than in the feedback group ($p = 0.04$). This interaction reflected a reliable difference in the change in false alarm rate between groups ($p = 0.007$).

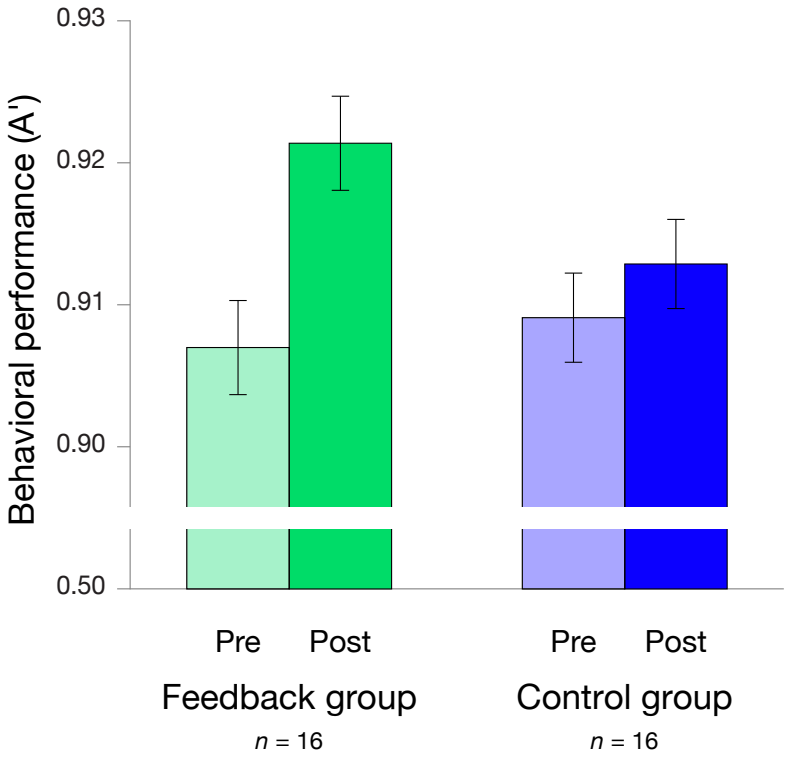

Figure 1.5: Change in behavior. Behavioral performance in the sustained attention task, as indexed by a non-parametric measure of sensitivity (A'), is plotted for the pre-training and post-training sessions. Participants who received accurate neurofeedback about their attentional state improved as a result of training, even though the feedback was no longer present in the post-training session. Control participants who received neurofeedback from other participants' brains did not improve. A reliable group difference in improvement shows that accurate feedback boosted performance above and beyond practice effects and stimulus exposure. Error bars represent ±1 within-subject s.e.m.

Rather than a benefit of accurate neurofeedback, the difference between groups could reflect a generic practice effect in the feedback group that was stymied by inaccurate neurofeedback in the control group. If so, then an improvement in sensitivity should be found even without feedback. We therefore ran a behavioral experiment with a new group of 16 participants who completed the same procedure but received only stable blocks. Unlike the

feedback group, their sensitivity did not increase from pre- to post-training ($p = 0.67$), inconsistent with this alternative account. In addition, accurate neurofeedback may have been effective simply because it resonated with a participant's attentional state and increased task engagement and motivation. If so, then feedback about any reliable measure of attention should be useful for training. We therefore ran a second behavioral experiment that was closely matched to the fMRI study, in which a new group of 16 participants received feedback based on RT (which was robustly related to attention; see Figure 1.6), along with a new group of 16 control participants who received yoked RT feedback. Unlike in the fMRI study, the change in sensitivity from pre- to post-training was not stronger in the feedback group than in the control group ($p = 0.29$), suggesting that not all correlates of attention are sufficient for training.

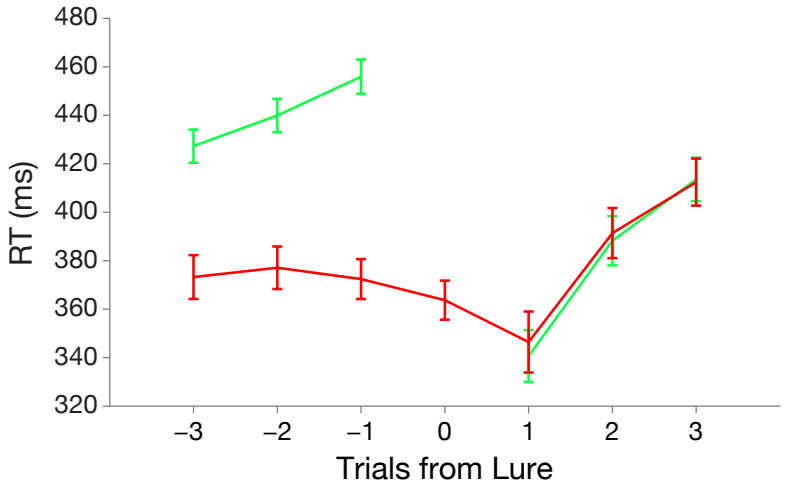

Figure 1.6: Average RTs surrounding lures. Green lines correspond to trials around correct rejections (CRs), where there was no behavioral response to the lure (presented at time = 0). Red lines correspond to trials around false alarms (FAs), where participants mistakenly responded to the lure. RTs were significantly slower prior to CRs than FAs (all timepoints, ps < 0.00001), consistent with the idea that FAs occurred when participants started responding habitually and were less attentive to the task. Error bars represent $\pm$ 1 s.e.m.

1.4 Training effects in the brain

One advantage of using whole-brain fMRI to provide feedback is that we could also gain insight into the neural changes induced by training. In particular, we hypothesized that learning via neurofeedback might strengthen and differentiate the two attentional states, such that they would become more discriminable in the brain from pre- to post-training. This might occur both in areas that represent attended stimulus features(Al-Aidroos et al., 2012; O'Craven et al., 1999) and in areas that represent task goals and control attention(Miller & Cohen, 2001; Noudoost et al., 2010). Although MVPA over the whole brain can be hard to interpret(Todd et al., 2013), this differentiation analysis tested for a very specific effect that would be hard to explain parsimoniously on the basis of generic confounds: namely,

regions that showed an improvement in classification as a result of training that was greater for the feedback relative to control groups.

We trained and tested classifiers on the first and last run of the rtfMRI training session (stable blocks) to distinguish attention to faces versus scenes and measured the change in cross-validation accuracy. Whole-brain classification (the basis for neurofeedback) showed a greater increase in accuracy from pre- to post-training in the feedback group than in the control group ($p = 0.01$). This interaction was present when the same analysis was performed separately in an anatomical mask of the frontal lobe ($p = 0.02$) and occipital lobe ($p = 0.04$), and it was trending in the temporal and parietal lobes ($p = 0.09$ and 0.08, respectively). Searchlight analyses further identified specific areas where activity patterns showed this interaction (Figure 1.7). The largest clusters were found in fusiform and parahippocampal gyri of ventral temporal cortex and in subcortical structures including the basal ganglia (striatum, globus pallidus) and amygdala (all $p < 0.05$, randomization test with threshold-free cluster correction).

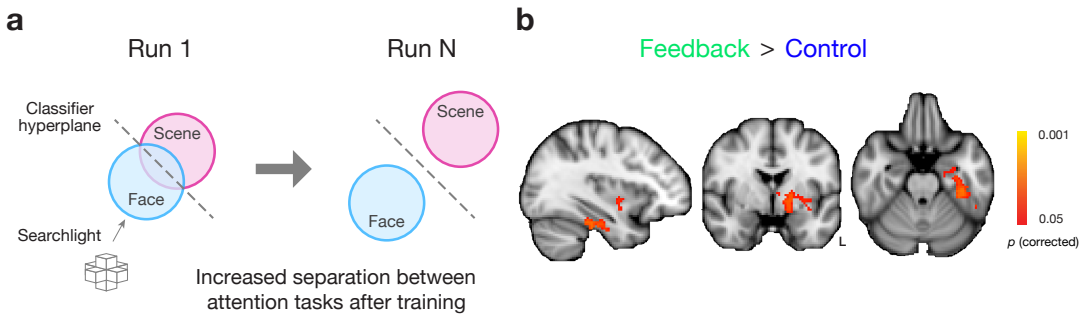

Figure 1.7: Searchlight analyses. (a) Voxel-wise analyses were conducted to identify brain regions whose surrounding activity patterns for the two attentional states became more separable after neurofeedback training. We computed cross-validation accuracy for classifiers trained to decode face and scene attention from RT-residualized BOLD data using a sphere with a 1-voxel radius centered on each voxel. Increased separability was quantified as the difference in accuracy between the end (run n) and start (run 1) of the fMRI session. (b) A greater increase in classifier accuracy for the feedback group relative to the control group ($p < 0.05$, randomization test with threshold-free cluster correction; Montreal Neurological Institute (MNI) x, y, z coordinates in mm) was observed in left ventral temporal cortex $(-34, -24, -25)$ and left basal ganglia $(-18, -4, -5)$. Small clusters (not shown) were obtained in left lateral temporal cortex $(-50, -45, -25; -51, -36, -20; -48, -42, -28)$ and left anterior temporal lobe $(-26, 22, -32)$.

1.5 Contribution of specific brain systems

Beyond investigating the consequences of training in the brain, the fMRI data can also be used to infer which brain regions were involved in the training process itself. Specifically, we examined which regions contributed to the whole-brain feedback and how these contributions affected behavioral training. This analysis consisted of three steps. First, we identified neural signals that could have been used to provide more targeted feedback from particular brain regions. Second, for each participant, we correlated these signals with whole-brain classifier output to quantify the extent to which the actual feedback that the participant received reflected information that was present within these regions. Third, we related these correlations to individual differences in the behavioral training effect to as-

sess which regions were most useful for training. In sum, if we had based the feedback on specific brain regions rather than the whole brain, how similar, and how useful, would this feedback have been?

Perhaps the simplest neural measure of when attention is allocated to faces versus scenes is the relative univariate activity of visual areas with selectivity for these categories, the fusiform face area (FFA) and the parahippocampal place area (PPA) (O'Craven et al., 1999), respectively. This is analogous to the approach used in several previous rtfMRI studies, in which the average activity from one or more regions of interest (ROIs) was returned as feedback (deCharms et al., 2005; Weiskopf et al., 2004; Yoo et al., 2012). The difference in univariate activity for task-relevant versus task-irrelevant ROIs (for example, PPA minus FFA for scene attention) was weakly but reliably correlated with the difference in whole-brain multivariate evidence for these categories over time (mean $r = 0.25$, s.e.m. $= 0.02$; $p < 0.00001$). That is, on average, 6% of variance in the whole-brain signal used for real-time feedback was explained by the relative activity levels of FFA and PPA. Individual differences in the size of the behavioral training effect were unrelated to this reliance on information in FFA versus PPA for feedback ($r = -0.04, p = 0.89$).

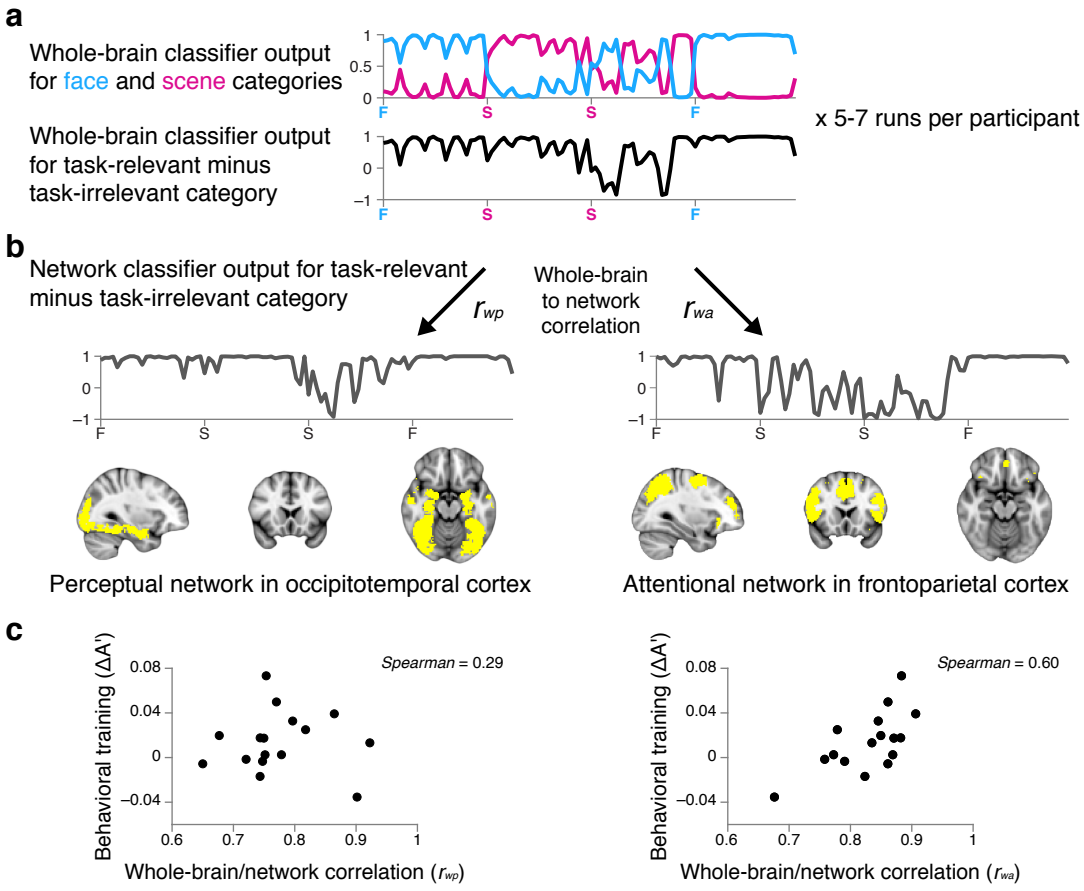

Figure 1.8: Potential sources of feedback. (a) Real-time whole-brain classifier output from the feedback blocks of a representative run for a single participant: evidence for each category (top) and evidence for the task-relevant minus task-irrelevant categories (bottom). (b) Offline classifier output for the same blocks from a perceptual network in occipitotemporal cortex (left) and an attentional network in frontoparietal cortex (right). The output from the whole-brain classifier was correlated with the outputs of the perceptual network classifier (rwp) and attentional network classifier (rwa) over time during the feedback blocks of each run. These correlations were averaged across runs within each participant to produce a measure of the extent to which the participant's real-time feedback relied on information in each network. (c) This measure of reliance on each network was in turn correlated with the change in behavioral A′ from pre- to post-training to assess whether feedback from each network was useful for training.

Information about visual categories is also represented outside peak category-selective areas, in patterns of activity distributed over regions of occipitotemporal cortex (Haxby

et al., 2001). Likewise, under attentionally demanding conditions, distributed activity patterns over frontoparietal regions linked to cognitive control contain stimulus-specific information(Woolgar et al., 2011). Therefore, classifiers applied to these 'perceptual' and 'attentional' networks of regions in occipitotemporal and frontoparietal cortices, respectively (constrained functionally, see Online Methods), provide additional neural measures of when attention was allocated to faces versus scenes (Figure 1.8). The difference in multivariate evidence for task-relevant versus task-irrelevant categories in the whole brain was reliably correlated with the difference in multivariate evidence for these categories in the perceptual network (mean $r = 0.77$, s.e.m. $= 0.02$; $p < 0.00001$) and attentional network (mean $r = 0.83$, s.e.m. $= 0.01$; $p < 0.00001$), although the correlation with the attentional network was significantly stronger ($p = 0.04$).

To assess whether the information contained in each network served as useful feedback for training, we related individual differences in these whole-brain/network correlations to the size of the training effect across participants. Behavioral improvement was unrelated to reliance on the perceptual network ($r = 0.29, p = 0.27$) but positively related to reliance on the attentional network ($r = 0.60, p = 0.02$).

1.6 Discussion

We found that moment-to-moment feedback about attentional state could enhance sustained attention abilities. We used closed-loop neurofeedback from MVPA as a type of cognitive prosthetic, facilitating participants' ability to detect neural signals that indicated an impending attentional lapse by displaying them visually in a form that was directly relevant to the task. In other words, we provided a neural error signal so that participants could

learn to better monitor and evaluate the state of their attention.

MVPA has become widespread because of its ability to read out the informational contents of the brain (Norman et al., 2006; Turk-Browne, 2013). However, classifiers exploit any predictive variance that distinguishes between classes, and they are thus susceptible to confounding factors (Todd et al., 2013). Our design allowed us to assess whether whole-brain classifier output truly reflected attentional state — the cognitive variable of interest — using behavior as the yardstick: when provided as feedback, classifier output was useful for improving attention-dependent performance.

By using fMRI for cognitive training, we gained important insights about the underlying neural mechanisms. We first identified brain regions that were affected by training, including frontal cortex, ventral temporal cortex and basal ganglia (striatum and globus pallidus), which came to represent the attentional states more distinctively as a result of feedback. We interpret the increased neural separation in these regions as reflective of the two component processes in our sustained attention task. First, participants needed to select the image from the task-relevant category when confronted with a composite stimulus. Increased neural separation of face and scene attention in frontal cortex may reflect learning of better task or control representations for each category, which in turn enabled stronger top-down modulation of category-selective visual representations in ventral temporal cortex, biasing processing toward the task-relevant image and thereby facilitating its selection (Al-Aidroos et al., 2012; Baldauf & Desimone, 2014; Miller & Cohen, 2001; Noudoost et al., 2010; Reddy et al., 2009). Second, participants needed to inhibit their prepotent response when the selected image came from the infrequent lure subcategory. Increased neural separation of face and scene attention in basal ganglia may reflect enhanced learning of different stimulus-

response rules within each category, with the striatum directly gating responses to targets and indirectly blocking responses to lures via inhibitory projections to the globus pallidus (Kravitz et al., 2010; Miller & Cohen, 2001; O'Reilly & Frank, 2006; Wiecki & Frank, 2013).

We next identified brain regions supporting the training process itself, simulating how feedback from these regions related to the real-time feedback. Univariate activity in the FFA and PPA was weakly correlated with whole-brain multivariate evidence, whereas multivariate evidence from a perceptual network in occipitotemporal cortex and multivariate evidence from an attentional network in frontoparietal cortex were strongly correlated. These results suggest that distributed activity patterns contributed more to the training effect than punctate responses in category-selective visual areas. We tested this more directly by exploiting variance in the training effect across participants. Training was predicted by the extent to which feedback relied on information in the attentional network, but not by reliance on information in the perceptual network or the relative activity levels of FFA and PPA. The importance of frontoparietal feedback can be interpreted as evidence that our attention training involved the reinforcement of task representations(Miller & Cohen, 2001; Noudoost et al., 2010), rewarding good states by reducing difficulty and punishing bad states by increasing difficulty. This interpretation is consistent with the observed multivariate changes in frontal cortex and basal ganglia, as these regions and their interaction are critical for feedback-based task learning (Foerde & Shohamy, 2011; Frank & Badre, 2012). It remains an open question whether feedback restricted to occipitotemporal cortex or to FFA and PPA (that is, without access to frontoparietal information) could lead to training.

The goal of our study was to derive basic science insights into the neural basis of top-down attention and the plasticity of attention-related behavior. In the future, the kind of

approach we pursued may find potential applications for training sustained attention in occupational settings (for example, baggage screeners and truck drivers) (Biggs & Mitroff, 2014) and clinical disorders (for example, attention deficit hyperactivity disorder and negative attentional biases in depression)(Disner et al., 2011; Shallice et al., 2002).

1.7 Methods

1.7.1 Participants

Eighty adults (45 female, 75 right-handed, mean age = 20.3 years) participated in the study for monetary compensation. This included 16 participants in each of the following: the fMRI experimental group, the fMRI control group, the no-feedback behavioral group, the RT-feedback behavioral group and the RT-control behavioral group. Each participant in the fMRI control and behavioral groups was matched as closely as possible to the demographics (age, gender and handedness) of a participant in the fMRI experimental group. Power analyses were not performed because of the use of a new paradigm and unknown behavioral and neural effect sizes. The sample size was chosen because it is fairly common for an fMRI study, especially one with multiple groups and sessions. Three additional fMRI participants were excluded because of technical problems with real-time data acquisition, one additional fMRI participant was excluded for falling asleep during several runs and one additional behavioral participant was excluded for low overall performance (3.2 s.d. below the mean in pre-training). For the fMRI participants, the experimenter was not blind to group assignment because of the complexity of data acquisition and analysis, especially the need to ensure that the real-time classification and feedback system was functioning. However, every fMRI participant received the same scripted instructions. All participants had

normal or corrected-to-normal visual acuity and provided informed consent to a protocol approved by the Princeton University Institutional Review Board.

1.7.2 Stimuli

Images consisted of grayscale photographs of male and female faces and indoor and outdoor scenes. These images were combined into composite stimuli by averaging pixel intensities using various weightings (for example, 60% face, 40% scene). The stimuli were displayed on a projection screen at the back of the scanner bore (subtending 10 □ 10° of visual angle) and viewed with a mirror attached to the head coil.

A fixation dot was superimposed on the images and presented during the inter-block intervals of each run, except when text instructions were displayed. Participants were instructed to fixate on this dot, they received practice doing so during their first session, and they were reminded about the importance of fixation before scanning. We did not use an eye tracker to ensure fixation because of the technical complexity of the real-time apparatus and analysis. Aside from this, eye tracking is rarely used in attention studies with overlapping face/scene stimuli(Al-Aidroos et al., 2012; Chadick & Gazzaley, 2011; Moore et al., 2013), and when it has been used, no differences in eye movements or position across categories were observed(Serences et al., 2004). In fact, only one participant reported using an eye-movement strategy to perform the task, and this participant showed the smallest training effect of anybody in the feedback group.

1.7.3 Procedure

Participants completed three sessions on different days. The first day was a behavioral pretraining session with two runs of the sustained attention task. The second day was an fMRI

session with several runs of the modified real-time neurofeedback version of the sustained attention task. The number of runs varied across participants, depending on how many they could complete within 2 h (range 6–9 runs). The third day was a behavioral post-training session, otherwise identical to the first session. We attempted to conduct the sessions on three consecutive days, but this was not always possible because of scanner availability and participants' schedules. All participants completed the study within 5 d. The average number of days (and s.e.m.) between the first and second sessions was 1.19(0.09) and between the second and third sessions was 1.25(0.09).

Each task run contained eight blocks. Each block began with a text cue for 1 s that instructed participants which subcategory was the target to which they should respond and, by extension, which category was to be attended. Four of the blocks involved attending to faces and the other four involved attending to scenes. The target subcategories were held constant within each participant (for example, "male" and "indoor") but were counterbalanced across participants. The cue was followed by 1 s of fixation and then a series of 50 trials. Each trial contained a composite face/scene image presented for 1 s with no interstimulus interval. Responses were recorded during the first 850 ms of stimulus presentation to allow computation time at the end of the trial. The trial structure followed a response inhibition task design(Robertson et al., 1997; Rosenberg et al., 2013): 90% of images contained the target subcategory (for example, an indoor scene after an 'indoor' cue) and required a response; the other 10% contained the non-target subcategory (in this case, an outdoor scene) to which responses needed to be withheld. The distribution of the subcategories was the same for the unattended category (for example, 90% male and 10% female after an 'indoor' cue), although these images were irrelevant for determining whether to

respond or not. After the last trial, there was a 4–6 s of fixation before the next block.

The first run of the rtfMRI session was identical to the runs of the behavioral sessions, with all composite stimuli being an equal mixture of face (50%) and scene (50%) images. Starting with the second run, the first four, 'stable' blocks kept the same equal mixture, but the final four, 'feedback' blocks had variable mixture proportions that depended on the participant's attentional state. Text instructions appeared before the first feedback block to alert participants that neurofeedback was starting. Each of these blocks started with an equal mixture for the first three trials. The mixture proportions for the remaining trials were determined on the basis of real-time MVPA of the fMRI data. They ranged from 17% to 98% of the task-relevant category (83% to 2% of the task-irrelevant category). For half of the participants, the last run of the rtfMRI session was identical to the first run, with all eight blocks using an equal mixture.

1.7.4 Data acquisition

Experiments were run using the Psychophysics Toolbox for Matlab (http://psychtoolbox.org/). Neuroimaging data were acquired with a 3 T MRI scanner (Siemens Skyra) using a 16-channel head coil. We first collected a scout anatomical scan to align axial functional slices to the anterior commissure-posterior commissure line. Functional images were acquired using a gradient-echo, echo-planar imaging sequence (2 s repetition time, 28 ms echo time, 3 mm isotropic voxel size, 64×64 matrix, 192 mm field of view, 36 slices) that covered most of the brain. At the end of the fMRI session, a high-resolution magnetization-prepared rapid acquisition gradient-echo (MPRAGE) anatomical scan was acquired for offline spatial registration. To improve registration, an additional coplanar T1 fast low angle shot anatomical

scan was also acquired.

1.7.5 Statistics

Because some of the data violated the assumption of normality needed for parametric tests, we used nonparametric tests throughout to determine statistical significance. Subject-level bootstrap resampling(Efron & Tibshirani, 1986) was used to assess random-effects reliability for comparisons of a small number of variables to chance or each other; one-sided tests were used for directional hypotheses and two-sided tests for nondirectional hypotheses. Correlations between two variables were estimated with Spearman's rank correlation after applying robust methods to eliminate the disproportionate influence of outliers in small samples (Pernet et al., 2013). Significance testing on voxel-wise brain maps was conducted with a permutation test in FSL's 'randomise' function (Nichols & Holmes, 2002) and corrected for multiple comparisons using threshold-free cluster enhancement (Smith & Nichols, 2009). Each control participant was matched to one experimental participant in all respects except for the key manipulation (that is, on demographics, stimuli and number of runs), and so comparisons across groups were performed with a matched-pairs sample design.

1.7.6 Real-time analyses

Preprocessing

During the fMRI session, data were reconstructed on the scanner. Prospective acquisition correction and retrospective motion correction were applied. Each motion-corrected volume was transferred to a separate analysis computer in real time. A brain mask was applied

to eliminate non-brain voxels. The volume was spatially smoothed in Matlab using a Gaussian kernel with full-width half-maximum (FWHM) = 5 mm. After each grouping of four stable blocks, the BOLD activity of every voxel was z-scored over time. The same normalization was applied during feedback blocks in real time, using the mean and standard deviation from the most recent four stable blocks.

Multivariate pattern analysis

During the fMRI session, we conducted MVPA using penalized logistic regression with L2-norm regularization (penalty = 1). The classifier was trained to distinguish top-down attention to faces and scenes from whole-brain activity patterns. The training examples for the classifier were obtained from a trailing window of stable blocks. For half of the participants, this trailing window included the twelve previous stable blocks and the classifier was trained during a 70-s fixation period between blocks 4 and 5 of the current run. For the other half, the trailing window did not include the stable blocks from the current run and the classifier was trained between runs; the fixation period between blocks 4 and 5 was reduced to 6 s. There were no reliable differences between these groups, and so they were analyzed together. For training the model, all regressors were shifted 4 s forward in time to adjust for the hemodynamic lag.

The trained model was tested in real time on brain volumes obtained during the feedback blocks. For each volume, the classifier estimated the extent to which the brain activity pattern matched the pattern for the two attentional states on which it was trained (from 0 to 1). The neurofeedback was based on the difference of classifier outputs for the task-relevant category minus task-irrelevant category. These outputs are perfectly anticorre-

lated in a two-class classifier, such that the difference ranged from −1 to 1. As a result of the anticorrelation, it is difficult to disentangle less attention to the task-relevant category from more attention to the task-irrelevant category (and vice versa). Differences of −1 and 1 should thus be interpreted in relative terms as more attention to the task-irrelevant and task-relevant categories, respectively. Note that if participants were not in either attentional state, the brain activity patterns would contain no signal that the classifier could identify and the classification would be driven by noise; the difference would then be 0 on average. We therefore interpret positive and negative values away from 0 as evidence of selective attention to one category of the composite stimulus.

Neurofeedback

The output of the classifier was used to determine the proportion of the images from the task-relevant and task-irrelevant categories in the composite stimulus on the next trial. The preprocessing and decoding of volume i were performed during volume $i + 1$ and the classifier output was used to update the stimulus mixture for the two trials in volume $i + 2$. This resulted in a minimum lag of 2 s (two trials) between data acquisition and feedback. Moreover, classifier output was averaged over a moving window of the preceding three volumes ($i - 2$, $i - 1$ and i for feedback in volume $i + 2$), meaning that feedback was based on brain states 2 − 8 s in the past. Because sustained attention fluctuates slowly (Rosenberg et al., 2013), we reasoned that this smoothing would provide a better estimate of attentional state by reducing high-frequency noise.

The average classifier output was mapped to a proportion of the task-relevant category using a sigmoidal transfer function (Figures 1.4 and 1.9). The maximum output value

(0.98) meant that the task-irrelevant image could be almost completely absent from the composite stimulus. The minimum output value (0.17) meant that the task-relevant image was always present to some degree, even when participants had lapsed completely (giving them a foothold to recover). The inflection point of the logistic function was centered at an input greater than chance (0.60), based on the mean decoding accuracy of a group of pilot participants. This placed the sensitive range of the feedback closer to the typical real-time classifier output values. For follow-up analyses that assessed how the feedback changed over training, we computed the average proportion of the image from the task-relevant category in each training run and calculated the linear slope across runs for each participant.

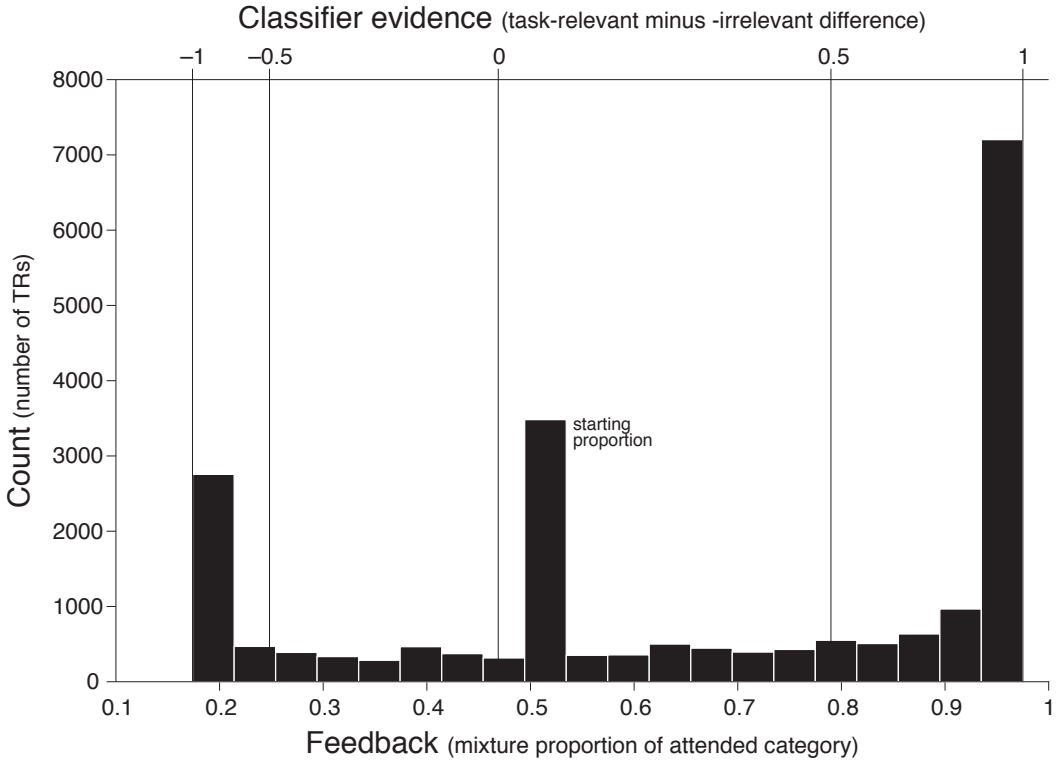

Figure 1.9: Histogram of feedback. The bottom x-axis refers to proportion of the image from the task-relevant category in each composite stimulus. The y-axis refers to the number of TRs that contained stimuli with this mixture (in bins of width = 0.04) across all feedback blocks from all participants in the feedback group. The top x-axis depicts the correspondence between classifier output and feedback (computed using the transfer function in Figure 1.4). The most frequent values were in the highest and lowest bins, as well as in the bin including feedback of 0.5. The highest bin reflects cases in which the image from the task-relevant category was 98% of the composite stimulus. The lowest bin reflects cases in which the image from the task-irrelevant category was 83% of the composite stimulus. The bin with 0.5 was frequent because every block began with an equal mixture of the images from the task-relevant and task-irrelevant categories.

Participants were aware of the feedback manipulation. Before the fMRI session, they were told that the images in the second half of most runs would change depending upon their attention, as measured from their brain. Specifically, the task would get easier if they were paying attention and it would get harder if they became inattentive. They were shown examples of how a composite stimulus could change on the basis of whether they were do-

ing a good or bad job of paying attention. Critically, control participants received the exact same instructions. After the study, participants completed a debriefing questionnaire, which included the question: "Did you feel that you could control the image with your brain?" Overall, 11 of 16 participants in the feedback group reported feeling some degree of brain control, compared to 4 of 16 participants in the control group. Interestingly, the feedback that the control participants received was positively correlated on average with what they would have received on the basis of their own brain activity patterns (mean $r = 0.29$, s.e.m. $= 0.04, p < 0.00001$). This correlation suggests that their attentional state was affected by the sham feedback, which in turn determined what feedback they should have received next. Unlike the feedback participants, however, the control participants were only reacting to the feedback and not driving it.

1.7.7 Offline analyses

General procedures.

Using FSL (http://fsl.fmrib.ox.ac.uk/), the data were temporally high-pass filtered (200 s period cut-off), motion corrected again, and spatially smoothed with a Gaussian kernel (5 mm FWHM). They were then transformed into standard Montreal Neurological Institute (MNI) space by linearly registering to the MPRAGE images and to the MNI152 standard brain. We conducted offline MVPA using the Princeton Multi-Voxel Pattern Analysis Toolbox (http://www.pni.princeton.edu/mvpa/), with z-scoring over time within each run and the same type of classifier as in the real-time analyses (penalized logistic regression using L2-norm regularization, penalty = 1).

Decoding accuracy

We assessed our ability to decode attentional state within individual participants by classifying the stable blocks, which were uncontaminated by stimulus-based feedback. (In fact, these data served as the training set for real-time classification, but were never subdivided into training and test sets so that classifier accuracy could be estimated with cross-validation.) We trained a classifier using the stable blocks from $n-1$ runs and tested it on the left-out run, then repeated n times. By averaging over these folds, we obtained a measure of how well we could decode the attentional state of each participant and assessed reliability in the group relative to chance (0.5). We interpreted this decoding accuracy as reflecting the neural separability of attentional states rather than the precision with which the classifier algorithm captured these states per se. That is, low decoding accuracy for a participant does not necessarily mean that his or her classifier itself was inaccurate, but rather that it was accurately tracking poor neural separation between attentional states. Such separation may be related to individual differences in attentional abilities, with poor separation reflecting weaker selection of task-relevant information and/or increased distraction by task-irrelevant information. To verify this interpretation, we correlated decoding accuracy across participants with behavioral sensitivity from the pre-training session. We used A′ to index sensitivity because of its robustness to the high hit rates that we expected to obtain because of the greater frequency of targets than lures (Donaldson, 1992).

Predicting behavioral accuracy

For classifier output to provide useful feedback for training purposes, (1) it should be related to behavior on a trial-by-trial basis within participant and (2) this relationship should

hold without artificially shifting trials back in time to correct for the hemodynamic lag (which cannot be done in real time). To judge whether these criteria were satisfied, we examined whether the classifier output before a lure trial (averaged over the three preceding volumes, as used to calculate feedback) predicted whether participants correctly withheld their response or incorrectly responded. This relationship was tested with a logistic regression (correct rejection = 1, false alarm = 0), whose slope was reliably positive at the group level (mean slope $=$ 0.67, s.e.m. $=$ 0.11, $p < 0.00001$). That is, more classifier evidence from volumes 2-8 s in the past — most influenced by neural events 6-12 s in the past, assuming a hemodynamic peak at 4 s — predicted behavioral accuracy on the current trial.

The average RT from the six trials during these volumes also predicted behavioral accuracy in a logistic regression (mean slope $=$ 0.01, s.e.m. $=$ 0.0009, $p < 0.00001$; Figure 1.6). To remove this confound, we averaged the two RTs from each volume, regressed this average out of the raw classifier output and behavioral accuracy across volumes, and then repeated the analysis above in the residuals (using partial correlation rather than logistic regression because behavioral accuracy was no longer binary). The positive relationship between classifier output and behavioral accuracy remained reliable (mean r $=$ 0.06, s.e.m. $=$ 0.01, $p < 0.00001$; Figure 1.10).

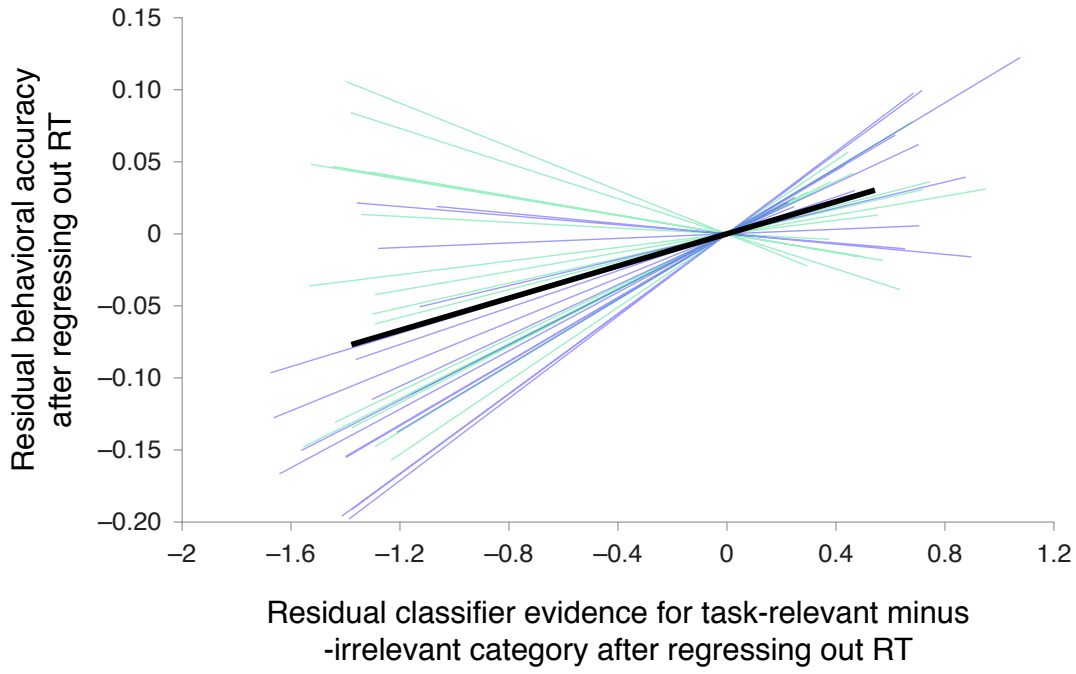

Figure 1.10: Removing the influence of RT. Both RT and classifier evidence for the task-relevant vs. task-irrelevant category were higher preceding CRs. To verify that the classifier was predictive of behavioral accuracy and not merely correlated with RT, we regressed RT out of classifier output and behavioral accuracy across trials and performed a partial correlation analysis. Green lines correspond to regression fits for feedback participants, and blue lines correspond to regression fits for control participants. The black line is the average fit. The relationship between classifier output and behavioral accuracy remained robust ($p < 0.00001$).

1.7.8 Changes in neural discriminability

We performed several analyses to examine whether attention training increased the separation between neural representations of the face and scene attentional states. We operationalized neural separation with decoding accuracy, comparing the first and last runs of the rtfMRI training session to assess training-induced changes. For each of these runs, we trained a classifier to decode attentional state from the stable blocks using a split-half cross-

validation procedure. To ensure that classification was not confounded by RT, we averaged the two RTs in every volume of the stable blocks and regressed out the resulting RT time course from the brain data before analysis27. We calculated the difference in decoding accuracy as the last minus first run for each participant and assessed the reliability of this change at the group level, comparing feedback and control groups. This analysis was performed over the whole brain, within each of the four lobes (defined using the MNI atlas in FSL), and over spherical searchlights (1 voxel radius, 7 voxel maximum volume) centered on every voxel in the brain.

Simulated feedback

For the univariate analysis of FFA and PPA, we localized these areas within each participant by contrasting face and scene attention blocks in the first fMRI run (which had no feedback). FFA and PPA ROIs were defined as 5-mm spheres around the peak face- and scene-selective voxels in right and left lateral fusiform gyri and collateral sulci/parahippocampal gyri, respectively. We then collapsed over hemispheres and averaged the time series of the voxels in each bilateral ROI for all feedback blocks (starting in the second fMRI run). For each block, we calculated the difference over time between the task-relevant ROI (FFA and PPA for face and scene attention, respectively) and the task-irrelevant ROI (PPA and FFA for face and scene attention, respectively) and then concatenated these differences across blocks within each run. To estimate how FFA and PPA activity related to the real-time feedback in each run, we correlated this ROI-derived time series with the difference in whole-brain classifier output for the task-relevant minus task-irrelevant categories (face minus scene evidence for face attention and scene minus face evidence for scene attention)

from the same blocks and then averaged across runs. These FFA/PPA to whole-brain correlations were in turn correlated across feedback participants with the improvement in A' from pre- to post-training.

For the multivariate analysis of perceptual and attentional networks, we first defined each network using functional and anatomical criteria. The functional criterion was based on forward-inference meta-analyses from http://neurosynth.org/ using the search terms "faces" OR "scenes" for the perceptual network and "attention" for the attentional network. The results were downloaded as voxel-wise masks and thresholded at $z = 2.3$. The anatomical criterion was based on the MNI atlas in FSL, which was used to generate binary masks of the occipital and temporal lobes for the perceptual network and the frontal and parietal lobes for the attentional network. Voxels that survived the intersection of the functional and anatomical masks were used to train a separate classifier for each network. Other than being performed offline rather than in real time, the rest of the classification procedure was identical to that of the whole-brain classifier, with training for each run based on a moving window of stable blocks and testing occurring volume by volume in the feedback blocks. The classifier output for each network was correlated with the whole-brain classifier output to estimate how much information in that network contributed to the real-time feedback in every run, and these correlations were averaged across runs within each participant. These network/whole-brain correlations were then correlated with the improvement in A' from pre- to post-training across feedback participants to assess the usefulness of relying on information in the perceptual and attentional networks for training.

1.7.9 Behavioral control experiments

We recruited three behavioral control participants for each fMRI participant from the feedback group of the main study (total $n = 48$). They were all demographically matched to the fMRI participant in handedness, gender and age (± 1 year). In addition, they received the same experimental design, in terms of stimulus order, block order, number of training runs, number of sessions and number of days between sessions. The training session was conducted in a behavioral testing room rather than the scanner. To emulate the contextual change experienced by the fMRI participants in switching environments between pre-/post-training and training, the pre- and post-training sessions were run in a different room in a different part of our building. Across the three groups of participants, we manipulated the nature of the feedback that they received during training. The assignment of control participants to each of these feedback conditions was randomized.

The no-feedback experiment ($n = 16$) was identical to the fMRI study, other than being conducted outside the scanner and replacing all feedback blocks with stable blocks. That is, the stimulus mixture proportion remained constant at 50% for both categories during all blocks. As a result of removing the feedback, participants also did not receive instructions about how to interpret varying stimulus proportions and there was no separate, yoked control group.

The RT-feedback experiment was identical to the fMRI study, other than being conducted outside the scanner and having the feedback controlled by RT rather than whole-brain classifier output. This experiment contained two between-subject conditions, the RT-feedback group ($n = 16$) and the RT-control group ($n = 16$). Participants were assigned in matched pairs, with the RT-control participant in each pair receiving feedback

yoked to that generated by the matched RT-feedback participant. Thus, by definition, the RT-feedback participant was run before their match. However, the RT-feedback and RT-control participants from different pairs were interleaved and run in the same cohort. Participants in both groups received the same instructions, which were slightly modified from the fMRI study to remove scanner-related references. Not only were participants blind to their condition, but the experimenter was also blind (other than to the first and last participants, who were necessarily RT-feedback and RT-control participants, respectively). A different researcher conducted participant recruitment and scheduling, resulting in a double-blind procedure. (Note that because of the lack of feedback in the no-feedback experiment, the experimenter was aware of whether a participant ended up in that particular group.)

The feedback regime for the RT-feedback condition was very similar to the fMRI experiment, using RT (instead of fMRI category evidence) as a measure of attentional state. During feedback blocks, the proportion of the task-relevant category increased when participants responded slowly and decreased when they responded quickly. We chose this mapping because RT was faster on trials preceding false alarms than correct rejections (Figure 1.6), consistent with habitual responding and worse sustained attention (Robertson et al., 1997). More specifically, the feedback was based on the participant's deviation from their average RT, calculated over a trailing window of stable blocks (the same set as used for training data in the fMRI classifier). The stimulus mixture proportion was adjusted using the average of the RTs from the previous six trials, which was equivalent to the three brain volumes that were used for feedback in the fMRI version. A sigmoidal transfer function transformed this value into a stimulus mixture proportion.

2

Forgetting from lapses of sustained attention*

When performing any task for an extended period of time, attention fluctuates between good and bad states. These fluctuations affect performance in the moment, but may also have lasting consequences for what gets encoded into memory. Experiment 1 establishes this relationship between attentional states and memory, by showing that subsequent memory for an item can be predicted by a response time index of sustained attention before the item even appears. Combined with this temporal predictiveness, Experiment 2 strength-

*The contents of this section were presented at the Vision Sciences Society annual meeting with coauthors K.A. Norman and N.B. Turk-Browne

ens the causal interpretation of the relationship by treating the sustained attention index as an independent variable to trigger the appearance of encoding trials, with better memory when items are presented during good attentional states. Together, these findings suggest that sustained attention can have downstream consequences for what we remember, and they highlight the inferential utility of adaptive experimental design. By continuously monitoring attention, we can influence which information will be later remembered.

2.1 Introduction

It is often necessary to sustain attention for long periods of time, such as when listening to a lecture, driving a car, or eating a meal with loved ones. Despite this ubiquitous need, and the potential educational, health, and social consequences of losing focus, humans are notoriously bad at sustaining attention. Such lapses occur frequently in a wide range of laboratory studies, such as continuous performance and response inhibition tasks (deBettencourt et al., 2015; Robertson et al., 1997) and low-prevalence visual search (Fleck & Mitroff, 2007; Wolfe et al., 2005). Indeed, sustained attention is characterized by continuous fluctuations between good and bad attentional states (Esterman et al., 2013; Rosenberg et al., 2013). The purpose of this study was to investigate the longer-term consequences of such fluctuations for memory.

Attention and memory are intricately related in general (see Aly & Turk-Browne, in press). However, attention is a broad term that encompasses a wide range of tasks (see Chun et al., 2011), and the relationship between attention and memory has been primarily examined for two varieties of attention. First, studies of divided attention have shown that performing a secondary task during encoding and/or retrieval impairs recognition mem-

ory (Craik et al., 1996). Second, studies of selective attention have shown that memory is enhanced for target features/locations and suppressed for distractors (e.g., Moray, 1959; Turk-Browne, 2013; Uncapher et al., 2011; Yi & Chun, 2005).

Here we explore the relationship between another attentional construct – sustained attention – and subsequent recognition memory. Compared to divided and selective attention, there has been relatively less work on the mnemonic consequences of sustained attention. One study found that mindfulness correlated with both sustained attention in a response inhibition task and memory for distractor scenes presented in the background, which was used to argue that performance in the task did not depend on selective attention (Rosenberg et al., 2013). Another study using a related task found that memory on lure trials in which responses need to be inhibited was suppressed relative to items on target trials (Chiu & Egner, 2015b).

Here we aim to more directly link sustained attention to memory encoding. That is, rather than studying this relationship in terms of how individual differences correlate over participants (Rosenberg et al., 2013) or how control demands influence memory for information encountered in the moment (Chiu & Egner, 2015b), we test the predictive hypothesis that the state of sustained attention leading into a trial determines the mnemonic fate of the content of that trial. We operationalize the state of sustained attention using a moving window of the response times (RTs) on preceding trials. This is based on response inhibition tasks in which faster RTs are associated with an increased error rate on upcoming lure trials, indicating habitual responding and a lapse of sustained attention (deBettencourt et al., 2015; Robertson et al., 1997). In Experiment 1, we show that this measure of attention predicts whether upcoming information will be encoded successfully. In Experiment 2, we

seek to reinforce the causal nature of this relationship using a real-time design in which the attentional measure is treated as an independent variable for triggering encoding opportunities.

2.2 Experiment 1

The goal of this experiment is to test whether the state of sustained attentional going into a lure trial predicts memory for that trial. Specifically, we hypothesize that faster preceding RTs will be associated with worse encoding of upcoming information.

2.2.1 Participants

Thirty-two undergraduates (22 female, 10 male; mean age = 20.45 years) from Princeton University participated in the experiment for course credit or a $10 payment. This sample size was chosen to be twice as large as our previous studies of sustained attention (deBettencourt et al., 2015), because the effects on memory were anticipated to be smaller. One additional participant was excluded for task performance more than 3 SDs below the mean (error rate on infrequent trials = 0.88). Participants reported normal or corrected-to-normal visual acuity and color vision, and they provided informed consent to a protocol approved by the Princeton University Institutional Review Board (IRB).

2.2.2 Stimuli

Color scene images (550 indoor and 550 outdoor) were selected from the SUN database (Xiao et al., 2010). They subtended approximately 7° in the center of a gray background, with a black central fixation dot (0.1°) superimposed. The fixation dot turned white after each button response.

2.2.3 Apparatus

Participants were seated approximately 70 cm from a CRT monitor (refresh rate = 100 Hz). Stimuli were presented using MATLAB (The MathWorks, Natick, MA) and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

2.2.4 Procedure

The design consisted of two phases: sustained attention and surprise memory test. In the sustained attention phase (Figure 2.1a), each participant viewed 500 unique images for 1000 ms each with no interstimulus interval. Overall, 90% of the images (450) were from one "frequent" category (e.g., outdoor) and 10% (50) were from the other "infrequent" category (e.g., indoor). The frequent category was counterbalanced across participants. Participants were instructed to press the "h" button with their right index finger for images from the frequent category and "j" with their right middle finger for images from the infrequent category. They first completed a practice block with novel images, until they achieved 80% accuracy.

In the surprise memory test phase (Figure 2.1b), each participant viewed 200 unique images (100 each from the frequent and infrequent categories). These images were equally likely to have appeared in the sustained attention task ("old") or not ("new"). All 50 old images from the infrequent category were included, along with 50 of the 450 old images from the frequent category (10 images per 100 trials to balance onset time). Participants were instructed to press number buttons with their right index finger to indicate their confidence that the image had been presented previously ("1" = high-confidence old, "2" = low-confidence new, "3" = low-confidence old, and "4" = high-confidence old). A response key

was presented below each image. The images and response key remained on the screen until response, after which the response was also shown for 500 ms, followed by a blank 500-ms interstimulus interval with a fixation dot. Participants were instructed to balance their responses across the buttons.

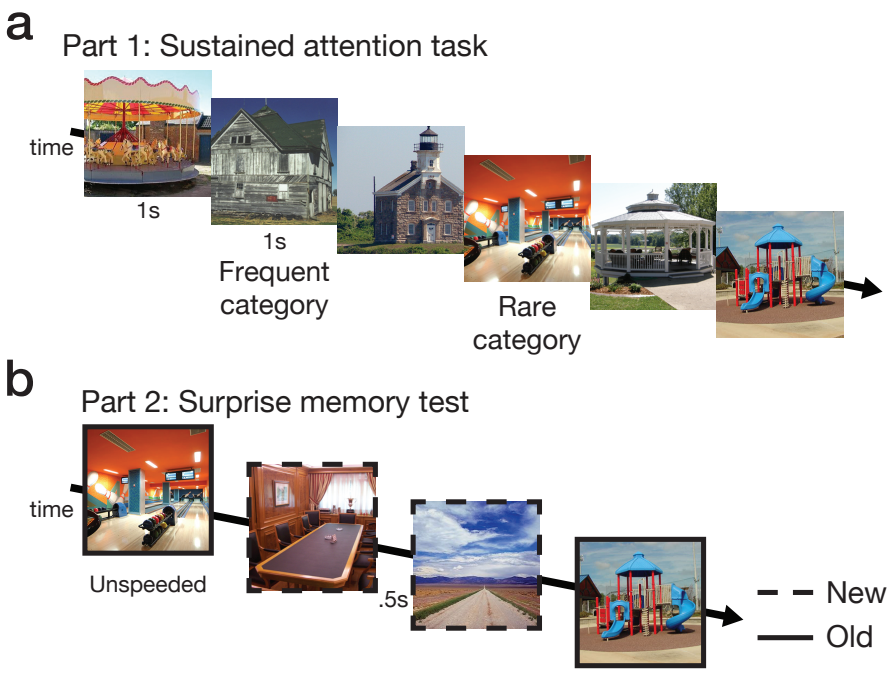

Figure 2.1: Experimental design. (a) Participants first completed a sustained attention task in which they viewed trial-unique scene images and made an indoor/outdoor judgment. 90% of the scenes were from one of these categories (e.g., outdoor) and 10% were from the other category (e.g., indoor). Because of this imbalance responding correctly to the infrequent category required inhibiting the prepotent response to the frequent category. (b) Participants then completed a surprise memory test in which they reported their confidence that each scene had appeared in the first part of the experiment. 50% of the images were from the sustained attention task (old) and 50% were novel to the experiment (new). Among these, 50% were from the frequent category and 50% were from the infrequent category.

2.2.5 Analysis

The state of sustained attention going into a trial (i) was operationalized as the average RT over the three preceding trials ($i-3$, $i-2$, $i-1$). Before relating this measure to memory on a trial-by-trial basis, we removed the linear drift in RTs over the course of the sustained attention task to control for generic time-dependent effects (e.g., practice, fatigue, etc.) that could otherwise complicate comparisons between trials from different points in the experiment. For the subsequent memory analysis, high-confidence old responses were treated as remembered and all other responses were treated as forgotten (as is standard; e.g., Kim et al., 2014; Wagner et al., 1998). Across infrequent items within participant, logistic regression was used to predict this binary memory variable at test from the normalized RT index of sustained attention at encoding.

All statistics were computed using a nonparametric random-effects approach in which participants were resampled with replacement 10,000 times (Efron & Tibshirani, 1986). Because this generates a sampling distribution of the true mean, null hypothesis testing was performed by calculating the proportion of the iterations in which the bootstrapped mean was in the opposite direction from the true effect. Unless stated otherwise, the true mean and 95% confidence interval of the bootstrapped distribution are reported as descriptive statistics.

2.2.6 Results

During the sustained attention task, the overall sensitivity of the category judgments was above chance (mean $A' = 0.92$, 95% CI = 0.90–0.93; vs. chance $= 0.5$, $p < 0.00001$). The error rate was higher for trials from the infrequent category (0.30, 0.26–0.34) vs. fre-

quent category (0.02, 0.02–0.03; $p < 0.00001$), reflecting the need to inhibit a prepotent response on these trials. Furthermore, our RT measure of sustained attention – average RT over preceding three trials – was predictive of the success of response inhibition on infrequent trials: responses were slower before a correct response (519 ms, 502–536) than an incorrect response (444 ms, 420–468; $p < 0.00001$). The timecourse of this effect over preceding timepoints is depicted in Figure 2.2a.

During the surprise memory test, sensitivity was above chance for items from both the frequent category (mean $A' = 0.66$, 0.62–0.70; $p < 0.00001$) and the infrequent category (0.81, 0.79–0.83; $p < 0.00001$), although sensitivity was higher for the infrequent category ($p < 0.00001$). Among the old infrequent items, the basis of the subsequent memory analysis, 38% (33–44) were classified as remembered (hits) and 62% (56–67) were classified as forgotten (misses). The critical test of our hypothesis concerned the relationship between this measure of memory at test and the measure of sustained attentional state at encoding. As illustrated in Figure 2.2b, this was evaluated with logistic regression across items within participant. The resulting slopes describing this relationship were reliably positive across participants (mean $\beta = 3.43$, $X.XX$–$X.XX$; $p = 0.039$). The distribution of these slopes is shown in Figure 2.2c.

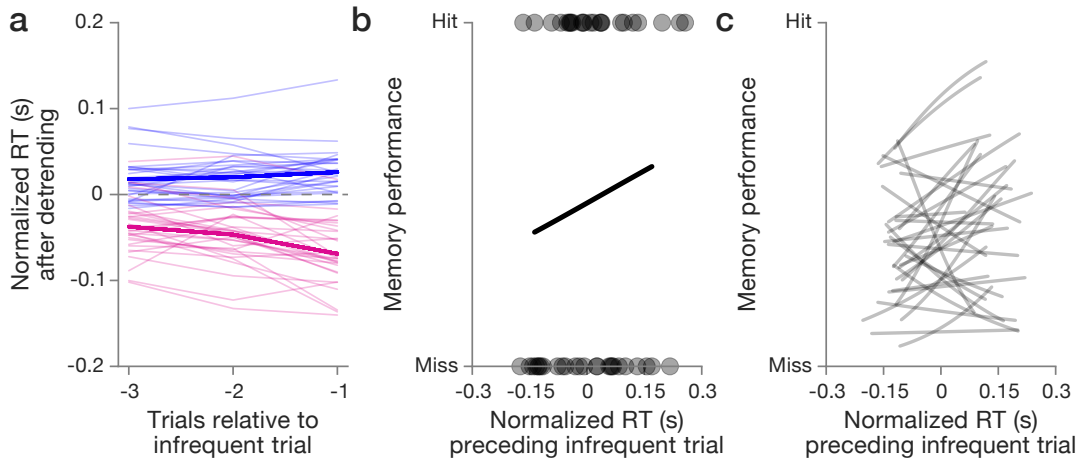

Figure 2.2: Relating sustained attention and memory. (a) RTs were slower before a correct (blue) vs. incorrect (pink) response to an infrequent trial (all time points, ps < 0.00001). Individual participants are plotted in thinner lines and the average in thicker lines. Raw RTs are reported in the text for this analysis but normalized RTs are depicted here because they were the input to the subsequent memory analysis shown in the other panels (statistics were unaffected). (b) Illustration of approach for quantifying relationship between sustained attention and memory in one representative participant. For every item from the infrequent category, the average RT over the three preceding trials in the sustained attention task was fit to the binarized recognition judgment from the surprise memory test using logistic regression. Each dot is one item and the line is the fitted logistic function. (c) The logistic functions for all participants are plotted, revealing a reliable positive slope on average. That is, slower preceding RTs correlated with better memory.

2.2.7 Discussion

This experiment demonstrates that it is possible to predict whether an image will be encoded into memory before it even appears, based on the state of sustained attention. This temporal precedence is consistent with attentional state being part of the causal chain that determines which memories are formed. However, this analysis is ultimately akin to an observational study, where a correlation is based on idiosyncratic fluctuations in attention not under experimental control. We seek to strengthen the directional interpretation of this relationship in Experiment 2 using a real-time, adaptive design in which attentional state

serves as an independent variable.

2.3 Experiment 2

The goal of this experiment is to test whether triggering encoding trials based on participants being in a good or bad sustained attentional state will influence what they remember. Specifically, we hypothesize that images presented while participants are responding faster than usual will be remembered worse.

2.3.1 Participants

Twenty-four undergraduates (15 female, 9 male; mean age = 19.22 years) from Princeton University participated for course credit. This sample size was chosen to be smaller than Experiment 1 because we anticipated that the triggering design would lead to larger effects. One additional participant was excluded for task performance more than 3 SDs below the mean (error rate on infrequent trials = 0.83).

2.3.2 Stimuli and Apparatus

Same as Experiment 1.

2.3.3 Procedure

As in Experiment 1, participants completed a sustained attention task and a surprise memory test, with the same display and response procedure. The sustained attention task was similar in design, with each participant viewing 500 unique indoor or outdoor images. The first 50 trials and last 50 trials had the same distribution of categories as Experiment 1, with 90% from the frequent category and 10% from the infrequent category. The middle 400

trials (trials 51 to 450) began with 100% of trials from the frequent category. Up to 40 of those 400 trials could be replaced with images from the infrequent category, depending on real-time measurements of behavior. Specifically, infrequent trials were inserted when our RT measure of sustained attention – average RT over the previous three trials – deviated slower than an upper bound or faster than a lower bound. These bounds were recomputed for each trial i from the real-time period in several steps: Analogous to Experiment 1, linear drift in the RTs was removed from trials 1 to $i - 1$. The overall mean and standard deviation (SD) of these trials was then calculated from the residuals. The average RT of just the preceding three trials ($i - 3$, $i - 2$, $i - 1$) was also calculated as a momentary index of sustained attention. If the moving-window RT was (a) slower than the mean plus one SD or (b) faster than the mean minus one SD, then the image for trial i was drawn from the infrequent category (Figure 2.3). Otherwise, the frequent image on trial i was not replaced. We required there to be a minimum of three frequent trials between infrequent trials (to not contaminate the moving-window RT). The average number of infrequent trials per participant during the real-time period was 27.8 (25.3–30.4). Participants were not informed that their RTs controlled when they would be shown images from the infrequent category.

In the surprise memory test, participants were shown all old images from the infrequent category (including triggered images), 50 old images from the frequent category, and a matched number of new images from each category (mean = 177, range = 154–204.

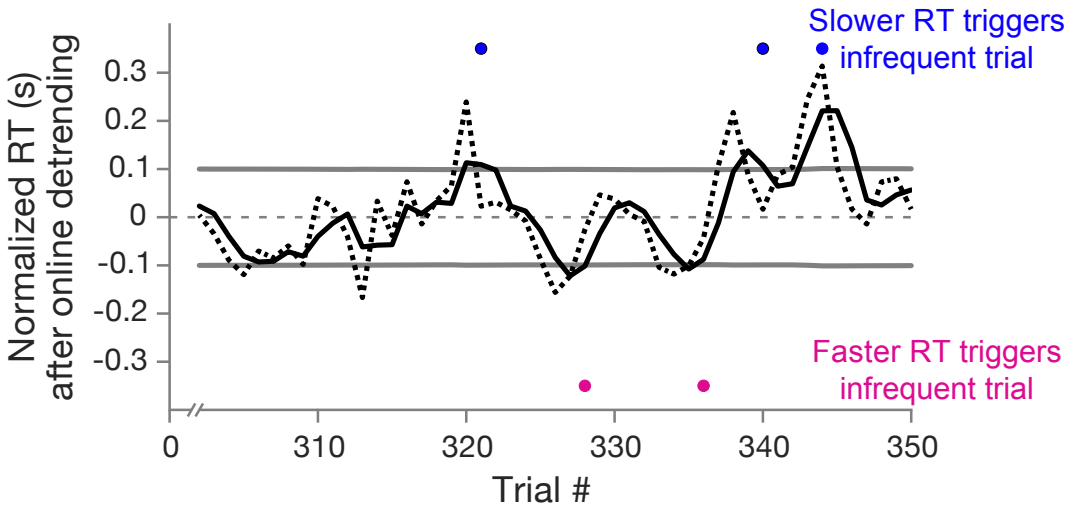

Figure 2.3: Online adaptive experimental design. The real-time triggering design is depicted for a period of trials from a representative participant. The dashed black line depicts the RTs to individual trials. The solid black line depicts the trailing window average RT over the 3 preceding trials. The bounds for triggering an infrequent trial, ± 1 SD, are plotted in a solid gray line. Infrequent trials were triggered if either the average RT was slower than the upper bound (blue dots) or faster than the lower bound (pink dots). Otherwise, frequent trials were shown.

2.3.4 Analysis

Infrequent trials from the real-time period were sorted based on whether they were triggered by fast or slow RTs. Otherwise, the analyses and statistics were similar to Experiment 1.

2.3.5 Results

During the sustained attention task, the overall sensitivity of the category judgments was above chance (mean $A' = 0.89$, 95% CI = 0.88–0.91; vs. chance = 0.5, $p < 0.00001$). For the infrequent category, the error rate was higher for images triggered because of fast RTs (0.66, 0.55–0.77) than to images that had been triggered due to slower RTs (0.26, 0.21–

0.31; $p < 0.00001$), as shown in Figure 4a.

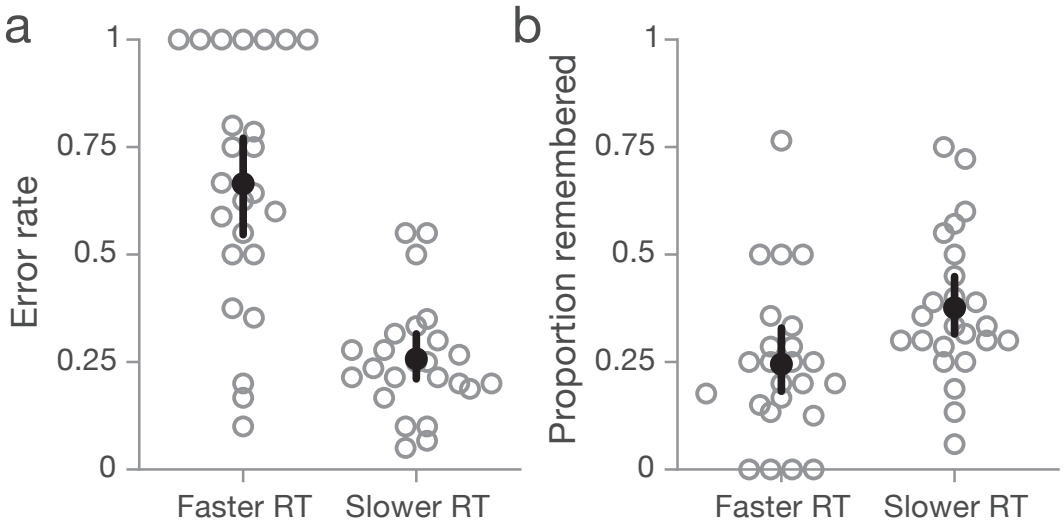

Figure 2.4: The fate of triggered trials. (a) Participants made more errors on infrequent trials that were triggered because of fast vs. slow RTs. (b) Participants remembered fewer images that were triggered because of fast vs. slow RTs. Individual participants are depicted in gray circles. The average is the solid black circle and the error bars depict 95% CIs.

During the surprise memory test, sensitivity was above chance for items from both the frequent category (mean $A' = 0.71$, 0.68–0.74) and the infrequent category (mean $A' = 0.79$, 0.77–0.82), although higher for the infrequent category ($p < 0.00001$).). The critical test of our hypothesis concerned the infrequent trials during the real-time period, and specifically whether memory was worse for trials that were triggered when we inferred that participants were in a bad vs. good sustained attentional state. As shown in Figure 4b, participants indeed remembered fewer of the images from trials triggered by fast RTs (0.24, 0.18–0.31) than slower RTs (0.38, 0.31–0.44; $p = 0.00024$).

One possible explanation for why RT triggering affected memory is that the state of sustained attention determines whether a behavioral error will be made to the infrequent

trial, and such errors prevent encoding into memory. Although this kind of indirect effect would be consistent with our claims about the causal role of sustained attention in memory encoding, we evaluated whether RT per se was additionally responsible by restricting the subsequent memory analysis to infrequent trials with correct responses. We included all participants who made at least one correct response in both attentional states ($n = 17$). Even in this restricted sample, memory was worse when trials were triggered from fast vs. slow RTs ($p = 0.014$). Note that this analysis was not possible in Experiment 1, as very few trials exceeded our bounds for fast and slow RTs idiosyncratically (beyond ± 1 SD). This is precisely a benefit of real-time adaptive designs, as we were able to ensure that all infrequent trials occurred at moments of clearly good and bad attentional states.

2.3.6 Discussion

This experiment employed a triggering design to more directly assess the relationship between attentional state and subsequent memory. Attentional state, operationalized based on fluctuations in RT, was used as an independent variable. The timing of each participant's session was customized on the fly, by monitoring attentional state and waiting until it was particularly good or bad to trigger critical probe trials. In doing so, we conclude that this experiment provides stronger support for the claim that sustained attention can control memory encoding.

2.4 General discussion

The goal of this study was to test the nature of the relationship between moment-to-moment fluctuations in sustained attentional state and episodic memory. In Experiment 1, we con-

firmed that quicker responses predicted upcoming errors in a sustained attention task and then used this measure of attentional state to predict how well stimuli from the task had been encoded incidentally. In Experiment 2, we treated response time as an independent variable in determining the timing of the experimental design, providing an endogenous manipulation of attentional state that allowed us to more precisely probe encoding during good and bad states. By considering fluctuations in sustained attention, we were able to explain variance in memory and control what participants would remember.

The online and adaptive design employed in Experiment 2 represents a methodological advance inspired by real-time neuroimaging studies (Stoeckel et al., 2014; Sulzer et al., 2013). This approach could have broad applications for other cognitive domains in which performance can be predicted in advance, including attentional capture (Leber, 2010), cognitive flexibility (Leber et al., 2008), and memory recollection (Otten et al., 2006). Once such predictive measures have been identified in a correlative manner, they can then be monitored in real time to trigger the onset of trials and control performance experimentally. Although this has been done with fMRI (e.g., Yoo et al., 2012), we have shown that behavioral studies also stand to benefit, including from strengthened causal inference and from a cleaner separation of cognitive states of interest.

Indeed, even in the realm of sustained attention, there are other predictive measures of performance that could be evaluated. For example, variability in response time across trials predicts accuracy in tasks like ours (Esterman et al., 2013; Rosenberg et al., 2013, 2015). Such variability is often calculated over a large time window, extending both before and after the trial of interest. For a real-time application, variability would need to be fully indexed in advance such that trials could be triggered in a leading-edge manner.

We found that slower preceding responses predicted both better inhibition of a prepotent response and better subsequent memory, which differs from previous studies that observed worse memory for items to which responses needed to be withheld (Chiu & Egner, 2015a,b). However, in our study we manipulated the trial proportions to establish a prepotent response, which affected the distinctiveness of the infrequent category relative to the frequent category. Therefore, we examined primarily the memory differences only within the infrequent trials (rather than between infrequent and frequent trials). It remains an open question whether variance in memory across infrequent trials can be predicted by preceding attentional state.

In sum, this study demonstrates that attentional state can be monitored behaviorally using response time and that such states can have profound consequences for memory. By tracking attention in real time, it may be possible to avoid bad attentional states during learning and to reduce the likelihood of forgetting.

3

Externalizing mental context reinstatement in real time*

3.1 Abstract

When episodic memories are encoded, they are associated with their context, both internal and external. Reinstating the same context that was present at encoding boosts memory recall performance. In a series of two experiments, we sought to characterize the relationship between mental context reinstatement and memory retrieval by asking participants to rein-

*Various portions of the contents of this chapter have been presented at the Context and Episodic Memory Symposium, the annual meeting of Society for Neuroscience, and the International Conference on Memory with coauthors N.B. Turk-Browne and K.A. Norman

state context prior to a memory recall test, via closed-loop neurofeedback. In Experiment 1, we demonstrated that this process of externalizing the fluctuations in context affects subsequent memory recall. In Experiment 2, we demonstrated that this relationship between context reinstatement and memory retrieval is not present when context reinstatement is provided without linking it to internal mental state. Together, these experiments demonstrate that neurofeedback can be a powerful tool for investigating memory phenomena and, through neurofeedback, we can affect memory retrieval.

3.2 Introduction

Returning to a specific context in which a memory was experienced, such as revisiting a particular restaurant, makes associated memories more accessible. This allows related experiences to be retrieved, such as the meal that was ordered during the previous visit. Experimental studies have demonstrated that memory performance is improved when mental context at the time of retrieval matches what was present at encoding (Manning et al., 2015; Smith & Vela, 2001). Moreover, this relationship between context match and memory retrieval is continuous. That is, the greater the degree of contextual match, the better the memory (Polyn et al., 2009; Staudigl et al., 2015).

Despite the evidence of contextual benefits for memory, we have yet to fully characterize how context exerts such an effect on memory. Not the least because items and their associated contexts are not entirely separable (Hsieh et al., 2014; Manning et al., 2011) although they may rely upon at least partly different neural mechanisms (Davachi, 2006; Ranganath, 2010). One possibility is that mentally time back in time to an episode prior to retrieval boosts recall of that episode, perhaps by rendering all memories encoded in the same con-

text more accessible. A previous study demonstrated that activity patterns associated with an item were reinstated prior to recall of that item (Polyn et al., 2005). However, this study was unable to disambiguate items from context, as the context was the category of that item, which was also shared with other items. Therefore, these results did not rule out the possibility that it was the items themselves that were being reactivated prior to recall. Here, we sought to more directly test the hypothesis that directed context reinstatement can affect memory retrieval.

We sought to drive context reinstatement by providing real-time feedback about how well participants were succeeding at reinstating a desired context. We used real-time analysis of functional Magnetic Resonance Imaging (fMRI) data to track how much a particular context was active, on a second-by-second basis (Stoeckel et al., 2014; Sulzer et al., 2013). Previous studies have demonstrated that real-time feedback about fMRI activity has profound effects on neural activity and behavioral performance (Amano et al., 2016; deBettencourt et al., 2015; Shibata et al., 2011). Here, we wanted to provide feedback while participants attempted to reinstate contexts associated with specific memories. Providing closed-loop neurofeedback makes it possible to continuously adjust the stimulus and task according to online measures of neural activity (deBettencourt et al., 2015).

Providing closed-loop neurofeedback makes it possible to continuously adjust the stimulus and task according to online measures of neural activity. In this way, neurofeedback can create a tight relationship between neural measures of context, the stimulus that is cueing memory retrieval, and the behavioral manifestation of memory (Figure 3.1a). With our implementation of neurofeedback, any fluctuations in the classifier's guess about the participant's mental context will cause corresponding fluctuations in in the physical visibility

of the cued context, and, in turn, the participant's internal mental context. This manipulation should therefore increase the relationship between classifier output and contextual reinstatement, including both behavior and neural indicators of memory system engagement. Therefore, we can use closed-loop neurofeedback to exaggerate subtle perturbations in neural state in order to better track contextual reinstatement and bias memory recall.

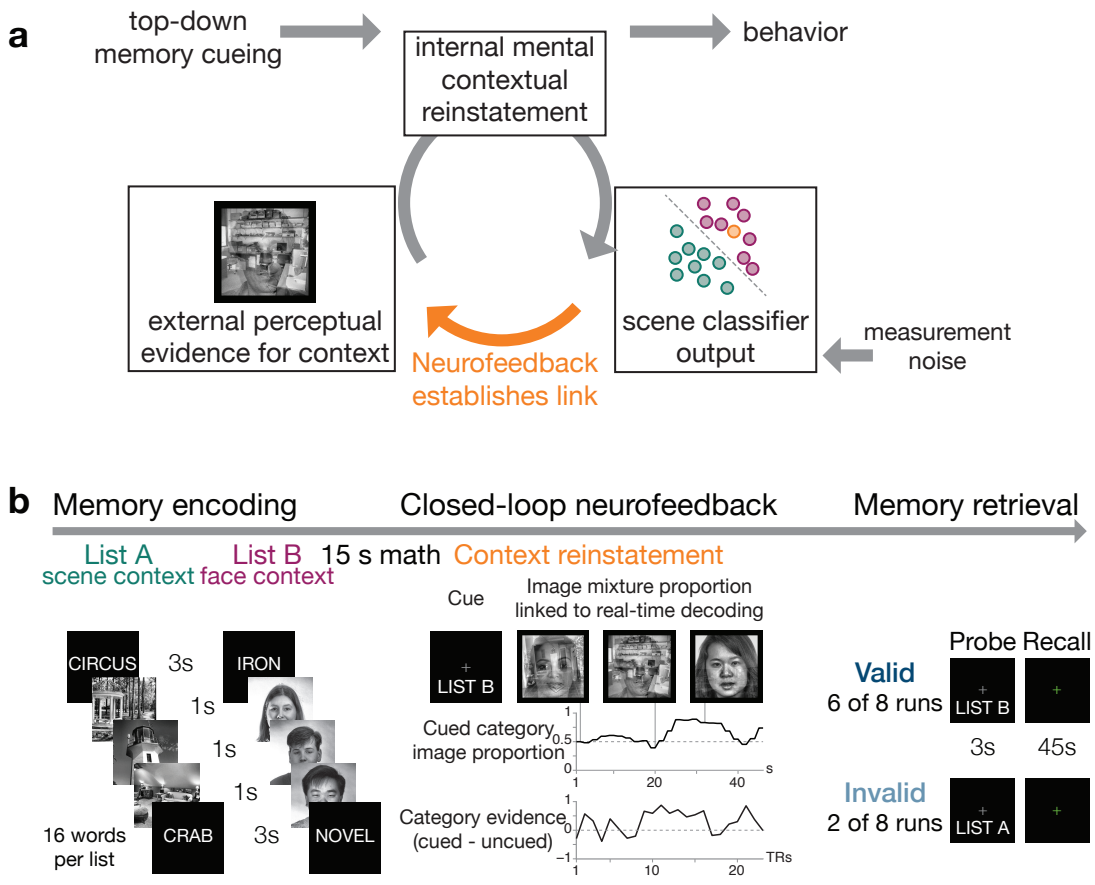

Figure 3.1: Study procedure. a. Closed-loop neurofeedback. Participants were provided a cue to guide mental state towards a particular context. Closed-loop neurofeedback linked the neural evidence for the cued context (measured by a multivariate pattern classifier in real time) to external stimuli. Changing the external stimuli influenced the internal mental state via bottom-up mechanisms (e.g. more external information consistent with that internal state). In turn, the internal mental state affects memory retrieval success. b. An example run of the task. During each run, participants studied two lists of sequentially presented words, List A & B. Each of the lists was embedded in a different context by interleaving the words with images of a single category (scenes or faces). After encoding, participants were provided a list name (e.g. List B) as a cue for context reinstatement. They were instructed to reinstate the appropriate context by thinking about the images that had appeared between the words of that list. Then, participants were shown composite face/scene images that reflected the neurofeedback, initialized at equal proportion face and scene. This mixture proportion was adjusted during the context reinstatement period to reflect the evidence for the cued context. Greater evidence for the cued context resulted in more of that category in the composite image (and less evidence resulted in less of that category). Finally, participants were presented with a list name as a memory probe. After the probe, they were instructed to freely recall as many words as possible from the probed list. In validly cued runs, the memory probe was the same as the context cue. In invalidly cued runs, the memory probe was different from the context cue.

We developed a task in which participants encoded memories in particular contexts, mentally reinstated one of the context, and then finally retrieved memories (Figure 3.1b). First, participants studied two lists of words, each of which was embedded in a context (either faces or scenes) by interleaving the words with images from that category (Gershman et al., 2013). After studying both lists, participants were cued to reactivate a context associated with one of the lists. Then, participants were presented with composite face/scene images, each comprising two images that had been each shown as context for one of the lists. These images, and the relative mixture proportions of the two categories, were updated in real time based on our evidence that the participants were reinstating the appropriate contexts. This evidence was provided in real time by testing a subject-specific multivariate pattern classifier (deBettencourt et al., 2015), which had been trained in separate localizer runs. When the classifier output was high (and thus context reinstatement was high), we increased the relative proportion of the cued category in the composite stimulus for the subsequent time point. In this way, we reinforced context reinstatement by rewarding successful attempts. Conversely, when classifier output was low, we decreased the relative proportion of the cued category in the composite stimulus (and increased the proportion of the uncued category), to communicate that participants were not succeeding. After the context reinstatement period concluded, participants were probed to freely recall as many items as possible from one of the lists. We manipulated whether the memory probe was the same list as had been cued: On validly cued runs, we tested memory for the same list for which neurofeedback had been provided; on invalidly cued runs, memory was tested for the other list.

We predicted, first, that this manipulation would establish a strong link between our

measure of mental context (based on classifier output) and the internal mental context of the participant, and, second, that this link would help participants get into the target mental context and would thus benefit memory recall. We aim to show that neurofeedback can elucidate the relationship between context and memory, and can indeed be used to influence memory retrieval.

3.3 Real-time neurofeedback

As an initial proof of concept, we evaluated to what degree the cued context was measured during the period of context reinstatement. If participants were reinstating the appropriate context, we predicted that there would be greater evidence for the cued context, compared to the uncued context. Consistent with this prediction, the multivariate pattern classifier reliably decoded the cued context during the context reinstatement period (mean accuracy $= 0.5847 \pm 0.010$ s.e.m.; chance $= 0.5$; $p < 0.00001$; Figure 3.2). This provides evidence that participants were able to reinstate the cued category.

However, this in itself does not indicate that neurofeedback was successful: one possibility is that participants are unable to control the neurofeedback, and any success at context reinstatement reflects initial biases or internal factors unrelated to the feedback loop. To account for this, we examined whether there was reliable evidence for the cued category at the start of the context reinstatement period, before the neural signals should be detectable by fMRI. There was no reliable evidence for the cued category (mean evidence = 0.4955 ± 0.02 s.e.m.; chance = .5; p = 0.847, two-tailed). This suggests the process of neurofeedback helped participants reinstate the cued context.

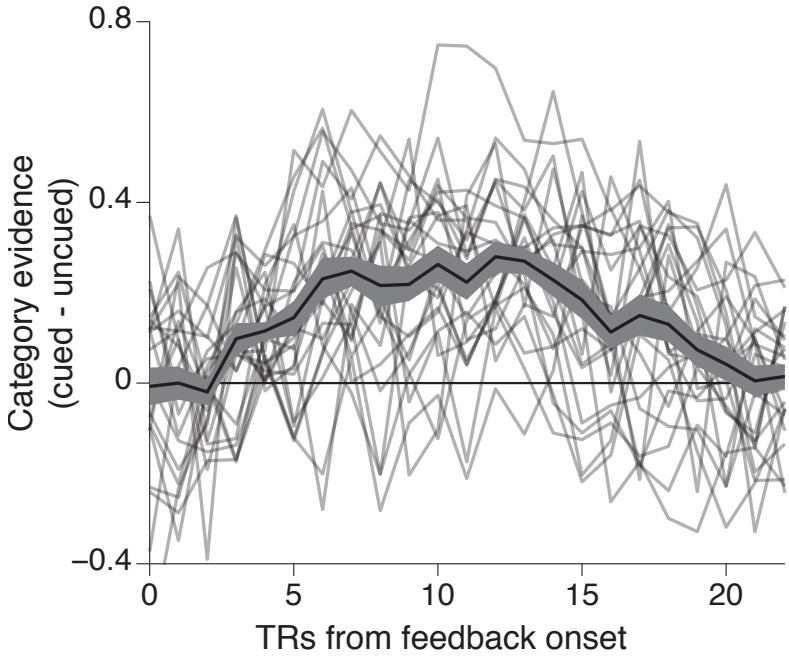

Figure 3.2: Classifier evidence during feedback. During the context reinstatement period, the classifier was tested at every TR. The feedback that the participants received was derived from the real-time TR-by-TR classifier evidence for the cued category minus the un-cued category. Each thin transparent lines depicts the average time course of classifier evidence for one participant (n=24) during the feedback period (TRs = 23). The average time course across participants is depicted in a thick black line, and the shaded region represents $\pm$ s.e.m.

3.4 Behavioral effects

Our behavioral measure of the efficacy of context reinstatement was the difference in memory performance when neurofeedback matched the list participants were asked to recall (i.e. validly cued runs) versus when neurofeedback did not match the memory probe (i.e., invalidly cued runs). If neurofeedback helps context reinstatement, and context reinstatement helps memory, memory recall should be higher on valid versus invalid runs. Consistent with the idea that reinstating an appropriate context boosts recall, recall performance

was better in the valid versus invalid condition ($p = 0.042$; Figure 3.3).

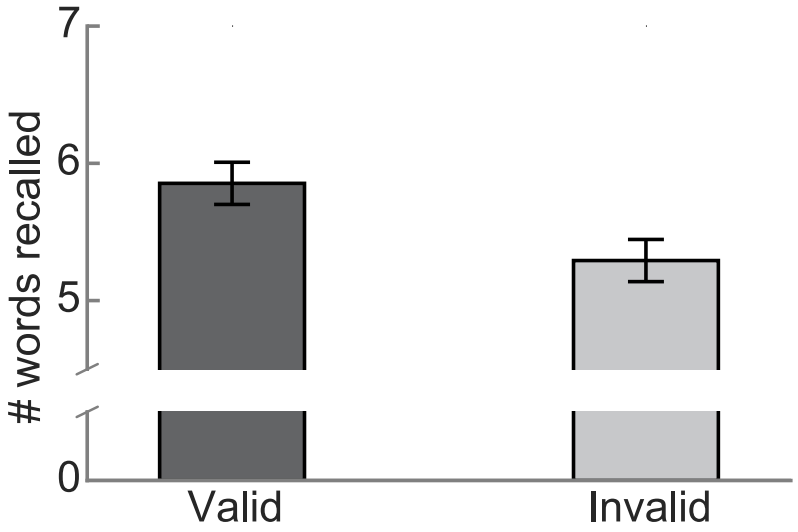

Figure 3.3: Effects of cue validity on memory recall performance. Number of recalls for the tested list. For validly-cued feedback runs, the cue at the start of the context reinstatement period matched the memory probe at the start of the free recall period. For invalidly-cued feedback runs, memory was tested for the other list (the one whose context was not cued at the start of the context reinstatement period). There was a reliable difference between the number of items recalled between these two conditions ($p < 0.05$). Error bars depict $\pm$ 1 within-participant s.e.m.

In invalid runs, participants only discovered that their memory would be tested for the other list at the time of the appearance of the memory probe. At that point, they might need to switch contexts in order to retrieve the words from the other list. We hypothesized that this switch might be reflected in a longer time to initiate recalls. In invalid runs, the time to first recall was slightly slower ($p = 0.08$), potentially supporting this contextual switch account.

Aside from population analyses, we can also look at individual levels of context reinstatement and its effect on behavior. For a single participant, there was variability across runs

in the degree to which the cued context was reinstated. We hypothesized that the more strongly the context was reinstated, the larger the effect on behavior. To calculate how strongly context was reinstated, we developed a context reinstatement score. First, across all participants, we identified the time points during the context reinstatement period that, on average, showed task-related activity. Then, for each run, we calculated the average evidence across those time points for the cued context relative to the evidence for the uncued context. As this was a binary classifier, the evidence for both categories at any time point sums to 1, and thus the score ranges from -1 to 1.

Next, we evaluated whether this measure of context reinstatement was related to memory performance. For each run, we plotted the context reinstatement score versus the total number of recalls for the tested list. For each participant, we measured the degree to which the context reinstatement affected recalls. To quantify this relationship, we fit a line that related these two variables (context reinstatement and number of recalls) and evaluated the reliability of the slope of this line across participants.

We predicted that the more context was reinstated on valid runs, the better memory performance would be (and thus, a positive slop between context reinstatement and recall). Indeed, there was a reliably positive slope across (mean slope $= 3.38 \pm 1.50$ s.e.m.; $p = 0.0049$; Figure 3.4a). Thus, on valid runs, the more context was reinstated, the better the memory.

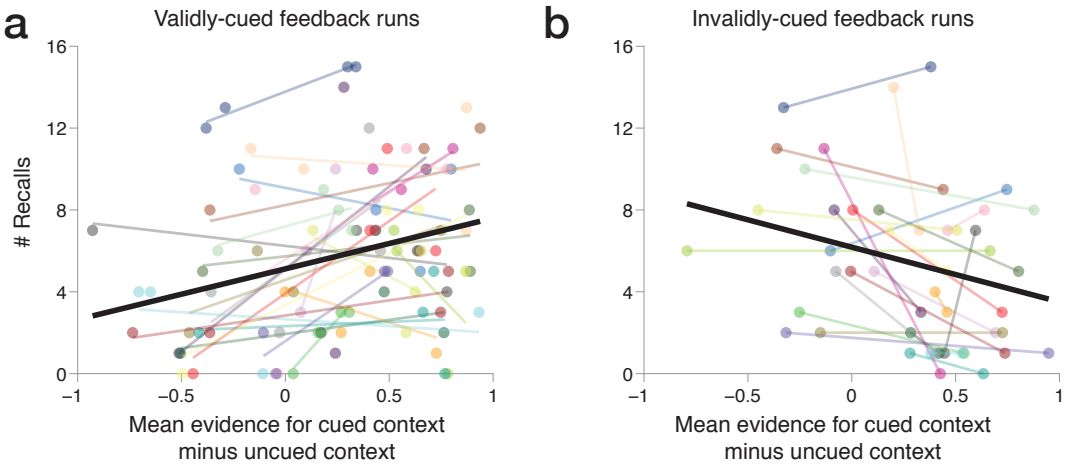

Figure 3.4: Context reinstatement correlates with behavior. (a) Evidence for the cued context in the valid condition was positively related to the number of recalls. Each dot represents data from a single run. On the x-axis is the mean evidence for the cued context minus uncued context during the context reinstatement period. On the y-axis is the total number of recalls for that run. Each participant (n=24) is depicted in a unique color with the linear fit of these points in the same color. The median fit across the population is depicted in the thick black line. The evidence for the cued context reliably predicted the number of recalls. (b) Evidence for the cued context in the invalid condition was negatively related to the number of recalls. Similar to the valid runs, the evidence for the cued context is depicted against the total number of recalls. In addition, the linear fit to the two points for each participant is depicted. The median fit across the population is depicted in the thick black line. Within the invalid condition, there was a negative relationship: memory was worse when there was more evidence for the other (non-tested) list during context reinstatement.

This positive relationship between context reinstatement and recalls could be due to any general effect (i.e. motivation, arousal) that varied across runs. If so, we would expect this relationship to persist for the invalid runs as well. However, if this effect reflects a consequence of context reinstatement on memory retrieval, we should expect the opposite for invalid runs: the more a participant reinstates the context for the non-tested list, the harder it would be to remember the list that was actually tested. Thus, there should be a negative relationship between context reinstatement and memory retrieval. Indeed, across participants, we found that there was a reliably negative relationship (mean slope $= -4.24 \pm 1.27$

s.e.m.; $p = 0.00011$, $n = 22$, two outlier participant excluded; Figure 3.4b). That is, for invalid runs, when the cued context was strongly reinstated, there were fewer recalls for the tested list (which, by design, was associated with a different context). Moreover, the relationship between context reinstatement and memory for valid and invalid runs was significantly different ($p = 0.00001$). Thus, the reactivation of contexts critically affect memory recall.

To assess whether these effect were driven by any outlier participants, we first normalized both context reinstatement and recalls within condition for each participant. Within the valid runs, the positive slope remained reliable after normalizing (mean slope $= 0.35 \pm 0.11$ s.e.m.; $p = 0.0016$). We also computed the fixed-effects correlation across all runs from all participants and found that to be positive (Spearman's $r = 0.36, p = 0.00037$). Finally, we assessed the reliability of this correlation by conducting a bootstrap analysis in which we reresampled participants with replacement and calculated the correlation on each new sample, and the relationship remained positive ($p = 0.0014$). These results suggest that the effects were robust across the population, and not driven by outlier performance in either context reinstatement or memory retrieval.

Similarly, for the invalid runs, the slopes remained negative after normalization (mean slope $= -0.54 \pm 0.16$; $p = 0.0010$). Additionally, both the fixed-effects correlation (Spearman's $r = -0.36, p = 0.011$) and the bootstrapped correlation ($p = 0.0039$) were negative. Finally, the difference between the valid and invalid conditions remained significantly different ($p = 0.00006$). Taken together, these results lend additional support to the hypothesis that reinstating context with neurofeedback prior to recall affects subsequent memory performance.

3.5 Replaying context

Despite our consistent findings that context reinstatement with neurofeedback affects recall, it was not clear whether this effect was specifically due to neurofeedback or reinstatement in general. The closed-loop neurofeedback experimental design purposefully coupled stimulus display and neural measurements. However, in a typical experiment, the stimulus display would not be contingent on neural activity. The closed-loop neurofeedback design purposely coupled the external stimulus display and neural measurements of internal state: for example, participants saw more face information when their brain state was more highly in the face context. Thus, it is unclear whether neurofeedback *per se* drove the effects we observed, or if they were a result of changes in the stimulus display, irrespective of the participant's internal state.To investigate this, we ran a follow-up experiment that included a control condition in which stimulus information was not modulated by the participant's mental context. To maximize power, we made slight changes to the experiment design, while keeping the basic structure intact: the cues were 100% valid in this experiment, the contexts were always scenes, and the runs were reduced to a single list. Based on offline analyses, we expanded the region of interest to include occipital cortex as well as temporal cortex. For half of the runs, context reinstatement was identical to the first experiment, with participants receiving neurofeedback with composite face/scene images that changed in terms of mixture proportions. For the other half of the runs, participants viewed all of the images that had appeared between words during the encoding phase at full coherence without any competitive information on the screen. Rather than adjusting the mixture proportions of images in real time, this was meant to provide the strongest possible context reinstatement.

By design, these conditions differed in the overall amount of context-relevant information on the screen. Given this, it is perhaps unsurprising that they differed in the total amount of classifier evidence decoded during the context reinstatement period. During the runs without feedback, when participants viewed unmixed images, the classifier's accuracy was 89% ± 1). During the feedback runs, when participants viewed composite face/scene mixtures, the classifier's accuracy was 57% ± 5.

In the runs without feedback, participants were presented with fully coherent and strong context cues during the reinstatement phase – all images were replayed to the participant without anything else overlaid. With neurofeedback, there was overall weaker context evidence on the screen, but there is a link between what the participant sees and what there internal context is. Could memory performance be improved had we just given them greater amounts of context-related information? In fact, we found no reliable difference between the control runs and neurofeedback runs ($p = 0.22$). Therefore, replaying the context it its entirety was not sufficient to improve memory above and beyond neurofeedback, suggesting that memory behavior is not simply a consequence of how much external context information is present

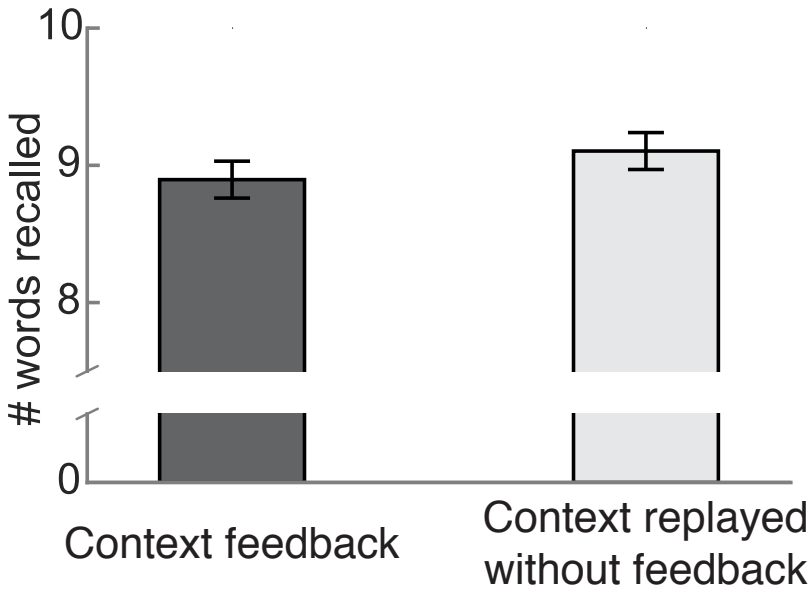

Figure 3.5: Effects of feedback on memory recall performance. Number of recalls for the tested list. For feedback runs, participants were presented with composite face/scene images during context reinstatement period, and the image mixture proportions varied via closed-loop neurofeedback. For no feedback runs, participants were presented with fully-coherent images during the context reinstatement period, in order to replay the cued context. There was no reliable difference between these runs. Error bars depict ± 1 within-participant s.e.m.

Our critical prediction is that the process of internally reinstating context, and the link between internal context and external context, predicts memory. Since this link only exists for the neurofeedback runs, such a relationship should only be found for this condition. We first repeated the analyses from the first experiment, examining whether context reinstatement during feedback runs correlates with memory behavior. Indeed, we replicated this effect (mean slope $= 1.67 \pm 0.66$; $p = 0.016$; Figure 3.6a). The critical question for this experiment was whether this same relationship would occur if there was no link between internal and external context. There was no such relationship between context reinstatement and memory performance in the runs without feedback (mean slope $= -2.28$,

$p = 0.87$; Figure 3.6b). The difference between the slopes relating context reinstatement and behavior between the feedback and non-feedback runs was reliable ($p = 0.032$). This provides further evidence that externalizing the dynamics of context reinstatement via neurofeedback reveals a relationship between prior context and memory recall.

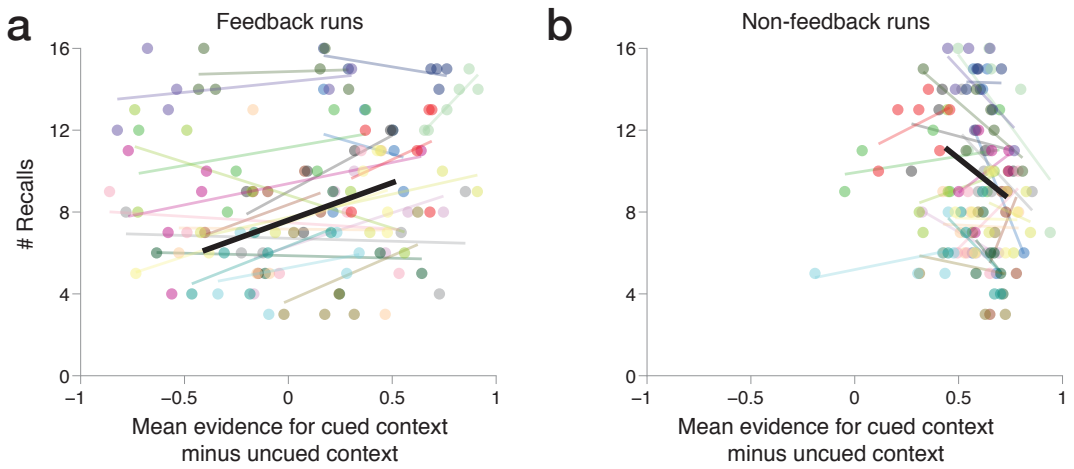

Figure 3.6: Feedback mediates the link between context reinstatement and memory. (a) Evidence for the cued context in the feedback condition was positively related to the number of recalls. Each dot represents data from a single run. On the x-axis is mean evidence for the cued context. This was calculated by averaging the evidence for the cued category (relative to the uncued category) during the context reinstatement period. On the y-axis is the number of recalls for that run, normalized within the runs in the feedback condition. The feedback runs (and their linear fit) for every participant are depicted in a unique color for that participant. The black line is the median fit across the population. In feedback runs, the evidence for the cued context reliably predicted the number of recalls, replicating the effect in the valid feedback condition of Experiment 1 (b) Evidence for the cued context in the condition in which context is replayed, but without a link to internal mental state. The axes of the plot are the same, but the data are from the runs in which feedback was not provided. In these runs, the relationship was not significant, suggesting the link between internal context and memory recall is mediated by neurofeedback.

Though the non-feedback control condition did not positively affect recall, we found higher overall classifier evidence compared to the mixed images. One possible interpretation of the lack of relationship between context reinstatement and memory is that the

greater overall amounts of classifier evidence pushed it beyond the sensitive range. That is, presenting fully-coherent images created a ceiling effect for neural measurements. To investigate whether the variability in the possible classifier values affected the detection of a relationship between context and recall, we correlated the variability of context reinstatement across the runs (quantified as the standard deviation) with the slope of the relationship. We did not find a significant relationship in the runs without feedback ($r = 0.21, p = 0.36$) nor in the runs with feedback ($r = -0.37, p = 0.10$). Therefore, these analyses suggest that the variability in classifier outputs in runs without feedback was not sufficient to explain the lack of a relationship.

3.6 Whole-brain activity differences during feedback

We examined the univariate differences in whole-brain activity during the context reinstatement period, to elucidate the neural mechanisms that might be at play in linking internal states, external states, and memory. We hypothesized that these areas would (1) be active during feedback and (2) be more active during feedback than no feedback. We found that many areas, including perceptual regions as well as regions typically associated with memory retrieval (including ventrolateral prefrontal cortex) were more active during the context reinstatement period with feedback, compared to context reinstatement without feedback (Figure 3.7).

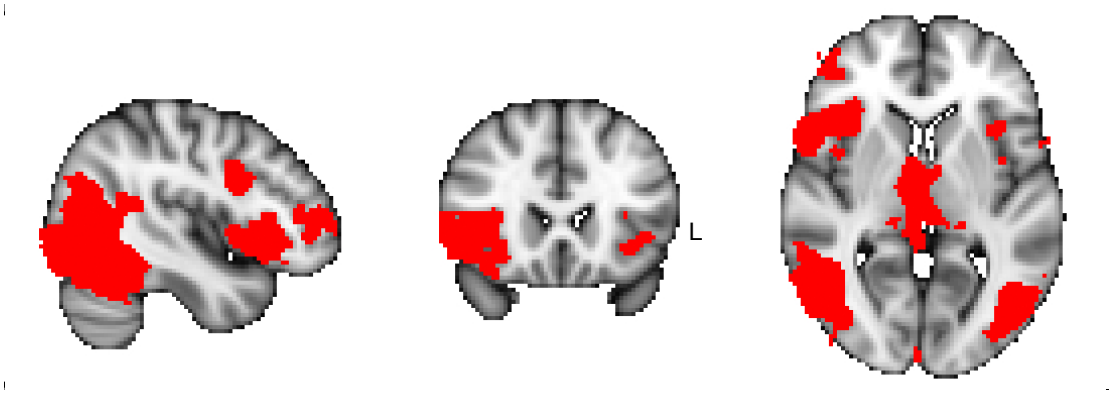

Figure 3.7: Whole-brain univariate activity during context reinstatement. Voxels depicted in red are those voxels that were activated in feedback runs and that were more active in feedback runs than non-feedback runs. Voxels were thresholded with FSL's randomise cluster correction ($p < 0.001$).

To understand the independent contributions of different areas and networks in the brain, we looked specifically at the correlation between the real-time measurement of context (the time series that constituted the feedback) and all of the voxels in the brain. We found that large clusters in ventral temporal cortex were correlated with the multivariate pattern evidence for the context. These regions have been found to represent these types of images (Epstein & Kanwisher, 1998) and are important for representing context (Davachi, 2006; Ranganath, 2010).

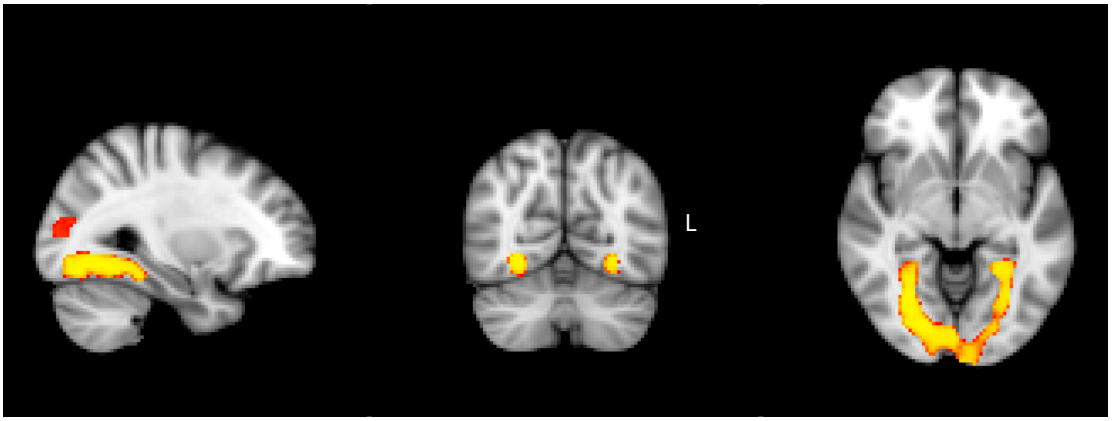

Figure 3.8: Voxelwise correlations with feedback. Regions displayed are those in which classifier information during context reinstatement was significantly correlated with multivariate activity ($p < 0.05$, FSL's randomise cluster correction).

3.7 Discussion

This study demonstrated that neurofeedback may provide a means to externalize and manipulate the otherwise internal mental process of context reinstatement. Moreover, the link between internal mental context and memory was present only during closed-loop neurofeedback: internal mental context was not correlated with memory in the absence of an externalized representation of that context. Previous studies have demonstrated that using real-time neurofeedback to achieve a desired mental state was associated with behavioral effects that were not otherwise present when naturally occurring neural fluctuations were not linked to feedback (Schafer & Moore, 2011). Here, we demonstrate that neurofeedback can be used to more tightly link internal mental context and the outward expression of memory, and may therefore be used to influence memory retrieval.

These results lend further support that context reinstatement can affect memory recall. Although mental context and the associated memories are intricately related (Davachi,

2006; Ranganath, 2010), here we attempted to focus on context specifically. One possibility is that during the reinstatement phase, the process of reinstatement also brought the words to mind, even though we instructed participants to focus on the images. However, the neurofeedback was derived from a pre-trained multivariate pattern classifier, which had been trained on an independent localizer period without any words. Therefore it is unlikely that any item-related information could have driven the observed behavioral relationships.

This study showed that reinstating mental context prior to a memory test can affect recall performance. This builds upon other literature that suggests that a preceding mental state at one moment can profoundly affect subsequent behavior, including sustained attention errors (deBettencourt et al., 2015), reaction time in a vigilance task (Hinds et al., 2013), attentional control (Leber, 2010), cognitive flexibility (Leber et al., 2008), and memory encoding (Yoo et al., 2012). Here, via a novel experimental paradigm, we demonstrate some of the first evidence that preceding neural state influences memory retrieval. This demonstrates that volitionally activating a specific mental state can influence subsequent mental processing.

Our control condition without feedback indicated that traditional means of probing the relationship between context reinstatement and memory were not sufficient to reveal a relationship between context reinstatement and subsequent memory retrieval. Future control experiments may be able to fully characterize the effects of providing feedback on context reinstatement. These include providing no images during the context reinstatement period, non-closed-loop feedback, yoked-control feedback, feedback from other brain regions, and uninstructed feedback. We will consider each of these in turn, as well as describe some of our predictions.

A blank context reinstatement period without any images would require the participant to generate the internal mental context by themselves, without external visual cues from the experimenter. This would likely prevent the ceiling effect observed in Experiment 2, but, because participants are not receiving feedback, they might not be able to achieve sufficiently accurate contextual states to affect their memories.

While our form of closed-loop feedback entangles mental processes with the stimulus by design (deBettencourt et al., 2015; Mishra & Gazzaley, 2015), control experiments could be run using a more straightforward means of providing feedback. Traditionally, thermometers or other visual displays have indicated to participants their level of BOLD activity in real-time fMRI designs (Stoeckel et al., 2014; Sulzer et al., 2013). This manipulation would pair the top-down cue with feedback that was not stimulus-related, to generalize to whether feedback itself (without perceptual cues) could also create a relationship between mental context display and neural data.

An additional concern might be that feedback might carry with it particular demand characteristics and expectations. To combat this, previous studies have included groups of control participants who were given identical instructions, except that they received feedback based on data from another participant or run, rather than their own online measures of context reinstatement (deBettencourt et al., 2015). A possible disadvantage is that misleading feedback might actually be detrimental, and might discourage participants from utilizing the feedback in subsequent runs. This control could help characterize the balance between perceptual and internal context in the proposed mechanism for closed-loop neurofeedback.

In these studies, feedback was based on classifier evidence in the temporal lobe (Exper-

iment 1) and the temporal + occipital lobe (Experiment 2). Additional experiments could include conditions where feedback could be provided from other areas (e.g. the posteromedial system that has been associated with mental context, (Ranganath & Ritchey, 2012) or a frontoparietal attentional network thought to be important for feedback about goal-oriented signals (deBettencourt et al., 2015). Feedback could even be provided from areas that are not thought to be involved in the context reinstatement process. These approaches could assess the relative contributions of different regions and networks to mental context reinstatement, and how these regions independently guide subsequent recall.

In the current study, we explicitly told participants to think of the list contexts during neurofeedback. An alternative approach is to not provide any cues or instructions to the participants with respect to what they should be thinking about during the neurofeedback phase. Incepted feedback has still been found to result in behavioral changes for perceptual learning (Shibata et al., 2011) and lower-level visual associations (Amano et al., 2016; deBettencourt & Norman, 2016). This experiment could be modified such that participants are not aware of the goals of the context reinstatement period. During the context reinstatement period, participants could be told to maximize the size of a feedback meter (e.g., the diameter of a green dot). Unbeknownst to the participants, the green dot would correspond to the amount of evidence for the cued category. This design would assess the relationship between internal mental context and brain decoding, removing the factors of awareness of the manipulation and the match between internal and external context. However, participants might not be able to achieve strong enough contextual states to effect behavior.

In this study, we have demonstrated that closed-loop neurofeedback allows us to more

directly perturb the neural activity and probe theories of memory retrieval. This technique could be expanded to other experiments in which context has been implicated, for example, rather than assisting memory retrieval, as we have demonstrated here, this same technique could be used to aid participants in forgetting (Manning et al., 2016; Sahakyan & Kelley, 2002). Eventually, it might be possible to further develop this technique to provide training for context reinstatement, and possibly eventually be useful for psychiatric disorders that involve context such as addiction and post-traumatic stress disorder.

Acknowledgments

This work was supported by US National Institutes of Health grant R01EY021755, US National Science Foundation (NSF) grant BCS1229597, NSF fellowship DGE1148900 and the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of these funding agencies.

We would like to thank Paula A. Pacheco for assistance in annotating the recall data, and Anne Mennen and Mariam Aly useful comments on the manuscript. We would also like to thank the PNI computing staff, including John Wiggins, Garrett McGrath, Ben Goldberg and Randee Tengi for technical support related to the real-time analyses.

3.8 Methods

3.8.1 Participants

Forty-eight adults (twenty-four per experiment) participated in the study for monetary compensation (25 female, 44 right-handed, mean age = 20.1 years). Power analyses could not be performed because of the use of a new paradigm and unknown behavioral and neu-

ral effect sizes. The sample size was chosen before the start of the experiment and is fairly common for an fMRI study. Six additional fMRI participants were excluded from Experiment 1: two because of technical problems with real-time data or audio acquisition, one for falling asleep during several runs, and four for excessive motion, defined both within ($\geq$ 3 mm) as well as across run ($\geq$ 5 mm), due to the lack of motion correction across runs during the fMRI session. Five additional fMRI participants were excluded from Experiment 2: one due to lack of understanding the instructions, and four for excessive motion, using the same standards as in Experiment 1. All participants had normal or corrected-to-normal visual acuity and provided informed consent to a protocol approved by the Princeton University Institutional Review Board.

3.8.2 Stimuli

Word lists

There were 16 lists of words, each comprising 16 words (Manning et al., 2016). Lists were randomly paired with a context of faces or scenes, and the order of the words within lists was randomized for each participant. For Experiment 2, participants completed 12 runs, each with a single list. Participants were presented with 12 of the 16 word lists in the same manner as before.

Images

Images consisted of grayscale photographs of male and female faces and indoor and outdoor scenes (Phillips et al., 1998; Xiao et al., 2010). These images were combined into composite stimuli by averaging pixel intensities using various weightings (for example, 60%

face, 40% scene). The stimuli were displayed on a projection screen at the back of the scanner bore and viewed with a mirror attached to the head coil.

3.8.3 Procedure

Localizer runs

Participants completed two localizer runs, viewing blocks of scene, face, and object images. Each block consisted of 12 images, presented for 1s with a 0.5s period of fixation between each image. A total of 12 blocks were presented, with 4s of fixation separating each block. Participants were asked to detect back-to-back image repetitions and respond by pressing a button.

Memory runs

Each memory run consisted of three phases: study, context reinstatement, and recall. Before the first list, each run began with 38s fixation. The name of the list (either List A or List B) was presented for 3s, followed by 1.5s fixation. Each word was presented for 3s. Between each word, 3 images (either faces or scenes) were presented, each for 1s. In total, a list was comprised 16 words and 45 images, with 6s of fixation at the end. In Experiment 1, there were two lists: one was paired with a scene context and the other with a face context. In Experiment 2, there was only a single list (List A) which always a scene context. After the study phase, there was a brief period of math problems (15s) to distract participants and prevent rehearsal. First, the response mapping for the math problems was displayed for 1.5s. This was followed by 9 math problems (1.5s each). Participants viewed an addition problem consisting of two single digits (e.g. $3 + 4$) and responded whether the sum was even or odd.

After the math section was complete, there was 3s fixation.

The cue for the context reinstatement period (either List A or List B) was displayed for 3s. This was followed by 33.75s of images, each displayed for .75s. For feedback runs, these images were composite face and scene images. During feedback runs, the first two images (1.5s) were always initialized at 50% face and 50% scene. The mixture proportions for the remaining trials were determined on the basis of real-time multivariate pattern analysis (MVPA) of the fMRI data, ranging from 17% to 98% of the category of the cued context (83% to 2% of the category of the uncued context)(deBettencourt et al., 2015). In Experiment 1, the images presented during context reinstatement were composite pictures of the faces and scene images that had appeared during the study phase in a random order. In feedback runs during Experiment 2, it was the same, but the face images were randomly selected as there was only a single list and it was always paired with a scene context. In non-feedback runs during Experiment 2, the images during context reinstatement were 100% scene images from the study phase in random order. At the end of the context reinstatement period, before the memory probe appeared, there was .75s fixation.

The recall phase began with 3s of the memory probe, indicating the name of the list to recall (either List A or List B). Then, the fixation dot on the screen turned green to indicating the start of the recall period. Participants were given 45 seconds to recall the items from one of the lists in any order. At the end of the recall period, the fixation dot turned back to white for 4s. Then, they were presented with a score from that run's context reinstatement period, corresponding to their performance during the context period. No feedback was provided on their recall performance. At the end of the experiment, participants received up to $10 extra corresponding to their cumulative scores on all the runs.

In Experiment 1, participants completed 8 runs of the task (6 valid, and 2 invalid). This imbalance between conditions was designed so that they wouldn't disregard the context reinstatement period. To ensure that participants trusted the cue, the first two runs of the experiment were always valid. One invalid run occurred during runs 3-5, and the other invalid run occurred during runs 6-8, and they were not permitted to occur back-to-back (i.e. runs 5 and 6). Given the temporal imbalance between conditions, runs 1 and 2 were excluded from additional analyses. The two invalid runs were counterbalanced for list cue (either "List A" or "List B") and cued context category (either scene or face).

3.8.4 Data acquisition

Experiments were run using the Psychophysics Toolbox for Matlab (http://psychtoolbox.org/). Neuroimaging data were acquired with a 3 T MRI scanner (Siemens Skyra) using a 20-channel head and neck coil. We first collected a scout anatomical scan to align axial functional slices to the anterior commissure-posterior commissure line. Then, a high-resolution magnetization-prepared rapid acquisition gradient-echo (MPRAGE) anatomical scan was acquired to use for real-time spatial registration. Functional images were acquired using a gradient-echo, echo-planar imaging sequence (1.5 s repetition time, 29 ms echo time, $3 \times 3 \times 3.5$ mm voxel size, 64×64 matrix, 192 mm field of view, 27 slices).

3.8.5 Statistics

Subject-level bootstrap resampling (Efron & Tibshirani, 1986) was used to assess random-effects reliability for comparisons of a small number of variables to chance or each other, by resampling over 100,000 iterations. One-sided tests were used for directional hypotheses and two-sided tests for nondirectional hypotheses.

3.8.6 Real-time analyses

Preprocessing

At the start of the fMRI session, the anatomical region-of-interest (ROI) was registered to the native functional space using FSL's FLIRT. For experiment 1, the ROI was the temporal lobe, and for Experiment 2, the ROI was expanded to also include the occipital lobe. During the fMRI session, functional data were reconstructed and prospective acquisition correction and retrospective motion correction were applied. After motion correction, the file was transferred to a separate analysis computer for the remainder of the analyses. The anatomical ROI mask was applied to reduce the voxel dimensionality. The volume was spatially smoothed in Matlab using a Gaussian kernel with full-width half-maximum (FWHM) = 5 mm. In Experiment 1, a high-pass filter adapted from FSL (cutoff = 100s) was applied in real time, in Experiment 2, no such high-pass filter was applied (since the run length was shorter). After each localizer run, the BOLD activity of every voxel was z-scored over time. During memory runs, the BOLD activity of each voxel was z-scored starting after the study period based on the study period mean and standard deviation.

Multivariate pattern analysis

A multivariate pattern classifier was trained on data from the face and scene blocks from both localizer runs. Labels were shifted 6s forward in time to account for the hemodynamic lag. For Experiment 1, we conducted MVPA using penalized logistic regression with L2-norm regularization (penalty = 1). Offline analyses on these data suggested that L1-norm regularization might have greater accuracy, so for Experiment 2, we used penalized logistic regression with L1-norm regularization (penalty = 1).

The trained model was tested in real time on brain volumes obtained during the context reinstatement period. For each volume, the classifier estimated the extent to which the brain activity pattern matched the pattern for the two categories (face and scene) on which it was trained (from 0 to 1). The neurofeedback was based on the difference of classifier outputs for the task-relevant category minus task-irrelevant category. However, given that this was a binary classifier, the evidence for each category is perfectly anticorrelated. However, the subtraction means that the evidence ranges from -1 (complete evidence for the wrong state) to 1 (complete evidence for the correct state).

Neurofeedback

The output of the classifier was used to determine the proportion of the images from the cued and uncued categories in the composite stimulus on the next trial. The preprocessing and decoding of volume i were performed during volume $i + 1$ and the classifier output was used to update the stimulus mixture for the two trials in volume $i + 2$. This resulted in a minimum lag of 2 s (two trials) between data acquisition and feedback. Moreover, classifier output was averaged over a moving window of the preceding three volumes ($i - 2$, $i - 1$ and i for feedback in volume $i + 2$), meaning that feedback was based on brain states $2 - 8$ s in the past. The average classifier output was mapped to a proportion of the task-relevant category using a sigmoidal transfer function(deBettencourt et al., 2015)).

Participants were aware of the feedback manipulation. Before the fMRI session, the were given instructions about the experiment, which included the feedback manipulation. They were told that the images in the context reinstatement period would reflect our measurements of their internal mental context, and that the images would get easier if they were

reinstating context well. They completed a short run of the task outside the scanner to familiarize themselves with the experimental design. During that run, they were shown examples a composite stimulus, and how the mixture proportion could change due to our measurements.

3.8.7 Behavioral analyses

Vocal responses were recorded during the recall period using a customized MR-compatible recording system (FOMRI II; Optoacoustics Ltd.). We used the Penn TotalRecall tool (http://memory.psych.upenn.edu) to score and annotate the audio. All annotations were completed without knowledge of the experimental design and were verified by an independent scorer who had no knowledge of the experimental manipulation or hypotheses.

Decoding accuracy

Multivariate pattern classifiers were trained using the face and scene blocks from both localizer runs. Classifier performance was assessed by testing the classifier on TRs during the context reinstatement period. Two measures were used, evidence for the cued category and accuracy. For a logistic regression classifier, evidence for the cued category can be considered the distance to the discriminating hyperplane. A TR was labeled as accurate if the evidence for the cued category was greater than the evidence for the uncued category. To assess whether there was any bias at the start of the context reinstatement period, the classifier accuracy was calculated for the data from the last TR during cue presentation (1.5s after cue onset). To evaluate classifier performance during the context reinstatement period, accuracy was computed over all TRs that constituted the feedback. Chance was assessed by permuting labels 100,000 times and recalculating accuracy for each of these permutations.

Relationship to behavior

To explain how context reinstatement related to memory behavior, we first obtained summary measures for each of these components for each run: (1) how successful they were at reinstating the cued context and (2) their memory performance. Context reinstatement was calculated by averaging the evidence for the cued context (relative to the uncued context) during the period of context reinstatement. To ensure that we were maximally sensitive to the period of time that context evidence was measurable, and to equate between the different contexts (face and scene), we first identified the time points where on average, across participants, there was was significant evidence for that category. However, the observed effects were insensitive to the exact selection of time points (e.g. over the entire context reinstatement period: $p_{valid} = 0.0016; p_{invalid} = 0.0009; p_{diff} = 0.0001$). Memory performance was calculated as total number of independent words that were recalled for the probed list. When necessary, to permit comparisons across participants with different levels of memory behavior and/or context reinstatement, we z-scored both context reinstatement scores and memory recalls, separately for valid and invalid runs. Correlations between two variables were estimated with Spearman's rank correlation after applying robust methods to eliminate the disproportionate influence of outliers in small samples (Pernet et al., 2013).

Univariate analyses

Functional data were preprocessed using FSL (http://fsl.fmrib.ox.ac.uk/), including slice-timing correction and spatially smoothing (5 mm FWHM). They were then transformed into standard Montreal Neurological Institute (MNI) space by linearly registering to the MPRAGE images and then to the MNI152 standard brain. To identify significant univari-

ate changes during the context reinstatement period, a single regressor was modeled for the entire context reinstatement period. Group level effects were calculated for all feedback runs from all participants. The next step was to assess the differences between runs in which feedback was provided and runs in which there was no neurofeedback. These two types of runs were independently modeled within subjects, and the paired differences across the population was calculated. Significance testing on voxel-wise brain maps was conducted with a permutation test in FSL's 'randomise' function (Nichols & Holmes, 2002) and corrected for multiple comparisons using threshold-free cluster enhancement (Smith & Nichols, 2009).

Voxelwise correlation with feedback

The goal for this analyses was to compute how much a voxel contributed to the real-time decoding and feedback. For each voxel, we extracted the timeseries of the classifier output during the context reinstatement period and correlated this with each voxel's activity during the context reinstatement period. We repeated this for all runs, and then averaged across runs for each condition (feedback and no feedback). Therefore, for each participant, we obtained a single value per voxel, which was the average correlation of that voxel with real-time decoding. For significance testing, we first Fisher z-transformed these correlation values and then conducted a permutation test in FSL's 'randomise' function (Nichols & Holmes, 2002) and corrected for multiple comparisons using threshold-free cluster enhancement (Smith & Nichols, 2009).

4

Conclusion and future directions

4.1 Conclusion

Real-time neurofeedback offers extraordinary potential to better understand cognitive functions, including attention and memory. In this dissertation, we have demonstrated that closed-loop neurofeedback about fluctuations in sustained attention can be used as a training signal, resulting in improvements in behavior. Better sustained attention is important is a critical component of the memory encoding process, and the amount of sustained attention is related to later memory success. We demonstrated that we can incorporate knowledge of real-time fluctuations of sustained attention to bias later memory for items. Even after encoding, we can provide feedback about the moment-to-moment progress in

mental context reinstatement to illuminate a relationship between context reinstatement and subsequent memory recall. Together, these results demonstrate the promise of incorporating information measured in real time for cognitive neuroscience.

4.2 Clinical applications

Although the experiments presented here are not specifically clinical in nature, there are many possible clinical applications of real-time fMRI and neurofeedback (deCharms et al., 2005; Stoeckel et al., 2014). Neurofeedback can shed light on and potentially correct aberrant cognitive processes, making this technique particularly useful for treating psychiatric disorders.

In particular, our findings involving neurofeedback on sustained attention could be beneficial for the numerous psychiatric disorders related to goal-directed attention. In particular, depressed individuals have been demonstrated to have a negative attention bias, in that they attend more to negative (e.g. sad) versus neutral or positive (e.g. happy) information (Disner et al., 2011). One early demonstration of this negative attention bias was through eye tracking data (Kellough et al., 2008). In this study, depressed individuals were presented with an array of images of different emotional valences, including negative images. Eye tracking fixations revealed that overt spatial attention was more frequently allocated to the negative images in an array. Previous work has shown that the negative attention bias can be mitigated via extensive behavioral training (Beevers et al., 2015). Moreover, this reduction in a negative attention bias was correlated with a reduction in depression severity, as measured through the Beck Depression Inventory (BDI).

One existing collaboration sought to ameliorate depression symptoms via reducing the

the negative attention bias in depressed individuals using closed-loop real-time fMRI neurofeedback (Schnyer et al., 2015), using an established protocol (deBettencourt et al., 2015). To adapt this protocol in the service of training depression, participants were asked to attend to the scene category while ignoring task-irrelevant faces. These faces were slightly modified to make them especially strong distractors for depressed individuals, by selecting sad faces from the Karolinska database (Lundqvist et al., 1998). Therefore, the closed-loop neurofeedback was intended to train attention to be sustained towards the neutral information and away from the negative information. Participants completed multiple sessions in the scanner, as well as pre and post behavioral sessions. This proof-of-concept study establishes the ability to adapt closed-loop neurofeedback for a clinical population.

4.3 Future directions

There are many lines of work in progress as we work towards a real-time understanding of cognition. Some existing collaborations seek to apply the techniques developed in this dissertation towards different aspects of cognition, including category learning, hippocampal encoding, memory weakening, and task space representations, among others. Additionally, given the temporal limitations of the hemodynamic signal, it seems promising to explore even more time-resolved feedback using EEG. All of these avenues are promising lines of work to help decipher the complexities of the brain in real time.

References

Al-Aidroos N., Said C.P. & Turk-Browne N.B. (2012) Top-down attention switches coupling between low-level and high-level areas of human visual cortex. *Proceedings of the National Academy of Sciences*, 109(36): pp. 14675–14680.

Aly M. & Turk-Browne N.B. (in press) How hippocampal memory shapes, and is shaped by, attention. In D.E. Hannula & M.C. Duff (Eds.), *The Hippocampus from Cells to Systems: Structure, Connectivity, and Functional Contributions to Memory and Flexible Cognition*. Springer, New York, NY.

Amano K., Shibata K., Kawato M., Sasaki Y. & Watanabe T. (2016) Learning to associate orientation with color in early visual areas by associative decoded fMRI neurofeedback. *Current Biology*, 26(14): pp. 1861–1866.

Baldauf D. & Desimone R. (2014) Neural mechanisms of object-based attention. *Science*, 344(6182): pp. 424–427.

Beevers C.G., Clasen P.C., Enock P.M. & Schnyer D.M. (2015) Attention bias modification for major depressive disorder: Effects on attention bias, resting state connectivity, and symptom change. *Journal of Abnormal Psychology*, 124(3): pp. 463–475.

Biggs A.T. & Mitroff S.R. (2014) Different predictors of multiple-target search accuracy between nonprofessional and professional visual searchers. *The Quarterly Journal of Experimental Psychology (Hove)*, 67(7): pp. 1335–1348.

Bosch S.E., Jehee J.F.M., Fernández G. & Doeller C.F. (2014) Different predictors of multiple-target search accuracy between nonprofessional and professional visual searchers. *The Journal of Neuroscience*, 34(22): pp. 7493–7500.

Brainard D.H. (1997) The psychophysics toolbox. *Spatial Vision*, 10(4): pp. 433–436.

Brewer J.B., Zhao Z., Desmond J.E., Glover G.H. & Gabrieli J.D. (1998) Making memories: brain activity that predicts how well visual experience will be remembered. *Science*, 281(5380): pp. 1185–1187.

Bubic A., von Cramon D.Y. & Schubotz R.I. (2010) Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, 4(25).

Büchel C., Coull J.T. & Friston K.J. (1999) The predictive value of changes in effective connectivity for human learning. *Science*, 283(5407): pp. 1538–1541.

Cerf M., Thiruvengadam N., Mormann F., Kraskov A., Quiroga R.Q., Koch C. & Fried I. (2010) On-line, voluntary control of human temporal lobe neurons. *Science*, 467(7319): pp. 1476–4687.

Chadick J.Z. & Gazzaley A. (2011) Differential coupling of visual cortex with default network or frontal-parietal network based on goals. *Nature Neuroscience*, 14(7): pp. 830–832.

Chiu Y.C. & Egner T. (2015a) Inhibition-induced forgetting results from resource competition between response inhibition and memory encoding processes. *The Journal of Neuroscience*, 35(34): pp. 11936–11945.

Chiu Y.C. & Egner T. (2015b) Inhibition-induced forgetting: when more control leads to less memory. *Psychological Science*, 26(1): pp. 830–832.

Chun M.M., Golomb J.D. & Turk-Browne N.B. (2011) A taxonomy of external and internal attention. *Annual Review of Psychology*, 62: pp. 73–101.

Cox R.W., Jesmanowicz A. & Hyde J.S. (1995) Real-time functional magnetic resonance imaging. *Magnetic Resonance in Medicine*, 33(2): pp. 230–236.

Craik F.I., Govoni R., Naveh-Benjamin M. & Anderson N.D. (1996) The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General*, 125(2): pp. 159–180.

Czeisler C.A., Walsh J.K., Roth T., Hughes R.J., Wright K.P., Kingsbury L., Arora S., Schwartz J.R.L., Niebler G.E. & Dinges D.F. (2005) Modafinil for excessive sleepiness associated with shift-work sleep disorder. *New England Journal of Medicine*, 353(2): pp. 476–486.

Davachi L. (2006) Item, context, and relational episodic encoding in humans. *Current Opinion in Neurobiology*, 16(6): pp. 693–700.

deBettencourt M.T., Cohen J.D., Lee R.F., Norman K.A. & Turk-Browne N.B. (2015) Closed-loop training of attention with real-time brain imaging. *Nature Neuroscience*, 18(3): pp. 470–475.

deBettencourt M.T. & Norman K.A. (2016) Neuroscience: Incepting associations. *Current Biology*, 36(14): pp. R673–R675.

deCharms R.C., Maeda F., Glover G.H., Ludlow D., Pauly J.M., Soneji D., Gabrieli J.D.E. & Mackey S.C. (2005) Control over brain activation and pain learned by using real-time functional MRI. *Proceedings of the National Academy of Sciences*, 102(51): pp. 18626–18631.

Dinges D.F. & Powell J.W. (1985) Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments & Computers*, 17(6): pp. 652–655.

Disner S.G., Beevers C.G., Haigh E.A.P. & Beck A.T. (2011) Neural mechanisms of the cognitive model of depression. *Nature Reviews Neurosciences*, 12(8): pp. 467–477.

Donaldson W. (1992) Measuring recognition memory. *Journal of Experimental Psychology: General*, 121(3): pp. 275–277.

Draganski B., Gaser C., Busch V., Schuierer G., Bogdahn U. & May A. (2004) Neuroplasticity: changes in grey matter induced by training. *Nature*, 427(6972): pp. 311–312.

Drew T. & Vogel E.K. (2008) Neural measures of individual differences in selecting and tracking multiple moving objects. *The Journal of Neuroscience*, 28(16): pp. 4183–4191.

Efron B. & Tibshirani R. (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1: pp. 54–75.

Eichenbaum H. & Cohen N.J. (2001) *From Conditioning to Conscious Recollection: Memory Systems of the Brain*. Oxford University Press, New York.

Epstein R. & Kanwisher N. (1998) A cortical representation of the local visual environment. *Nature*, 392(6676): pp. 598–601.

Esterman M., Noonan S.K., Rosenberg M. & Degutis J. (2013) In the zone or zoning out? Tracking behavioral and neural fluctuations during sustained attention. *Cerebral Cortex*, 23(1): pp. 2712–2723.

Fahle M. & Poggio T. (2002) *Perceptual learning*. MIT Press, Cambridge, MA.

Fleck M.S. & Mitroff S.R. (2007) Rare targets are rarely missed in correctable search. *Psychological Science*, 18(11): pp. 943–947.

Foerde K. & Shohamy D. (2011) The role of the basal ganglia in learning and memory: Insight from Parkinson's disease. *Neurobiology of Learning and Memory*, 96(4): pp. 624–636.

Folstein J.R., Palmeri T.J. & Gauthier I. (2013) Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex*, 23(4): pp. 814–823.

Frank M.J. & Badre D. (2012) Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. *Cerebral Cortex*, 22(3): pp. 509–526.

Gershman S.J., Schapiro A.C., Hupbach A. & Norman K.A. (2013) Neural context reinstatement predicts memory misattribution. *Journal of Neuroscience*, 33(20): pp. 8590–8595.

Goddard N.H., Hood G., Cohen J.D., Eddy W.F., Genovese C.R., Noll D.C. & Nystrom L.E. (1997) Online analysis of functional MRI datasets on parallel platforms. *Journal of Supercomputing*, 11(3): pp. 295–318.

Grill-Spector K., Henson R. & Martin A. (2006) Repetition and the brain: Neural models of stimulus-specific effects. *Trends in cognitive sciences*, 10(1): pp. 14–23.

Haxby J.V., Gobbini M.I., Furey M.L., Ishai A., Schouten J.L. & Pietrini P. (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539): pp. 2425–2430.

Hinds O., Thompson T.W., Ghosh S., Yoo J.J., Whitfield-Gabrieli S., Triantafyllou C. & Gabrieli J.D.E. (2013) Roles of default-mode network and supplementary motor area in human vigilance performance: evidence from real-time fMRI. *Journal of Neurophysiology*, 109(5): pp. 1250–1258.

Hsieh L.T., Gruber M.J., Jenkins L.J. & Ranganath C. (2014) Hippocampal activity patterns carry information about objects in temporal context. *Neuron*, 81(5): pp. 1165–1178.

Jiang Y. & Song J.H. (2005) Hyperspecificity in visual implicit learning: Learning of spatial layout is contingent on item identity. *Journal of Experimental Psychology: Human Perception & Performance*, 31(6): pp. 1439–1448.

Johnson K.A., Robertson I.H., Kelly S.P., Silk T.J., Barry E., D□ibhis A., Watchorn A., Keavey M., Fitzgerald M., Gallagher L., Gill M. & Bellgrove M.A. (2007) Dissociation in performance of children with ADHD and high-functioning autism on a task of sustained attention. *Neuropsychologia*, 45(10): pp. 2234–2245.

Kellough J.L., Beevers C.G., Ellis A.J. & Wells T.T. (2008) Time course of selective attention in clinically depressed young adults: an eye tracking study. *Behaviour Research and Therapy*, 46(11): pp. 1238–1243.

Kim G., Lewis-Peacock J.A., Norman K.A. & Turk-Browne N.B. (2014) Pruning of memories by context-based prediction error. *Proceedings of the National Academy of Sciences*, 111(24): pp. 8997–9002.

Kravitz A.V., Freeze B.S., Parker P.R.L., Kay K., Thwin M.T., Deisseroth K. & Kreitzer A.C. (2010) Regulation of parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry. *Nature*, 466(7306): pp. 622–626.

LaConte S.M. (2011) Decoding fMRI brain states in real-time. *Neuroimage*, 56(2): pp. 440–454.

Leber A.B. (2010) Neural predictors of within-subject fluctuations in attentional control. *Journal of Neuroscience*, 30(34): pp. 11458–11465.

Leber A.B., Turk-Browne N.B. & Chun M.M. (2008) Neural predictors of moment-to-moment fluctuations in cognitive flexibility. *Proceedings of the National Academy of Sciences*, 105(36): pp. 13592–13597.

Logothetis N.K., Pauls J., Augath M., Trinath T. & Oeltermann A. (2001) Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843): pp. 150–157.

Lorenz R., Monti R.P., Violante I.R., Anagnostopoulos C., Faisal A.A., Montana G. & Leech R. (2016) The automatic neuroscientist: A framework for optimizing experimental design with closed-loop real-time fmri. *Neuroimage*, 129: pp. 320–334.

Lundqvist D., Flykt A. & Öhman A. (1998) *The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience*. Psychology section, Karolinska Institutet. ISBN 91-630-7164-9.

MacInnes J., Dickerson K., kuei Chen N. & Adcock R.A. (2016) Cognitive neurostimulation: Learning to volitionally sustain ventral tegmental area activation. *Neuron*, 89(6): pp. 1331–1342.

Mackworth N.H. (1948) The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, 1: pp. 6–21.

Manning J.R., Hulbert J.C., Williams J., Piloto L., Sahakyan L. & Norman K.A. (2016) A neural signature of contextually mediated intentional forgetting. *Psychonomic Bulletin & Review*, (pp. 1–9).

Manning J.R., Norman K.A. & Kahana M.J. (2015) The role of context in episodic memory. In M. Gazzaniga (Ed.), *The Cognitive Neurosciences*. MIT Press, Cambridge, MA.

Manning J.R., Polyn S.M., Baltuch G.H., Litt B. & Kahana M.J. (2011) Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proc Natl Acad Sci*, 108(31): pp. 12893–12897.

McClelland J.L., McNaughton B.L. & O'Reilly R.C. (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3): pp. 419–457.

Miller E.K. & Cohen J.D. (2001) An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24: pp. 167–202.

Mishra J. & Gazzaley A. (2015) Closed-loop cognition: The next frontier arrive. *Trends in Cognitive Sciences*, 19(5): pp. 242–243.

Moore K.S., Yi D.J. & Chun M. (2013) The effect of attention on repetition suppression and multivoxel pattern similarity. *Journal of Cognitive Neuroscience*, 25(8): pp. 1305–1314.

Moray N. (1959) Attention in dichotic listening. *Quarterly Journal of Experimental Psychology*, 11(1): pp. 56–60.

Nichols T.E. & Holmes A.P. (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, 15(1): pp. 1–25.

Norman K.A., Polyn S.M., Detre G.J. & Haxby J.V. (2006) Beyond mind-reading: multivoxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9): pp. 424–430.

Noudoost B., Chang M.H., Steinmetz N.A. & Moore T. (2010) Top-down control of visual attention. *Top-down control of visual attention*, 20(2): pp. 183–190.

O'Craven K.M., Downing P.E. & Kanwisher N. (1999) fMRI evidence for objects as the units of attentional selection. *Nature*, 401(6753): pp. 584–587.

O'Reilly R.C. & Frank M.J. (2006) Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18(2): pp. 283–328.

Otten L.J., Quayle A.H., Akram S., Ditewig T.A. & Rugg2 M.D. (2006) Brain activity before an event predicts later recollection. *Nature Neuroscience*, 9(4): pp. 489–491.

Pagnoni G., Zing C.F., Montague P.R. & Berns G.S. (2002) Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, 5: pp. 97–98.

Papageorgiou T.D., Lisinski J.M., McHenry M.A., White J.P. & LaConte S.M. (2013) Brain-computer interfaces increase whole-brain signal to noise. *Proceedings of the National Academy of Sciences*, 110(33): pp. 13630–13635.

Parra L.C., Spence C.D., Gerson A.D. & Sajda P. (2003) Response error correction-a demonstration of improved human-machine performance using real-time EEG monitoring. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2): pp. 173–177.

Pelli D.G. (1997) The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, 10(4): pp. 437–442.

Pernet C.R., Wilcox R. & Rousselet G.A. (2013) Robust correlation analyses: false positive and power validation using a new open source matlab toolbox. *Frontiers in Psychology*, 3: p. 606.

Phillips P.J., Wechsler H., Huang J. & Rauss P.J. (1998) The FERET database and evaluation procedure for face recognition algorithms. *Image and vision computing*, 16(5): pp. 295–306.

Poldrack R.A., Clark J., Paré-Blagoev E.J., Shohamy D., Creso Moyano J., Myers C. & Gluck M.A. (2001) Interactive memory systems in the human brain. *Nature*, 414(6863): pp. 546–550.

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

Polyn S.M., Norman K.A. & Kahana M.J. (2009) Task context and organization in free recall. *Neuropsychologia*, 47(11): pp. 2158–2163.

Ramirez S., Liu X., Lin P.A., Suh J., Pignatelli M., Redondo R.L., Ryan T.J. & Tonegawa S. (2009) Creating a false memory in the hippocampus. *Science*, 341(6144): pp. 387–391.

Ranganath C. (2010) Binding items and contexts the cognitive neuroscience of episodic memory. *Current directions in psychological science*, 19(3): pp. 131–137.

Ranganath C. & Ritchey M. (2012) Two cortical systems for memory-guided behaviour. *Nature reviews neuroscience*, 13(10): pp. 713–726.

Reddy L., Kanwisher N.G. & VanRullen R. (2009) Attention and biased competition in multi-voxel object representations. *Proceedings of the National Academy of Sciences*, 106(50): pp. 21447–21452.

Redelmeier D.A. & Tibshirani R.J. (1997) Association between cellular-telephone calls and motor vehicle collisions. *New England Journal of Medicine*, 336(7): pp. 453–458.

Robertson I.H., Manly T., Andrade J., Baddeley B.T. & Yiend J. (1997) 'oops!': Performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia*, 35(6): pp. 747–758.

Rosenberg M., Noonan S., DeGutis J. & Esterman M. (2013) Sustaining visual attention in the face of distraction: a novel gradual-onset continuous performance task. *Attention Perception & Psychophysics*, 75(3): pp. 426–439.

Rosenberg M.D., Finn E.S., Constable R.T. & Chun M.M. (2015) Predicting moment-to-moment attentional state. *NeuroImage*, 114: pp. 249–256.

Rudy J.W. & O'Reilly R.C. (2001) Conjunctive representations, the hippocampus, and contextual fear conditioning. *Cognitive, Affective and Behavioral Neuroscience*, 1: pp. 66–82.

Sadtler P.T., Kristin M Quick M.D.G., Chase S.M., Ryu S.I., Tyler-Kabara E.C., Yu B.M. & Batista A.P. (2014) Neural constraints on learning. *Nature*, 512(7515): pp. 423–426.

Sahakyan L. & Kelley C.M. (2002) A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(6): pp. 1064–1072.

Sarter M., Givens B. & Bruno J.P. (2001) The cognitive neuroscience of sustained attention: Where top-down meets bottom-up. *Brain Research Reviews*, 35(2): pp. 146–160.

Schafer R.J. & Moore T. (2011) Selective attention from voluntary control of neurons in prefrontal cortex. *Science*, 332(6037): pp. 1568–1571.

Schapiro A.C., Rogers T.T., Cordova N.I., Turk-Browne N.B. & Botvinick M.M. (2013) Neural representations of events arise from temporal community structure. *Nature Neuroscience*, 16(4): pp. 486–492.

Schnyer D.M., Beevers C.G., deBettencourt M.T., Sherman S.M., Cohen J.D., Norman K.A. & Turk-Browne N.B. (2015) Neurocognitive therapeutics: from concept to application in the treatment of negative attention bias. *Biology of Mood & Anxiety Disorders*, 5: p. 1.

Schwartz S., Maquet P. & Frith C. (2002) Neural correlates of perceptual learning: a functional MRI study of visual texture discrimination. *Proceedings of the National Academy of Sciences*, 99(26): pp. 17137–17142.

Serences J.T., Schwarzbach J., Courtney S.M., Golay X. & Yantis S. (2004) Control of object-based attention in human cortex. *Cerebral Cortex*, 14(12): pp. 1346–1357.

Serruya M.D., Hatsopoulos N.G., Paninski L., Fellows M.R. & Donoghue J.P. (2002) Brain-machine interface: Instant neural control of a movement signal. *Nature*, 416(6877): pp. 141–1142.

Shallice T., Marzocchi G.M., Coser S., Del Savio M., Meuter R.F. & Rumiati R.I. (2002) Executive function profile of children with attention deficit hyperactivity disorder. *Developmental Neuropsychology*, 21(1): pp. 43–71.

Shibata K., Watanabe T., Sasaki Y. & Kawato M. (2011) Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Science*, 334(6061): pp. 1413–1415.

Smith S.M. & Nichols T.E. (2009) Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, 44(1): pp. 83–98.

Smith S.M. & Vela E. (2001) Environmental context-dependent memory: a review and meta-analysis. *Psychonomic Bulletin & Review*, 8(2): pp. 203–220.

Stadler M.A. & Frensch P.A. (1998) *Handbook of implicit learning*. Sage Publications, Inc, London.

Staudigl T., Vollmar C., Noachtar S. & Hanslmayr S. (2015) Temporal-pattern similarity analysis reveals the beneficial and detrimental effects of context reinstatement on human memory. *The Journal of Neuroscience*, 35(13): pp. 5373–5384.

Stoeckel L.E., Garrison K.A., Ghosh S.S., Wighton P., Hanlon C.A., Gilman J.M., Greer S., Turk-Browne N.B., deBettencourt M.T., Scheinost D., Craddock C., Thompson T., Calderon V., Bauer C.C., George M., Breiter H.C., Whitfield-Gabrieli S., Gabrieli J.D., LaConte S.M., Hirshberg L., Brewer J.A., Hampson M., Van Der Kouwe A., Mackey S. & Evins A.E. (2014) Optimizing real time fMRI neurofeedback for therapeutic discovery and development. *Neuroimage Clinical*, 5: pp. 245–255.

Sulzer J., Haller S., Scharnowski F., Weiskopf N., Birbaumer N., Blefari M.L., Bruehl A.B., Cohen L.G., deCharms R.C., Gassert R., Goebel R., Herwig U., LaConte S., Linden D., Luft A., Seifritz E. & Sitaram R. (2013) Real-time fMRI neurofeedback: Progress and challenges. *Neuroimage*, 76: pp. 386–399.

Sur M. & Rubenstein J.L. (2005) Patterning and plasticity of the cerebral cortex. *Science*, 310(5749): pp. 805–810.

Todd M.T., Nystrom L.E. & Cohen J.D. (2013) Confounds in multivariate pattern analysis: Theory and rule representation case study. *Neuroimage*, 77: pp. 157–165.

Tulving E. & Schacter D.L. (1990) Priming and human memory systems. *Science*, 247(4940): pp. 301–306.

Turk-Browne N.B. (2013) Functional interactions as big data in the human brain. *Science*, 342(6158): pp. 580–584.

Turk-Browne N.B. & Scholl B.J. (2009) Flexible visual statistical learning: Transfer across space and time. *Journal of Experimental Psychology: Human Perception and Performance*, 35(1): pp. 195–202.

Turk-Browne N.B., Scholl B.J. & Chun M.M. (2008) Babies and brains: Habituation in infant cognition and functional neuroimaging. *Frontiers of Human Neuroscience*, 2.

Uncapher M.R., Hutchinson J.B. & Wagner A.D. (2011) Dissociable effects of top-down and bottom-up attention during episodic encoding. *The Journal of Neuroscience*, 31(35): pp. 12613–12628.

Wagner A.D., Schacter D.L., Rotte M., Koutstaal W., Maril A., Dale A.M., Rosen B.R. & Buckner R.L. (1998) Building memories: remembering and forgetting of verbal experiences as predicted by brain activity. *Science*, 281: pp. 1188–1191.

Weiskopf N., Mathiak K., Bock S.W., Scharnowski F., Veit R., Grodd W., Goebel R. & Birbaumer N. (2004) Principles of a brain-computer interface (BCI) based on real-time functional magnetic resonance imaging (fMRI). *IEEE Transactions on Biomedical Engineering*, 51(6): pp. 966–970.

Weissman D.H., Roberts K.C., Visscher K.M. & Woldorff M.G. (2006) The neural bases of momentary lapses in attention. *Nature Neuroscience*, 9(7): pp. 971–978.

Wiecki T.V. & Frank M.J. (2013) A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological Review*, 120(2): pp. 329–355.

Wolfe J.M., Horowitz T.S. & Kenner N.M. (2005) Cognitive psychology: rare items often missed in visual searches. *Nature*, 435(7041): pp. 439–440.

Woolgar A., Hampshire A., Thompson R. & Duncan J. (2011) Adaptive coding of task-relevant information in human frontoparietal cortex. *The Journal of Neuroscience*, 31(41): pp. 14592–14599.

Xiao J., Hays J., Ehinger K.A., Oliva A. & Torralba A. (2010) SUN database: Large-scale scene recognition from abbey to zoo. *IEEE Conference on Computer Vision and Pattern Recognition*.

Xiao L.Q., Zhang J.Y., Wang R., Klein S.A., Levi D.M. & Yu C. (2008) Complete transfer of perceptual learning across retinal locations enabled by double training. *Current Biology*, 18(24): pp. 1922–1926.

Yi D.J. & Chun M.M. (2005) Attentional modulation of learning-related repetition attenuation effects in human parahippocampal cortex. *The Journal of Neuroscience*, 25(14): pp. 3593–3600.

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

Yoo J.J., Hinds O., Ofen N., Thompson T.W., Whitfield-Gabrieli S., Triantafyllou C. & Gabrieli J.D.E. (2012) When the brain is prepared to learn: enhancing human learning using real-time fMRI. *Neuroimage*, 59(1): pp. 846–852.

Yoss R.E., Moyer N.J. & Hollenhorst R.W. (1970) Pupil size and spontaneous pupillary waves associated with alertness, drowsiness, and sleep. *Neurology*, 20(6): pp. 545–554.

Zatorre R.J., Fields R.D. & Johansen-Berg H. (2012) Plasticity in gray and white: neuroimaging changes in brain structure during learning. *Nature Neuroscience*, 15(4): pp. 528–536.