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

2020

Exploring the potential of real-time fMRI to change thought trajectories

Mennen, Anne Caitlin

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp019306t222g

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

Exploring the potential of real-time fMRI to change thought trajectories

Anne C. Mennen

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

Adviser: Kenneth A. Norman

June 2020

© Copyright by Anne C. Mennen, 2020. All rights reserved.

Abstract

Real-time functional magnetic resonance imaging (rt-fMRI) is a non-invasive imaging technique with the potential to help us understand and train both normal and abnormal brain functioning. As fMRI records whole-brain, spatially-localized, patterns of activity in response to stimuli, neurofeedback capitalizes on this data by adjusting stimuli to drive neural responses. When applied to the clinical domain, rt-fMRI can provide specific training to the region or cognitive process thought to be deviant from healthy subjects. Given the potential of rt-fMRI, exploration must be done to develop new methods and uncover inherent limitations of this technique. This thesis represents work done to realize the potential of rt-fMRI for driving specific cognitive changes. While the current application of rt-fMRI is primarily focused on the treatment of depression, the experimental techniques can similarly be applied to other clinical populations.

Chapter 2 reviews differences in how depressed and control subjects allocate attention. Attention encapsulates many cognitive processes, from initial encoding to sustained maintenance of stimuli. The review highlights the variability in results that are dependent upon the task used. This motivates the exploration of different real-time experimental designs and tasks to improve mental health; two such projects are detailed in Chapters 3 and 4.

The study described in Chapter 3 directly aims to reduce depression severity through attention training, while the experiments presented in Chapter 4 explore the possibility of modifying story interpretations. In parallel, we developed an open-source cloud-based software framework for real-time processing. Our pipeline and experiments can be deployed at other institutions, with the hope that this encourages growth of the rt-fMRI community.

Altogether, the work presented in this thesis tests the feasibility of rt-fMRI to be used to alter thought through neurofeedback. The findings in the respective experiments help us to understand the capabilities of this technique, while the cloud software facilitates future use.

Contents

Abstract iii

Acronyms ix

1 Introduction 1
1.1 fMRI: Mapping thoughts . . . 2
1.2 Real-time fMRI neurofeedback . . . 3
1.2.1 General advantages . . . 3
1.2.2 Clinical applications . . . 5
1.2.3 Gaps in the rt-fMRI field . . . 5
1.3 How this thesis addresses the gaps . . . 7
1.4 Cloud computing . . . 9
1.5 Summary . . . 11

2 Negative attentional bias in depression 12
2.1 Introduction . . . 13
2.2 Internal Attention . . . 14
2.2.1 Cognitive control . . . 14
2.2.2 Emotion regulation . . . 16
2.3 External attention . . . 17
2.3.1 Spatial attention . . . 17
2.3.2 Disengagement . . . 19
2.4 Neurofeedback training . . . 21
2.5 Conclusions . . . 25

3 Closed-loop neurofeedback to improve negative attentional bias in depression 27
3.1 Introduction . . . 28
3.1.1 Current study . . . 30

3.1.2 Hypotheses . . . 31
3.2 Methods . . . 36
3.2.1 Participants . . . 36
3.2.2 Stimuli . . . 38
3.2.3 Procedure . . . 40
3.2.4 Data acquisition . . . 45
3.2.5 Real-time analyses . . . 46
3.2.6 Offline processing . . . 50
3.2.7 Primary analyses . . . 52
3.2.8 Secondary analyses . . . 55
3.3 Results . . . 60
3.3.1 Primary analyses . . . 60
3.3.2 Secondary analyses . . . 67
3.4 Discussion . . . 73

4 Real-time neurofeedback to alter interpretations using naturalistic stimuli **78**
4.1 Introduction . . . 79
4.1.1 Current study . . . 80
4.2 Experiment 1 . . . 81
4.2.1 Methods . . . 82
4.2.2 Results . . . 93
4.2.3 Discussion . . . 99
4.3 Experiment 2 . . . 100
4.3.1 Methods . . . 100
4.3.2 Results . . . 106
4.3.3 Discussion . . . 116
4.4 Discussion . . . 118

5 Conclusion and future directions **121**
5.1 Lessons learned . . . 122
5.2 Future directions . . . 123
5.3 Final conclusion . . . 124

Appendix A Supplementary Materials: Chapter 3 **125**
A.1 Correlation with specific networks . . . 125
A.1.1 Results . . . 126

Appendix B Supplementary Materials: Chapter 4 **131**
B.1 Designing the classifier 131
B.1.1 Neural data 132
B.1.2 Classification approach 132
B.1.3 Decision 1: how to preprocess the data 134
B.1.4 Decision 2: which time points in the story to use for neurofeedback 141
B.2 Instructions for Experiment 2 142

References **145**

Listing of figures

1.1 Real-time cloud diagram 10

2.1 Neural circuit for depression 15
2.2 Eye-tracking data showing the negative bias in depression 19
2.3 Example neurofeedback paradigm with depressed subjects 23

3.1 Gaze task trial structure 41
3.2 Go/no-go task stimulus conditions 42
3.3 Go/no-go task trial structure 43
3.4 Face-matching task trial structure 44
3.5 Go/no-go task run conditions 45
3.6 Cloud configuration 49
3.7 Transfer function 49
3.8 NF: Histogram showing customized attentional state bins 53
3.9 MADRS depression severity changes 60
3.10 NF: Average classification 61
3.11 NF: Sustained attention in different states 63
3.12 MADRS versus sustained negative attention 64
3.13 Gaze task: initial orientation 65
3.14 Gaze task: total viewing and maintenance 67
3.15 Go/no-go task: behavior 68
3.16 Face-matching task: LA response 70
3.17 Node connectivity results 72
3.18 Connectivity between FPN and LA 73

4.1 Experiment 1: Procedure 86
4.2 Cloud configuration 90
4.3 Experiment 1: Comprehension scores 94
4.4 Experiment 1: Interpretation scores 95

4.5 Experiment 1: Empathy ratings 96
4.6 Experiment 1: Interpretation versus empathy 97
4.7 Experiment 1: Classification results 97
4.8 Experiment 1: Neurofeedback scores 98
4.9 Experiment 2: Procedure 104
4.10 Experiment 2: Comprehension scores 106
4.11 Experiment 2: Interpretation scores 107
4.12 Experiment 2: Empathy ratings 108
4.13 Experiment 2: Interpretation versus empathy 109
4.14 Experiment 2: Probe responses 109
4.15 Experiment 2: Neurofeedback scores 111
4.16 Experiment 2: Classification divided by probe response 112
4.17 Experiment 2: Interpretation score versus classification 113
4.18 Experiment 2: Comprehension scores divided by classification accuracy . . 114
4.19 Experiment 2: Interpretation scores divided by classification accuracy . . . 115
4.20 Experiment 2: Empathy difference divided by classification accuracy 116
4.21 Experiment 2: Probe choices divided by classification accuracy 117

A.1 Network contributions: $\Delta A'$ versus network correlations 127
A.2 Network contributions: $\Delta A'$ versus Δnetwork correlations 129
A.3 Network contributions: ΔSustained negative attention versus Δnetwork correlations 129
A.4 Network contributions: ΔSustained correct attention versus Δnetwork correlations 130

B.1 Bootstrap pipeline 135
B.2 Bootstrap pipeline results 140
B.3 Selected classifier results over individual TRs 142
B.4 Selected classifier results across different classifier types 143
B.5 Experiment 2 instructions 144

Acronyms

ABM attention bias modification. 18, 28, 79

ACC anterior cingulate cortex. 15, 22, 74

ACPC anterior commissure-posterior commissure. 45, 85

BIDS Brain Imaging Data Structure. 50

BOLD blood-oxygen-level-dependent. 2, 9, 88, 89

BrainIAK Brain Imaging Analysis Kit. 10, 134

CBM-I cognitive bias modification for interpretation. 17, 18, 28, 79, 80

CBT cognitive behavioral therapy. 74, 75

CCT cognitive control training. 4, 16, 24

CFD Chicago Face Database. 39

CNDS Center for Neuromodulation in Depression and Stress. 37, 40

CSF cerebrospinal fluid. 50, 57, 88

DLPFC dorsolateral prefrontal cortex. 14, 15

DMN default mode network. 36, 58, 70–72, 130, 136

DMPFC dorsomedial prefrontal cortex. 136

EEG electroencephalography. 4, 24, 75

fMRI functional magnetic resonance imaging. iii, 1, 2, 4, 5, 9, 10, 21, 22, 24, 29, 75, 80, 81, 132, 137

FPN fronto-parietal network. vii, 36, 58, 59, 70–73

FU follow-up. 32

FWHM full-width half-maximum. 47, 56, 58

GM gray-matter. 50, 57, 88

HC healthy control. 36, 37, 47, 60–65, 67, 68, 70, 72, 73, 127, 129, 130

HRF hemodynamic response function. 53, 69

IAPS International Affective Picture System. 38

ISC intersubject correlation. 119

KDEF Karolinska Directed Emotional Faces. 39

LA left amygdala. vii, 57, 59, 68–71, 73

lIFG left inferior frontal gyrus. 136

lSTS superior temporal sulcus. 136

MADRS Montgomery-Åsberg Depression Rating Scale. vii, 32, 37, 40, 52, 60, 64, 66, 69, 70

MDD major depressive disorder. 37, 47, 57, 60–73, 76, 126–130

MNI Montreal Neurological Institute. 46, 51, 56, 57, 84–87, 89–91, 104, 126, 136

MPRAGE magnetization-prepared rapid acquisition gradient-echo. 46, 85

MRI magnetic resonance imaging. 37

NAPS Nencki Affective Picture System. 38

NF neurofeedback. vii, 32, 45, 65, 70, 91, 128, 130

rIFG right inferior frontal gyrus. 136

ROI region of interest. 23, 30, 46, 47, 57, 59, 68, 71, 74, 75, 117, 125–127, 129, 130, 133–136, 139, 140, 143

rt-cloud real-time cloud. vii, 10, 76, 89, 91, 120, 124

rt-fMRI real-time fMRI. iii, 2–11, 22, 29–32, 34, 45, 74, 80, 89, 99, 100, 110, 119, 122–124

SCID-5 Structured Clinical Interview for DSM-5 Disorders. 40

SCR skin-conductance response. 6

SRM shared response model. 132–135, 137–139

SVM support vector machine. 105, 139

tDCS transcranial direct current stimulation. 24

TOM theory of mind. 136, 139, 140, 143

TPJ temporoparietal junction. 136

TR repetition time. 9, 46, 48, 49, 52–54, 56–58, 69, 85, 91, 133, 137–143

VMPFC ventromedial prefrontal cortex. 136

wb whole-brain. 46, 125, 127, 129, 130

WM white-matter. 50, 57, 88

Acknowledgments

I COULD WRITE ANOTHER THESIS detailing my appreciation for the work, insight, and support from everyone who made this thesis possible.

First and foremost, I would like to thank my adviser, Ken Norman. Since our first meeting (ironically on Skype), his enthusiasm for science has inspired me. Six years later, I am still as eager to hear Ken share his ideas and/or his weekly Instacart musings. I cannot imagine how insurmountably difficult these experiments would have been had I not had such a caring, intelligent, and funny adviser to turn to weekly. Somehow, I would enter into Ken's office confused and frustrated and leave refreshed and excited to try new analyses. Above everything else, I am inspired by his commitment to being a kind person.

Next, I would like to thank the other members of my thesis committee for their continued help and insight throughout the years. Specifically, I thank Peter Ramadge for making the schlep down to PNI and Nick Turk-Browne for rolling his robot wheels over. In particular, I am very very thankful that Nick continued his involvement and invaluable support in our collaborations after he moved to Yale (and led me to Cameron/Nour and Natalia). Additionally, I am thankful for all that I have learned from Jon Cohen throughout my time at Princeton, from his advice during Intel meetings to all of his impassioned opinions.

I would like to thank Dr. Yvette Sheline for her unwavering research guidance and mentorship in our depression project. I so admire your achievements, expertise, and leadership in your laboratory, and have appreciated being given the chance to join your lab momentarily. Thank you to all CNDS lab members for welcoming me into the lab and particularly Adna, Darsol, Janet, Joyce, and Coco for helping with research.

To my collaborator, Grant Wallace, thank you for teaching me so much about coding and

being so dedicated to the fruition of our real-time cloud project. None of this would have been possible without your hard work to fix any of the issues that would arise unexpectedly. Thank you for persisting throughout all the inclement weather trips and frenzied Gchats.

Next, I would like to thank all of the PNI staff that has (quite literally) made all of this research possible. John Wiggins, Garret McGrath, Jawaan Lowery, Gene Piccola, and Randee Tengi – I don't know how I could possibly state how much I appreciate all of the crucial IT support you've given me over the past 6 years. To the PNI Scully Staff (Leigh Nystrom, Mark Pinsk, and Nick DePinto), thank you for tolerating things constantly breaking around me, fixing everything, and for giving me a reason to look forward to imaging meetings to catch up with all of you. Additionally, I would like to thank Penn Medicine's Mark Elliot for his continued help in setting up our experiment. To Alex Michaud and Ed Clayton, thank you both for being the people who can answer any question with a prompt solution and smiling face.

To Ben Goldberg and Dawn Tindall, thank you both for being the among the first at PNI to make me (and so many other new and nervous graduate students) feel welcome at Princeton. I will never forget your friendly faces and positive attitudes. May you both rest in peace.

Next, I would like to thank all of the Norman lab members. The personality of the Norman lab originally drew me into a rotation, and I was not let down in the slightest. From the OG members to the new additions, I have enjoyed getting to know all of you. You have been the smartest, silliest, and nicest group of people I have ever known, and have become my `compmem` family. Additionally, thank you particularly to all those who helped me with research – Megan, Neggin, Lizzie, Sam, Paula, Arlene, Jordan, Kevin, and Fred. Thank you to my thesis editors, Lizzie, Sam, Paula, and Manoj. I turn to you all because I have the utmost respect for you as scientists and friends. I want to give a special shout-out to my `Pygers` copilots (Sam, Lizzie, and Paula) for their selfless service, friendship, push ups, and tiger memes.

To all of the friends that I have met along the way during my time at Princeton, thank you so much for the laughs, hugs, scientific debates, GC brunches, tedious board games, workouts, weird mystery nights, concerts, glow parties, hikes, weddings, and everything else that people used to do before quarantine. To Jessie and Laura, thank you for being my curly hair sisters. To Nicole and Abby, thank you for being my non-curly hair sisters. To everyone involved in the skweekends/bweekends/schweekends, I can't wait to see you all soon (one

day).

I want to thank my family for being wonderful resources of fun and support along the way. I thank my siblings – Steph, Ethan, Elizabeth, and Lauren – for being my lifelong best friends and giving me the best motivation to work hard so that I can enjoy time with you all. Thank you to my parents for introducing me to science and giving me the love of learning. You two are the smartest and hardest-working people I know, and I hope it is clear how you have inspired me to make you proud. Thank you for always believing in me.

Last but not least, I want to thank my fiancé, Clay, for appearing in the right place at the right time and asking me to discuss my research over coffee. Finding you during graduate school will always be my greatest discovery.

1
Introduction

In fact, fMRI is not and will never be a mind reader, as some of the proponents of decoding-based methods suggest, nor is it a worthless and non-informative 'neophrenology' that is condemned to fail, as has been occasionally argued.

Nikos K Logothetis

1.1 fMRI: Mapping thoughts

Functional magnetic resonance imaging (fMRI) research has grown in complexity since the discovery that the blood-oxygen-level-dependent (BOLD) signal tracked neural responses to stimuli (Blamire et al., 1992; Ogawa et al., 1992). Initially, fMRI researchers localized where the brain processed information, such as categories (e.g.., faces in the fusiform face area; Grill-Spector et al., 2004) and emotions (e.g., fear in the amygdala; Armony & Dolan, 2002).

Since the initial focus on localization, neuroscientists have explored how mental states are represented as distributed patterns of activity across multiple regions (Haxby et al., 2001; Norman et al., 2006; Cohen et al., 2017). This updated framework increased the computational demands placed on researchers. As a result, fMRI researchers adopted machine learning tools that were suited for handling large data sets by finding signal in noisy, high-dimensional data with limited samples (e.g., Chen et al., 2015).

Following the advances in analysis techniques, fMRI researchers gained insight into cognitive processes such as memory (Polyn et al., 2005; Detre et al., 2013), communication (Hasson et al., 2012; Nguyen et al., 2019), decision-making (Schuck et al., 2016), and statistical learning (Schapiro et al., 2018). Growth in both fMRI expertise and computational power continues to drive the field forward. For instance, optimized algorithms on high-performance computing systems allow efficient calculation of correlations between every pair of volumetric pixels, or *voxels*, in the brain given a times series of fMRI data ($\approx 10^{10}$ samples; Wang et al., 2016; Cohen et al., 2017).

Capitalizing on the ability to measure complex brain states, fMRI researchers designed ways to harness, and try to control, functional activity. We discuss real-time fMRI (rt-fMRI),

first in its initial stages, and later how it is used in our research.

1.2 Real-time fMRI neurofeedback

While rt-fMRI is not a recent development (e.g., Cox et al., 1995; see reviews Sulzer et al., 2013; Cohen et al., 2017; Sitaram et al., 2017; Watanabe et al., 2017; Young et al., 2018; Thibault et al., 2018; Shibata et al., 2019), advances in computing hardware, scanning sequences, and analysis methods have led to significant expansion since the early implementations of this technique. Efficient computing systems have allowed rt-fMRI researchers to incorporate multiple neural regions into the neurofeedback signal, e.g., based on distributed patterns of activity (LaConte et al., 2007) or functional connectivity between two regions (Yamada et al., 2017; Ramot et al., 2017)

1.2.1 General advantages

The benefits of rt-fMRI are twofold: (1) rt-fMRI neurofeedback offers a direct measurement of cognition and (2) the feedback is customized to an individual's response.

In terms of the first advantage, using neural as opposed to behavioral feedback is a more direct way to manipulate underlying brain activity. Controlling the signal of interest directly allows feedback that is faster and more accurate. Of note, the recent behavioral training paradigms that adjusted stimuli in real-time have not yielded success consistently in clinical groups (e.g., using eye-movement and cognitive control training, as in Möbius et al., 2018 and Koster et al., 2017, respectively). While behavioral training is easier to deploy in clinical settings, behavior may be too far removed from the neural process of interest, and thus limited in the ability to manipulate the neural processes driving the behavior.

Previous research has suggested that fMRI is more sensitive to certain response differences between clinical and healthy groups. For instance, fMRI can reveal attentional biases in depression that are not captured with behavioral tasks alone. Discussed in Chapter 2 and Mennen et al. (2019), neural, but not behavioral, response differences are found between depressed and control subjects when negative stimuli are presented very quickly (Gotlib & Joormann, 2010; Mogg & Bradley, 2005; Suslow et al., 2010; Victor et al., 2013, 2010, 2012).

Additionally, neurofeedback with fMRI specifically has whole-brain recording capability, which allows the localization of subcortical structures. Compared to the less-costly electroencephalography (EEG)-based feedback, rt-fMRI can alter neural signals in fewer sessions (Thibault et al., 2018) and can localize regions with better resolution (Abreu et al., 2018; Meir-Hasson et al., 2014; Thibault et al., 2018; Kadosh & Staunton, 2019).

Customized feedback offers the flexibility to adjust to individual responses, unlike pre-programmed training. Adapting to a subject's performance in real-time can help to ensure continued participation in the task. Shortcomings in behavioral training used on depressed individuals may be in part due to the lack of engagement. Attention training tasks to reduce the negative bias in depression typically use a one-size-fits all approach, which may contribute to its mixed efficacy (e.g., see Hallion & Ruscio, 2011 and Jones & Sharpe, 2017). Additionally, the authors of Koster et al. (2017) concluded in their review of cognitive control training tasks that engagement predicted the clinical benefits (Siegle et al., 2014; Koster et al., 2017). Thus, even though rt-fMRI research requires more resources than other real-time methods (such as behavior or EEG), the benefits specific to using rt-fMRI warrant further exploration of this intensive technique.

1.2.2 Clinical applications

A sizable fraction of rt-fMRI experiments (18% of the studies included in Thibault et al., 2018) is dedicated specifically to clinical research, demonstrating widespread interest in harnessing rt-fMRI to improve mental health. Rt-fMRI is well-suited for clinical populations, as the customized training can adjust to heterogeneity across individuals.

Additionally, the spatial resolution of fMRI data allows researchers to study and regulate neural responses that are unique to the clinical condition (Stoeckel et al., 2014). Differences between clinical and healthy populations are well-documented in fMRI research; for instance, regarding depression, see Sheline et al. (2001, 2009, 2010); Siegle et al. (2007b); Yamada et al. (2017); Yu et al. (2019); Satterthwaite et al. (2016); Fales et al. (2008); Kaiser et al. (2015). Thus, established differences between clinical and healthy populations can serve as inspiration for future neurofeedback training.

As fMRI researchers continue to develop new ways to understand clinical conditions, rt-fMRI researchers can incorporate the state-of-the-art approaches into real-time training. For instance, Xia et al. (2018) determined specific behavioral symptoms that were the most highly correlated with pathological functional connectivity patterns. This data-driven approach to define functional *biomarkers* is one case where rt-fMRI neurofeedback can capitalize on this finding and develop specific treatments based on psychiatric symptoms.

1.2.3 Gaps in the rt-fMRI field

Although there has been major growth in rt-fMRI research, there are areas that warrant further exploration. A large portion of research seeks to regulate mean signal magnitude, instead of multivariate patterns of distributed activity. Decoded neurofeedback from individual task-

related multivariate patterns (e.g., food cravings) represented a minority of studies (12/99) included in the Thibault et al. (2018) review.

In addition to common processing pipelines, another limitation of rt-fMRI is a lack of robust behavioral findings, particularly after controlling for placebo and demand characteristics (Thibault et al., 2018). Out of the 41 studies reviewed in Thibault et al. (2018) that included a control group, only 24 out of 41 neurofeedback studies found stronger behavioral effects in the neurofeedback group. Maintaining significant results after including a well-suited control group is an important issue that the field must confront (Sorger et al., 2019).

One innovative example of rt-fMRI successfully regulating brain states *without* changing behavior is reported in Taschereau-Dumouchel et al. (2018). This study used neurofeedback to train phobic subjects to subconsciously associate reward with an animal they feared. The study successfully lowered the amygdala's response and skin-conductance response (SCR) to those animals. However, there was no report of lowering clinical measures of fear, which were subjectively reported. In the authors' more recent paper, Taschereau-Dumouchel et al. (2019), further analyses showed that the physiological responses to fear and subjective experience of fear were represented differently in the brain. As a result, the authors cautioned others against equating the two measurements, which has particular relevance for neurofeedback studies that target specific regions thought to be related to behavior.

By attempting to decode and subsequently regulate *cognitive states* with rt-fMRI, we hope to avoid the shortcomings of rt-fMRI research mentioned above. We hypothesize that regulating complex mental states can more robustly yield behavioral effects than regulating average regional activity or manipulating representations that are unrelated to the task (e.g.,

subconscious categorical representations). Below, we discuss how our studies provided individualized feedback at the moment our subjects were in a certain cognitive state (e.g., attending to a negative face).

Additionally, another branch of work in this thesis focused on a separate limiting factor in the rt-fMRI field: the large demand of resources required to set up and execute a neurofeedback study. To advance state-of-the-art infrastructure in this domain, we developed an open-source real-time cloud framework for easier sharing and execution of real-time software. We hope this software helps the community by reducing the technical barriers preventing widespread neurofeedback research.

1.3 How this thesis addresses the gaps

In each of the chapters to follow, we focus on how rt-fMRI neurofeedback can be used to reinforce one brain pattern to prevail over another.

First, in Chapter 2, we discuss clinical applications of attention training, as attentional allocation has been shown to differ between healthy and depressed participants (Kellough et al., 2008; Armstrong & Olatunji, 2012; Price et al., 2013; Sanchez et al., 2013; Joormann, 2004; Cisler & Koster, 2010). We explore one instance in detail: depressed subjects are more likely to attend to negative information. By characterizing the most robust differences in the attentional processes between depressed and control subjects, we can design more effective neurofeedback experiments.

In Chapter 3, we build on the existing literature and directly test the efficacy of rt-fMRI to 'unstick' depressed subjects from a sustained negative attentional state. In the experiment, we tracked brain signatures of attention through multivariate decoding of different image

categories. We then monitored how rt-fMRI training changed depressive symptoms, as well as other behavioral and neural markers related to depression. While rt-fMRI neurofeedback has been tested on depressed populations many times (Linden et al., 2012; Linden, 2014; Yuan et al., 2014; Young et al., 2014; Zotev et al., 2016; Mehler et al., 2018; Young et al., 2018; Thibault et al., 2018; MacDuffie et al., 2018; Yamada et al., 2017), this experiment differed in the training approach. First, we regulated multivariate representations of attention instead of mean signal magnitude. Second, the neurofeedback was *closed-loop*, in that the feedback display was integral to the task itself, instead of being a separate metric that subjects had to monitor alongside the task.

In Chapter 4, we describe another experiment aimed at regulating the interpretation of an ambiguous spoken story. As depressed subjects also have a tendency to interpret ambiguous information negatively (Everaert et al., 2017, 2018), we explored the potential of using rt-fMRI training to bias interpretations. Our study was based on Yeshurun et al. (2017), which set a precedent for decoding the interpretation of an ambiguous story (offline). We used the same story stimulus to train a classifier on the data reported in Yeshurun et al. (2017) to establish template brain states corresponding to the two different story interpretations, prior to recruiting real-time subjects. This project represents the first rt-fMRI experiment designed to manipulate the interpretation of a dynamical, naturalistic story.

Additionally, the specific application of our cloud-based real-time processing platform will be discussed in each corresponding chapter. Next, we discuss the background and motivation for its development.

1.4 Cloud computing

As mentioned above, most rt-fMRI studies train participants to regulate the mean BOLD signal magnitude within a particular region. Although multivariate-based methods generally yield greater sensitivity (Oblak et al., 2019), they are substantially more complex and computationally demanding. Thus, despite the widespread advancement of offline analyses, multivariate approaches have not been widely deployed in the real-time domain. This limitation is a major barrier to rt-fMRI research.

The core challenge in rt-fMRI research is that analyses must be completed rapidly in order to provide prompt feedback. These steps may include: detecting new fMRI images from the scanner within milliseconds of arrival, performing image registration, continually updating classification models over the course of a scanning run, translating classification results into feedback, updating the feedback display, and recording behavioral responses. If the neurofeedback signal is continuously updated (as in deBettencourt et al., 2015), all processing steps must be completed within a 1-2 second acquisition window for each volume or repetition time (TR). Even when intermittent feedback is used (as in Hellrung et al., 2018), the data processing rate remains high because larger batches of data are processed. The processing demands may then require cluster or cloud computing, which may only increase obstacles for researchers, as network communication adds more complexity.

A diverse skill set in software engineering is required in order to satisfy the technical requirements mentioned above. Our aim is to provide software to consolidate IT efforts, limit hardware costs, and reduce the expertise needed for doing rt-fMRI research. Without focused effort to expand the capabilities and ease of use, the field is at risk of stagnating and

never reaching its potential to understand and treat brain-based pathologies using fMRI.

With the adoption of an open-source real-time cloud (rt-cloud) framework, the field can benefit in several ways. The standardized pipeline allows easier sharing and replication of a given experiment. Using shared software and resources promotes cross-institutional collaboration and facilitates community maintenance and development of the code base. Additionally, experiments will be less expensive, as the cloud service relies on a pay-as-you-go model. In the clinical realm, lower costs and ease of use will be crucial to persuade medical institutions to adopt neurofeedback training protocols.

Both of the experiments detailed in Chapters 3 and 4 served as pilots for designing the rt-cloud framework, in collaboration with software engineer Grant Wallace, as part of the Brain Imaging Analysis Kit (BrainIAK; `https://brainiak.org`). Each experiment differed in rt-fMRI application, both in qualitative theoretical approach and quantitative method. The common use of our pipeline is a testament to the flexibility of our rt-cloud platform.

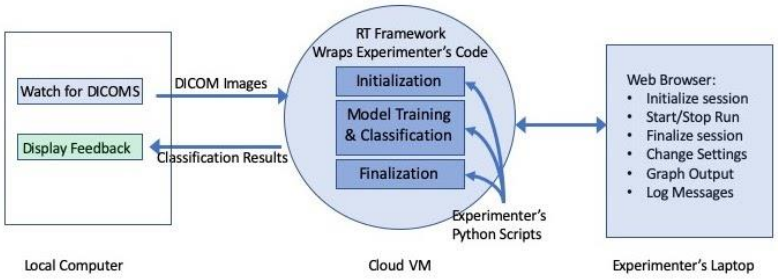


Figure 1.1: Diagram of cloud processing components. Reprinted with permission from Grant Wallace. **VM** = virtual machine; **RT** = real-time

1.5 Summary

Altogether, this thesis explores the potential of using rt-fMRI to modify thought trajectories. As thoughts are dynamic, distributed, and complex in nature, much of this work is devoted to properly quantifying the corresponding neural signal. In the following chapters, we discuss both the promising results and the need for follow-up studies to defend the specific contributions of neurofeedback. As we work to fulfill the latter requirement, cloud-based real-time processing will allow future experimental iterations to be executed efficiently.

2

Negative attentional bias in depression *

... [ATTENTION] ITSELF, THOUGH ITS GENESIS IS DOUBTLESS PERFECTLY NATURAL, MAKES EXPERIENCE MORE THAN IT IS MADE BY IT.

William James

One of the most common symptoms of depression is the tendency to attend to negative

*The contents of this chapter were published in Mennen et al. (2019)

stimuli in the world and negative thoughts in mind. This symptom is especially nefarious because it is also a cause — biasing processing to negatively valenced information, thus worsening mood, and exacerbating the condition. Here we attempt to systematize the diverse body of recent research on the negative attentional bias from across cognitive and clinical psychology in order to identify recurring themes and devise potential mechanistic explanations. We leverage theoretical progress in our understanding of healthy attention systems in terms of internal versus external components. With this lens, we review approaches to training attention that might reduce the negative attentional bias, including behavioral interventions and real-time neurofeedback. Although extant findings are somewhat mixed, these approaches provide hope and clues for the next generation of treatments.

2.1 Introduction

Negative attentional bias in depression — the tendency of depressed individuals to focus on negative stimuli and thoughts more than healthy individuals — has been the subject of research spanning multiple decades, much of it seeking to explain why the bias is not always observed (Gotlib & Joormann, 2010; Mogg & Bradley, 2005; Jones & Sharpe, 2017). This topic continues to inspire the development of new tasks and training paradigms that have leveraged insights from prior research and new technology. Here we review recent work in this field and explore how basic cognitive investigations of changes in perception and attention in depression may lead to progress in understanding the nature of the bias, and consequently, potential treatment approaches for ameliorating it.

2.2 Internal Attention

To begin characterizing the negative attentional bias and addressing inconsistencies in the literature, we leverage the framework of internal versus external attention (Chun et al., 2011). External attention involves allocating attention to stimuli in the current sensory environment, whereas internal attention operates over mental representations not directly supported by the environment, such as thoughts, emotions, memories, and tasks. Depressed patients perform differently from healthy individuals in both domains. In this section we address different forms of internal attention and related training approaches, before turning to external attention in the next section. First, we highlight two processes that involve attention to internal representations: cognitive control and emotion regulation.

2.2.1 Cognitive control

Relative to healthy individuals, depressed individuals show general deficits in executive functions such as rule shifting, verbal working memory, updating, planning, verbal fluency, psychomotor speed, and especially inhibition (Koster et al., 2017; Snyder, 2013). This behavioral difference is often attributed to prefrontal hypoactivity (see Snyder, 2013; Figure 2.1), as depressed individuals perform worse than healthy controls on cognitive control tasks when dorsolateral prefrontal cortex (DLPFC) activity is lower than controls, but not when DLPFC activity is greater than controls (Snyder, 2013; Holmes & Pizzagalli, 2007; Wagner et al., 2006). Impairments in cognitive control are less robust for at-risk subjects (those identified by a dysphoric mood and high tendency to ruminate) who are completing tasks with neutral materials, but they are consistently found in tasks employing negatively valenced stimuli (Koster et al., 2017). Additionally, deficits in memory performance are most apparent when tasks

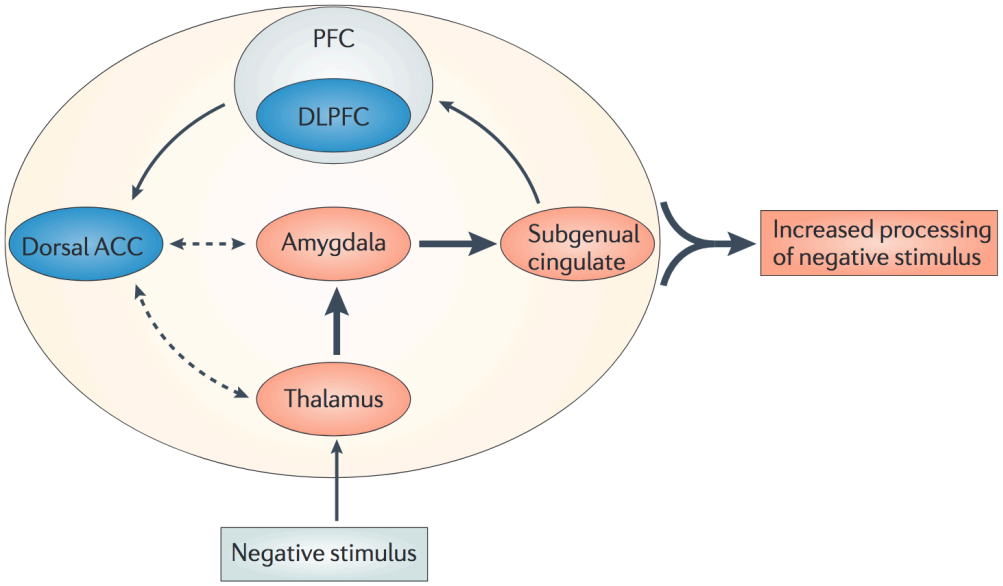


Figure 2.1: Neural circuit illustrating how both stimulus-driven and goal-directed pathways can drive the negative attentional bias in depressed subjects (dashed lines indicate weakened functional connections in depression). Negative stimuli trigger a cascade of increased activity (red) from thalamus $\rightarrow$ amygdala $\rightarrow$ subgenual cingulate. The DLPFC and thus ACC respond with reduced activity (blue), which in turn reduces inhibition of the amygdala and prolongs processing of negative stimuli. Reprinted with permission from Disner et al. (2011)

are relatively unconstrained by instructions, which is thought to allow depressed individuals to spontaneously ruminate on negative thoughts, especially about themselves (Gotlib & Joormann, 2010).

These executive deficits associated with depression led to the development of cognitive control training (CCT) — multiple training sessions on tasks devised to improve executive function by demanding inhibition of habitual responses. Tasks that require increased focus on neutral stimuli, such as sounds and numbers, reduce rumination (Siegle et al., 2007a) and depressive symptoms (Siegle et al., 2007a; Calkins et al., 2015), though not always in a generalized manner (Koster et al., 2017). A recent review that evaluated CCT across different depressed populations from subclinical to those in remission concluded that it was most effective when subjects had cognitive impairments before training but were also attentive to the task (Koster et al., 2017). Emotional stimuli are infrequently incorporated into CCT training, but this was shown to improve clinical results in one study (Iacoviello et al., 2014).

2.2.2 Emotion regulation

Depressed individuals show increased internal attention to negative representations of the past, which leads to extended rumination, ineffective cognitive reappraisal of negative situations (Gotlib & Joormann, 2010), and negative interpretations of affectively ambiguous scenarios (Gotlib & Joormann, 2010; Mogg & Bradley, 2005; Hallion & Ruscio, 2011). This latter negative interpretation bias tends to persist despite the revelation of new, positive information (Everaert et al., 2018) and is more pronounced when self-referential prompts are used (Everaert et al., 2017). Although depression is associated with attention deficits, it can also lead to better concentration on negative self-thought and improved memory for negative

information, in contrast to healthy individuals who are more likely to attend to and remember positive information (Gotlib & Joormann, 2010). A possible explanation for this finding is that depressed individuals are more likely to already be in a negative mood, and this congruent context makes negative information easier to remember (Gotlib & Joormann, 2010).

One approach for modifying the negative interpretation bias is to train patients to interpret ambiguous words or phrases in a positive way with cognitive bias modification for interpretation (CBM-I). CBM-I has been found to reduce the negative interpretation bias, though sometimes without transfer to other aspects of depression (LeMoult et al., 2018).

2.3 External attention

Tasks that elicit external attention — attention to one's immediate environment — also capture the negative bias in depression. Although these tasks invoke perceptual processes by presenting negative words and/or images, we do not assume that external attention is the *only* mechanism at play (e.g., these stimuli could trigger negative memories or rumination). Still, we distinguish these tasks from those discussed above for their separate designs, results, and training attempts. Specifically, we limit our focus to tasks that measure attention towards negative stimuli as well as inhibition of attention away from negative stimuli.

2.3.1 Spatial attention

Much of the research in this domain involves variants of the dot-probe task, in which a probe target immediately replaces one of two stimuli that differ in emotion (Gotlib & Joormann, 2010; Mogg & Bradley, 2005). Detection of this probe is faster when it replaces the stimulus being attended than when it replaces the unattended stimulus. In this way, negative bias can

be calculated by computing the difference in reaction time to detect a probe that replaces a negative stimulus versus a positive or neutral stimulus. Such biases are found more reliably when the images are presented for longer durations (e.g. 500–1000 ms; Gotlib & Joormann, 2010; Mogg & Bradley, 2005).

The dot-probe task has been adapted for a form of attention training known as attention bias modification (ABM). In ABM, the probe is placed behind the less-negative stimulus more often than the negative stimulus. Such training has been associated with decreases in depression severity (Beevers et al., 2015; Browning et al., 2012; Li et al., 2016a) and improvements in mood (Cooper et al., 2014). However, other studies found null results and identified potential methodological confounds, including using the same task for training and testing (Duque & Vazquez, 2018). Indeed, meta-analyses including studies using either ABM or CBM-I show that such training modifies the negative bias in depression inconsistently; CBM-I tends to yield larger effect sizes than those derived from ABM training, both for anxious and depressed samples analyzed together (Hallion & Ruscio, 2011) and analyzed separately (Cristea et al., 2015), whatever effect that exists is smaller for depressed than anxious individuals (Jones & Sharpe, 2017; Hallion & Ruscio, 2011) and not significant for either clinical population with outliers removed (Cristea et al., 2015), and the number of training sessions is inversely related to study efficacy (Cristea et al., 2015). Of note, the meta-analyses (Jones & Sharpe, 2017; Hallion & Ruscio, 2011; Cristea et al., 2015) included additional ABM paradigms apart from the dot-probe task, but the dot-probe task was the most prevalent (Jones & Sharpe, 2017). One possible explanation for these mixed results could be how researchers measure depressive symptoms, as the effects of training can be more robust if a stressor precedes the test (Hallion & Ruscio, 2011). Moreover, tasks that rely on spatial

probes to measure reaction times at the end of a trial cannot distinguish the roles of specific attention mechanisms (Duque & Vazquez, 2018); this is better accomplished by studies designed to capture inhibitory processes, discussed next.

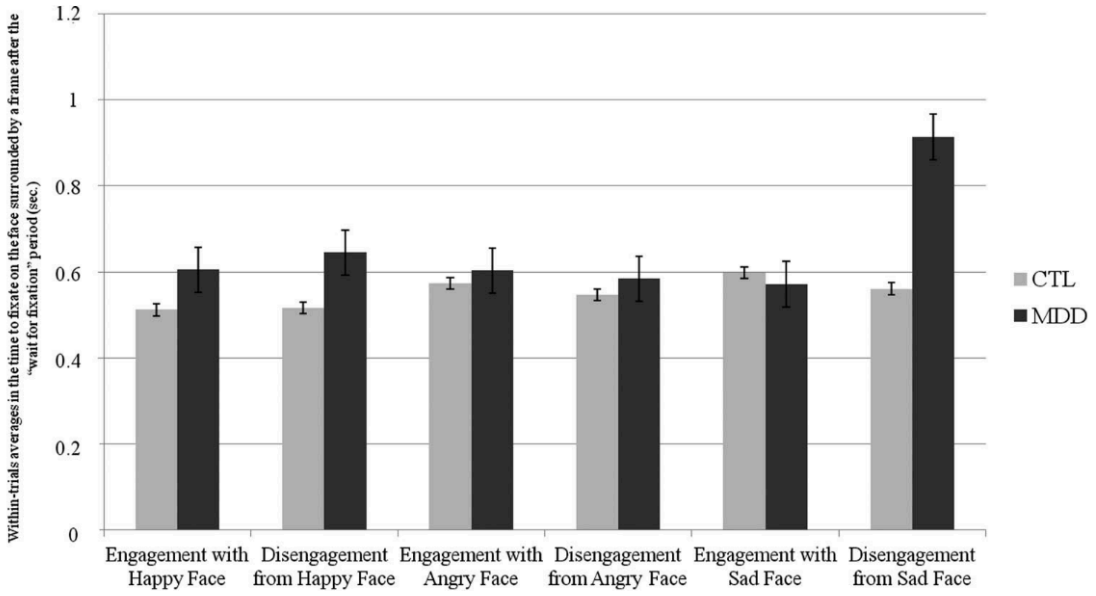


Figure 2.2: Eye-tracking data showing reaction times after either attending toward or away from an emotional face (paired with a neutral face), in both depressed **(MDD)** and control **(CTL)** participants. The results clearly show that the main impairment in depressed patients involves disengaging from sad faces. Reprinted with permission from Sanchez et al. (2013)

2.3.2 Disengagement

Where and when people look during natural viewing of emotional images provides a more continuous readout of behavior that can help quantify and distinguish different components of external attention, such as initial selection, sustained attention to relevant stimuli, and disengagement from irrelevant stimuli (Duque & Vazquez, 2018; Kellough et al., 2008; Eizenman et al., 2003; Caseras et al., 2007). Indeed, eye-tracking studies have shown that depres-

sion is associated with slower disengagement from negative stimuli (see Sanchez et al., 2013; Figure 2.2). That is, depressed and non-depressed individuals are equally likely to initially orient towards negative stimuli, but once these stimuli are in the focus of attention, depressed patients struggle to reorient attention elsewhere (Caseras et al., 2007). A meta-analysis of eye-tracking studies confirmed this finding and further showed that the bias for the initial selection component of external attention reflects reduced selection of positive stimuli in depression rather than increased selection of negative stimuli (Armstrong & Olatunji, 2012).

Other behavioral data point to a problem in inhibiting negative information (Everaert et al., 2017), which may prevent depressed individuals from being able to disengage. One way to measure inhibition is in terms of *negative priming* — the tendency to respond more slowly to an item after having ignored it on a previous trial when it served as a distractor. (Note: 'negative' in negative priming does not have any relation to valence, but is rather used to indicate that the effect is opposite to the typical facilitation by priming.) Larger negative priming indicates successful inhibition and smaller (or nonexistent) negative priming indicates unsuccessful inhibition. Compared to controls, depressed individuals show less negative priming when attending to negatively valenced words and images after having just ignored them as distractors. This suggests a failure to disengage from these stimuli, a tendency that was related to rumination (Gotlib & Joormann, 2010). This disengagement deficit in depression can be found for neutral stimuli as well (Gotlib & Joormann, 2010), consistent with reduced inhibition in depression more generally (Golomb et al., 2009). The attentional differences are not always detrimental: depressed individuals responded faster than controls on a go/no-go task when responding to negative words and ignoring positive words (Erickson et al., 2005).

Given the successful demonstration of negative biases with eye movements, eye-tracking has also been used for attention training, often by guiding overt external attention toward less negative stimuli (Vazquez et al., 2016; Ferrari et al., 2016; Möbius et al., 2018). These paradigms train subjects to look towards more positive stimuli by ending trials only after the participant has fixated on the desired stimulus for a minimum amount of time. Although eye movement training has succeeded in reducing time spent viewing negative stimuli, training has not yet been able to transfer to depression-related symptoms such as mood after a stress-inducing task (Ferrari et al., 2016; Möbius et al., 2018). Another promising training approach involves using eye-tracking to deliver real-time feedback about where a person is looking. In one such study (Sanchez et al., 2016), participants saw positive and negative words and had to unscramble them to form a positive sentence. Their fixations caused the font color to change to either green or red depending on whether they were looking at positive or negative words, respectively. The purpose of this was to help the subjects stay focused on the positive, with the goal of modifying their negative interpretation bias. Eye-tracking feedback might be more effective than merely guiding the eyes away from negative stimuli because incorporating the feedback might engage internal attention processes. Consistent with this, the benefit of eye-based feedback training for reducing the negative attentional bias was mediated by changes in cognitive control (Sanchez et al., 2016).

2.4 Neurofeedback training

Another class of cutting-edge training approaches does not fit cleanly under internal or external attention, and so we discuss it here separately. These approaches incorporate neural signals into the training, based on findings that fMRI can reveal signatures of the negative at-

tentional bias in depressed individuals not apparent in their behavior (Gotlib & Joormann, 2010; Mogg & Bradley, 2005; Suslow et al., 2010; Victor et al., 2013, 2010, 2012). We are not referring to using neural signals as a dependent measure of successful training or symptom change, as has been done in studies that collected fMRI scans before and after behavioral training (Beevers et al., 2015; Li et al., 2016a; Hilland et al., 2019; Wiers & Wiers, 2017). Rather, to the extent that specific brain regions are responsible for the negative attentional bias, directly training these regions with neurofeedback from real-time fMRI (rt-fMRI) may induce behavioral improvement. This implicates both internal and external attention because the neurofeedback is based on internal states of the brain related to tasks, goals, and strategies, but it is delivered as a salient visual stimulus in order to externalize these states and influence subsequent processing.

Rt-fMRI neurofeedback is an emerging field, but there is already substantial research that demonstrates the feasibility of decoding and altering different aspects of cognition. Monitoring attentional states based purely on neural data has proven possible in healthy subjects. One study found performance differences after trials were triggered in real-time during 'good' or 'bad' neural attentional states (Hinds et al., 2013). Another study aimed to reduce the extent to which depressed individuals attended to negative distractors (Schnyer et al., 2015), by decoding and providing feedback on how they oriented attention away from negative faces and towards neutral scenes. Other clinically motivated studies trained depressed subjects to regulate specific brain regions in real-time while performing a task. Examples of training paradigms and their respective clinical success include training depressed subjects to: (1) decrease anterior cingulate cortex (ACC) activity while performing individualized cognitive behavioral therapy strategies on negative autobiographical memories, which predicted self-

reported efficacy (post-training) of the cognitive behavioral strategy practiced in real-time (MacDuffie et al., 2018); (2) increase amygdala activity while recalling positive autobiographical memories, which reduced depressive symptoms (see Young et al., 2017, 2018; Figure 2.3); and (3) decrease salience network activity while viewing negative images, which lowered self-reported responses to negative stimuli (Hamilton et al., 2016).

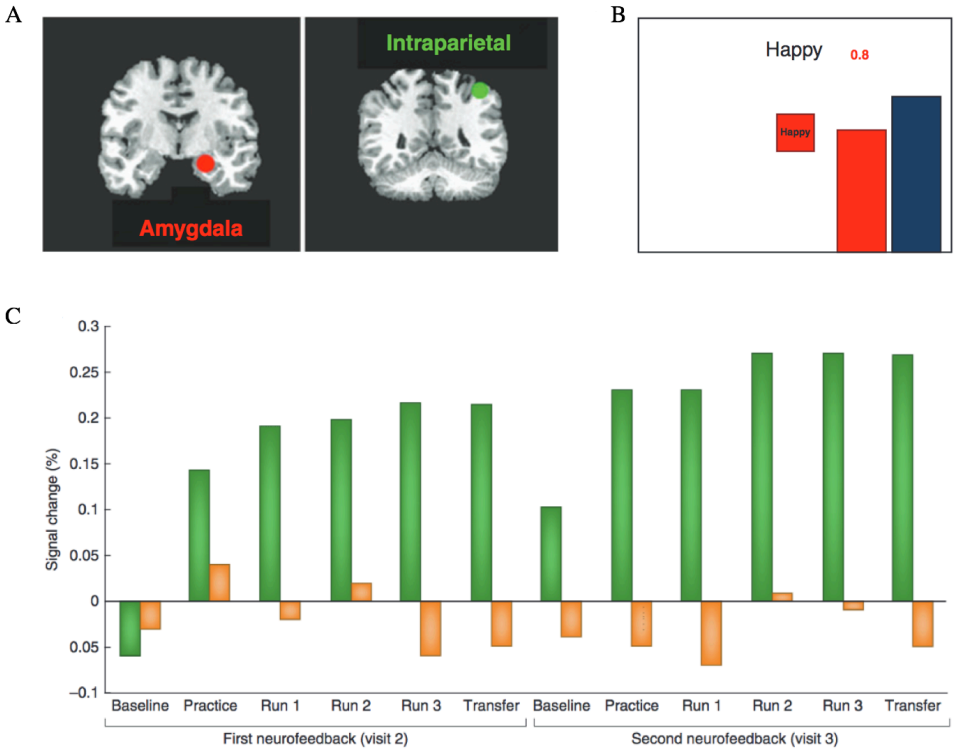


Figure 2.3: Neurofeedback training adapted with permission from Young et al. (2018). **(a)** Two neurofeedback target ROIs: left amygdala for the experimental group and left horizontal segment of the intraparietal sulcus for the control group. **(b)** Example feedback display when subjects think of happy memories and try to increase the red bar (signaling ROI activity) to the target blue bar height. **(c)** Results adapted with permission from Young et al. (2017) showing a change in amygdala activity over training for the experimental group **(green)** and not the control group **(orange)**, matching a behavioral improvement (lower depression severity scores) specific to the experimental group. *Reprinted with permission from the American Journal of Psychiatry (Copyright © 2017). American Psychiatric Association. All rights reserved.*

Neurofeedback based on functional connectivity (i.e. correlation in activity over time between brain regions) is also a promising avenue for attention training, as changes in functional connectivity are associated with reduction in depressive symptoms. For example, upregulating amygdala activity increases connectivity between the left amygdala and left cuneus in a way that predicts reduction of depressive symptoms (Yuan et al., 2014). More generally, attention training may increase connectivity within frontal areas and decrease connectivity between frontal and insular areas (Wiers & Wiers, 2017). Successful treatment has been linked to activity within the insular cortex itself (Hilland et al., 2019) and to connectivity between precuneus and medial frontal gyrus (Beevers et al., 2015). Even in healthy subjects, training on a neutral flanker task can increase resting-state connectivity between the amygdala and right inferior frontal gyrus (Cohen et al., 2016). This change in connectivity after cognitive control training could motivate future research to improve cognitive control through neurofeedback that rewards increased connectivity between these regions.

Although fMRI provides unparalleled access to brain states, other neurofeedback technologies have the advantage of being more portable and less costly. For example, the combination of CCT and transcranial direct current stimulation (tDCS) reduces depression symptoms even lasting through a follow-up visit (Segrave et al., 2014). This benefit was greater than either CCT or tDCS alone, highlighting potentially unique contributions of behavioral and neural attention training. Another approach to translating fMRI neurofeedback to more portable techniques like EEG is to first perform simultaneous fMRI and EEG to derive an EEG correlate of activity only visible to fMRI, such as from the amygdala (Meir-Hasson et al., 2014). This EEG correlate can then be used for neurofeedback outside of the scanner, for example, to train soldiers to downregulate amygdala activity in response to a computerized

stressful situation in preparation for combat (Keynan et al., 2019).

Although not all of the clinical studies mentioned above sought to train attention directly, they offer insight into neural manipulations that could be used to obtain a desired clinical outcome. In the future, internal or external attention could be selectively trained through deliberate design of the task and by targeting specific brain regions. For example, one could use paradigms that load heavily on internal attention, such as reappraising a negative memory to a positive one as in (MacDuffie et al., 2018), while targeting neural measures associated with internal attention, such as frontal activity, fronto-limbic connectivity, or decoded emotional states (Li et al., 2016b). To train external attention away from negative stimuli, one could present images that vary in terms of image category and valence (e.g. negative faces and neutral scenes as in Schnyer et al., 2015), and use a category-based classifier of neural activity to identify, and provide neurofeedback about, how much the negative stimulus is being attended (e.g., how much face evidence there is in the brain).

2.5 Conclusions

Depression and other mental health disorders are associated with biases and deficits in attention. Our focus in this review was on the negative attentional bias. We synthesized diverse behavioral and neural findings using modern theoretical frameworks from the study of attention, and then we considered various behavioral and neural training programs that have potential as innovative treatments. Our hope is that this progress will eventually help break the vicious cycle in depression of the bias resulting in increased processing of negative stimuli and thoughts, these experiences worsening mood, which exacerbates the bias and increases processing of negative information, further worsening mood, and so on. In this way, treat-

ing the negative attentional bias may lead to a broader reduction in a range of depression symptoms.

Acknowledgements

This work was supported by funding from the John Templeton Foundation (36751) to KAN and NTB, Intel Corporation to KAN and NTB, Canadian Institute for Advanced Research to NTB, and National Institutes of Health (T32MH065214) to ACM. The opinions expressed in this chapter are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

3

Closed-loop neurofeedback to improve negative attentional bias in depression *

...THE PROCESS OF GETTING STUCK IN YOUR THINKING, NOT BEING ABLE TO REDIRECT YOUR ATTENTION, IS AT THE CORE OF THIS DEBILITATING CHARACTERISTIC OF DEPRESSION.

Jutta Joormann

*Portions of this chapter were presented at the 2019 rtFIN Conference with coauthors N.B. Turk-Browne, D. Seok, M.T. deBettencourt, K.A. Norman, and Y.I. Sheline.

3.1 Introduction

Depressed individuals process negative stimuli differently from healthy subjects, leading to differences in attention, memory, and cognitive control (Gotlib & Joormann, 2010; Mogg & Bradley, 2005; Jones & Sharpe, 2017; Koster et al., 2017; Snyder, 2013). Neurally, depressed subjects tend to show larger and more prolonged responses to negative stimuli (Disner et al., 2011). This may manifest clinically as rumination, or the automatic replay of negative thoughts (Gotlib & Joormann, 2010; Nolen-Hoeksema et al., 2008). Given that depressed subjects attend more to negative information, researchers have designed paradigms to train participants to reduce this negative bias, and ultimately, depression severity (Gotlib & Joormann, 2010; Mogg & Bradley, 2005).

One approach typically used with depressed subjects is attention bias modification (ABM) training. Here, participants are trained to shift overt spatial attention away from negative stimuli and/or towards positive stimuli (Wells & Beevers, 2010; Browning et al., 2012; Li et al., 2016a; Cooper et al., 2014; Duque & Vazquez, 2018). A second approach, cognitive bias modification for interpretation (CBM-I; Mathews & Mackintosh, 2000), trains subjects to adopt the positive interpretation of an ambiguous situation (e.g., Joormann et al., 2015).

Following training, subjects are typically able to learn the reinforced behavior, for example, by attending less to negative stimuli (Wells & Beevers, 2010; Beevers et al., 2015). However, transfer to clinical measures (e.g., decreasing depression severity or self-reported rumination) have been found less consistently (see Hallion & Ruscio, 2011; Kruijt et al., 2013; Jones & Sharpe, 2017; Cristea et al., 2015). One limiting factor may be that this pre-programmed training approach is inflexible to individual differences, such as varying moments of atten-

tional lapses, effort, or understanding.

More recent approaches to attention training have used behavior-based real-time feedback, by presenting subjects with their mouse (Sanchez-Lopez et al., 2019a) or eye (Sanchez et al., 2016; Sanchez-Lopez et al., 2019b) positions in real-time. In these experiments, participants unscrambled sentences to form a positive meaning, while real-time feedback displayed where they were attending. The feedback added red or green outlines around individual words, depending on if the attended word was negative or positive, respectively. Though this training approach has yet to be tested in clinical populations, healthy participants showed promising improvements in reappraisal (Sanchez et al., 2016) and decreases in rumination (Sanchez-Lopez et al., 2019a,b). This use of real-time feedback brought awareness to current attentional states, which may yield clinical efficacy in the future.

While behavioral training has been the common approach to reduce attentional biases in depression, behavioral biases are a second-order effect, while differences in neural processing are the underlying cause. Functional magnetic resonance imaging (fMRI) may allow for observations that are "closer to the source" of the biases, and, thus, potentially a more sensitive measurement of these differences. Indeed, depressed subjects show neural – but not behavioral – evidence of increased processing of negative stimuli that are presented quickly (Mennen et al., 2019). In this vein, we seek to combine the advantages of customized, rapid feedback with the benefits of using neural measurements of attention.

A previous study that trained sustained attention in healthy subjects (deBettencourt et al., 2015) demonstrated the potential of real-time fMRI (rt-fMRI) neurofeedback. Participants received visual feedback based on their brain activity during an attention task. The task consisted of overlaid face and scene images with mixed opacity values. As subjects responded to

the cued, or *go* category, they had to simultaneously ignore the other un-cued, or *no − go*, category. Neurofeedback was closed-loop, or directly integrated with the task. For instance, if the neural data indicated that subjects were attending more to the incorrect category (e.g., faces), then that category would become more opaque (e.g., faces would become prominent and scenes would become more transparent). In this way, the task became *more difficult* during attentional lapses by forcing subjects to try harder to push themselves out of poor attentional states.

This study found that after a single rt-fMRI neurofeedback session, there were significant pre- to post-neurofeedback improvements in attention. Importantly, these changes were larger than those in a yoked control feedback group, who received feedback based on someone else's brain activity. Thus, this study proved that individualized feedback based on neural activity was advantageous to improve sustained attention.

3.1.1 Current study

We based our current study on the closed-loop neurofeedback procedure from deBettencourt et al. (2015). We chose this approach because of the direct involvement of sustained attention in both the task and neurofeedback signal. Other neurofeedback techniques that regulate region of interest (ROI) activity have found clinical success with depressed populations (Young et al., 2018). However, self-regulation of ROI activity is not necessarily tied to a specific cognitive state (Taschereau-Dumouchel et al., 2019).

In our current study, instead of trying to *improve sustained attention* to neutral stimuli in healthy subjects, our goal was to *reduce negative attentional bias* in depressed subjects. To adapt the study accordingly, we changed the neurofeedback task so subjects always had

to ignore negative faces. In this way, depressed subjects had to 'unstick' themselves when showing sustained negative attention to the prominent negative faces. Of note, a pilot study previously tested this neurofeedback task and found promising results for seven depressed participants (Schnyer et al., 2015).

We built upon previous versions of this task and made changes in multiple domains. First, we included two groups: a depressed group and a healthy control group. Second, we included three sessions of neurofeedback training. Lastly, we included additional behavioral and neural measures to capture changes over time.

Overall, our goals were to (1) decrease depression severity with rt-fMRI neurofeedback training, (2) determine which neural and behavioral tasks yielded robust pre-neurofeedback group differences, and (3) for the tasks found in (2), reduce the group differences after neurofeedback. Because we wanted to optimize our treatment first, we did not include a control group of depressed subjects that received sham neurofeedback. Future versions of this task can extract the most robust measures to develop focused training and test on future cohorts.

3.1.2 Hypotheses

As mentioned above, we included multiple tasks to understand how best to capture attentional differences between depressed and healthy participants. For this reason, we divided our expectations into primary and secondary hypotheses. We made this distinction based on how well we expected each task to capture biases in sustained attention. By including tasks both related and unrelated to neurofeedback training, we could test near and far transfer learning.

First, we present Table 3.1 as an overview of the study's full design. While we will include further details in the sections to follow, we provide the table as a reference when introducing

Visit	1 Pre NF	2 NF 1	3 Mid NF	4 NF 3	5 Post NF	6 1M FU	7 3M FU
Task 1	SCID-5	Resting-state	Gaze	**Go/no-go** (8 runs)	MADRS	MADRS	MADRS
Task 2	MADRS	Face-matching	Go/no-go	Face-matching	Gaze	Gaze	
Task 3	Gaze	**Go/no-go** (7 runs)	**Go/no-go** (9 runs)	Resting-state	Go/no-go	Go/no-go	
Task 4	Go/no-go						

Table 3.1: Full study design for all 7 visits: pre-neurofeedback, three neurofeedback sessions, post-neurofeedback, a one-month follow-up, and a three-month follow-up visit. Columns separate the different visits, while rows indicate the order of tasks performed at each visit. Cell color specifies the type of task: **pink** = clinical assessments; **beige** = behavior tasks completed outside of the scanner; **yellow** = scanner tasks. **NF** = neurofeedback; **FU** = follow-up. Go/no-go in **bold** is added to emphasize the rt-fMRI neurofeedback runs.

the tasks related to our hypotheses.

3.1.2.1 Primary hypotheses

We expected depression severity to decrease after neurofeedback training for the depressed group. We measured depressed scores with the Montgomery-Åsberg Depression Rating Scale (MADRS) structured clinical interview (Montgomery & Asberg, 1979; Williams & Kobak, 2008). We measured changes in depression severity by recording **MADRS** scores during 4 visits (Visit 1 and Visits 5-7; see Table 3.1).

Our next set of primary hypotheses were based on the neurofeedback task. We refer to this task as the **go/no-go - scanner task** to distinguish it from the behavioral version. As mentioned above, subjects were always instructed to attend to the neutral scenes and ignore the negative faces during neurofeedback. Thus, the relative strengths of the neural repre-

sentations of faces compared to scenes provided an index of the attended category. Before training, we hypothesized that depressed subjects would have difficulty disengaging attention away from negative faces. We operationalized this difficulty with a *sustained negative attention* measure that tracked the probability of remaining in the most negative attentional state (see Methods, Section 3.2.7.2 for further details).

Additionally, we hypothesized that neurofeedback training would reduce the negative attention bias for depressed subjects, which would make it easier for depressed subjects to 'unstick' themselves from negative states. Thus, we expected group differences in sustained negative attention to decrease over time. Additionally, if neurofeedback played a role in symptom improvement, we expected a positive relationship between the reduction of sustained negative attention and the reduction of depression severity.

The other primary task, which we will refer to as the **gaze task**, was also designed to measure biases in sustained attention, this time with eye-tracking instead of neural data. We chose an eye-tracking task because of its continuous measurement of attention through eye movements. Based off Kellough et al. (2008), each trial displayed four images representing different emotions (dysphoric, threatening, happy, and neutral) on the screen at once for 30 seconds. Recording eye movements throughout these long trials allowed us to assess which stages of attention (e.g., orientation, maintenance) differed the most between groups before training, and if those group differences changed over time.

Before neurofeedback, we expected depressed subjects to show difficulty disengaging attention away from the dysphoric images (Sanchez et al., 2013; Mennen et al., 2019). This would manifest as prolonged attention to the dysphoric images, as was the case in Kellough et al. (2008) and Armstrong & Olatunji (2012). Given initial group differences, we expected

the negative bias to decrease over time, demonstrating that the benefits of rt-fMRI neurofeedback transferred to this task.

We also hypothesized that before neurofeedback, depressed subjects might orient to positive images less often than control subjects, as in Armstrong & Olatunji (2012). Given an initial difference, we expected the group difference to decrease after training, though we were less certain because (1) orientation was a separate stage in attention compared to disengagement, and (2) our neurofeedback task trained attention away from negative images instead of towards positive ones.

3.1.2.2 Secondary hypotheses

Here, we describe additional tasks that we included to measure group differences over time. While these tasks were not designed to capture biases in sustained attention, we based them off successful applications in previous literature.

Similar to our neurofeedback task, subjects performed the **go/no-go - behavior task** to measure attentional biases through behavioral key presses. Again, the task consisted of overlaid face and scene images, though this time with equal opacity. Still, subjects always were instructed to attend one category (e.g., neutral faces) and ignore the opposite category (e.g., neutral scenes). For this behavioral version, we also included a condition where subjects attended to happy faces and ignored neutral scenes.

To quantify sustained attention through behavior, we used the same attention metric (A′) as in deBettencourt et al. (2015), as it was shown to improve after neurofeedback. In our study, however, instead of looking at overall A′ improvements, we calculated A′ differences between different emotional conditions (e.g., ignoring neutral faces versus ignoring negative

faces). Before neurofeedback, we expected that depressed subjects would have a more difficult time, and thus worse A′ performance, when ignoring negative faces compared to neutral faces. However, because differences in A′ performance across conditions was one step removed from the original finding of improved neutral A′ in deBettencourt et al. (2015), we explored this possibility as secondary hypothesis.

We also hypothesized, though with less certainty, that depressed subjects might show more difficulty (or lower A′ scores) when attending to happy faces compared to attending to neutral faces. While there is some evidence to suggest that depressed subjects show deficits attending to positive stimuli, attention to positive versus negative stimuli is usually compared directly (Kellough et al., 2008; Armstrong & Olatunji, 2012; Gotlib & Joormann, 2010). For this reason, it is unclear if the difference in attention is due to a lack of attention to positive stimuli or a bias for negative stimuli.

To include an additional metric of neural responses to emotional faces, we used the **face-matching task**. This task was adapted from Chai et al. (2015) to measure neural responses to negative, neutral, and happy faces in a block design. Originally, Chai et al. (2015) analyzed group differences in *average* amygdala responses to negative relative to neutral blocks. Before neurofeedback, we expected depressed subjects to show higher limbic system reactivity during negative face blocks (Chai et al., 2015; Chechko et al., 2013; Fales et al., 2008).

However, because we were interested in sustained attention, we were unsure if averaged block activity could capture negative biases related to disengagement. We looked to Sheline et al. (2009) for an additional analysis approach, which found that depressed subjects showed increased amygdala responses to negative images *only* towards the end of the trial. This motivated us to consider the time course of amygdala activity, as an increased response towards

the end of the block may indicate a failure to inhibit processing of negative stimuli. Given initial group differences in the amygdala response either averaged over the whole block or at the end of block only, we expected the group differences to decrease after neurofeedback.

Lastly, we included **resting-state** scans before and after neurofeedback. We expected depressed subjects to show different functional connectivity patterns from those of control subjects based on previous literature. Specifically, before neurofeedback, we expected the following group differences:

1. connectivity within the default mode network (DMN): depressed > healthy (Yu et al., 2019; Sheline et al., 2010; Greicius et al., 2007)

2. connectivity within fronto-parietal network (FPN): depressed < healthy (Yu et al., 2019; Lui et al., 2011; Hamilton et al., 2011)

3. connectivity between DMN and FPN: depressed > healthy (Yu et al., 2019)

4. frontal-amygdala connectivity: depressed < healthy (Shou et al., 2017; Satterthwaite et al., 2016)

For the measures that showed initial group differences, we expected those group differences to decrease over time.

3.2 Methods

3.2.1 Participants

Twenty-seven adults (14 female, mean age = 26.4 years) participated in the study. Twelve of the subjects were part of a healthy control **(HC)** group (6 female, mean age = 23.2 years;

6 male, mean age = 27.7 years) and 15 were part of the major depressive disorder **(MDD)** group (8 female, mean age = 26.3 years; 7 male, mean age = 28.4 years). Both groups completed the same tasks in the same order; the only difference between groups was the initial diagnosis requirements. Subjects were recruited from the University of Pennsylvania Center for Neuromodulation in Depression and Stress (CNDS) laboratory. The CNDS laboratory recruits individuals from the Mood and Anxiety Disorders program, student health, general adult psychiatry programs, and the general community. All subjects received monetary compensation for their participation. This study was approved by the University of Pennsylvania Institutional Review Board, as well as the Princeton University Institutional Review Board through an IRB Authorization Agreement.

To meet general eligibility requirements, all participants had to (1) be 18 - 60 years old, (2) be willing to not take psychotropic medications for the duration of the study, (3) be fluent in English, (4) be able and willing to provide written consent, and (5) be right-handed. Participants in the MDD group had to score at least a 16 on the MADRS clinical interview during Visit 1 to participate. Participants in the HC group had to meet the additional requirements of (1) having no history of MDD in their lifetimes, and (2) having no indication of current, significant depressive symptoms. The latter meant that participants could not score above an 8 on the MADRS interview during Visit 1.

Subjects were excluded from both groups if they (1) had a medical disorder that may cause depression or require medication that could cause depression symptoms, (2) were unwilling to provide informed consent, (3) had the diagnosis of concurrent DSM-5 psychiatric disorders (e.g., bipolar disorder, primary anxiety disorder, etc.), (4) had not completed the 10^{th} grade in school, (5) showed an active suicide risk, or (6) had MRI contraindications (e.g.,

foreign metallic implants).

3.2.2 Stimuli

In the following section, we describe the stimuli used for each task in the order of presentation.

Gaze-task

This task was based off Kellough et al. (2008), which measured sustained attention during one session. Thus, we had to adapt Kellough et al. (2008) to track changes over time in a multi-session design. For this reason, we created 4 different versions of the task to be completed on Visits 1, 3, 5, and 6 (Table 3.1). We supplemented images from the original study with stimuli from the International Affective Picture System (IAPS; LANG & PJ, 2005) and the Nencki Affective Picture System (NAPS; Marchewka et al., 2014). We included all IAPS images used in the original version (Kellough et al., 2008).

To normalize all images across databases, the images were cropped to be square. Next, images were sorted into four different versions while balancing for image entropy. We then made sure that all versions of the task did not differ significantly in entropy, contrast, or luminance. All subjects completed each version, but the version order was randomized for each participant.

Go/no-go task

Face and scene images for this task were adapted from deBettencourt et al. (2015). All scenes, comprising indoor and outdoor grayscale images, were taken directly from deBettencourt

et al. (2015). Because our task included additional positive and negative face conditions, we created our own set of faces. To collect adequate faces showing all emotions, we used the following databases: the Chicago Face Database (CFD; Ma et al., 2015), Karolinska Directed Emotional Faces (KDEF; Lundqvist et al., 1998), and NimStim (Tottenham et al., 2009).

These 3 databases allowed us to assemble a collection of male and female actors with happy, neutral, and negative (fearful or sad) images. Next, we normalized the images with the following steps: (1) cropping by the eyebrows and chin to ensure subjects would attend to emotion over hair alone while making male/female judgements, (2) grayscaling, (3) resizing, (4) intensity adjusting, (5) contrast normalizing, (6) luminance normalizing. For both the behavioral and scanner versions of this task, image sizes were adjusted such that the square image subtended a 10° visual angle as in deBettencourt et al. (2015). For more detailed information regarding how images were combined as composite face/scene stimuli, see deBettencourt et al. (2015).

Face-matching task

We included the same face and object images as in Chai et al. (2015). Face stimuli were adapted from the Radboud (Langner et al., 2010) and NimStim (Tottenham et al., 2009) stimulus sets. There were 2 versions of this task; one version drew images from the Radboud set and the other used images from the NimStim set. As subjects completed this task twice (Visits 2 and 4; Table 3.1), the order in which each subject completed the tasks was randomized by subject.

3.2.3 Procedure

Visits

The study consisted of 7 visits total, as illustrated in Table 3.1. The first 5 visits were part of the main study (pre-scanning, 3 neurofeedback sessions, post-scanning). Visit 6 was a behavior-only 1-month follow-up. Visit 7 was a 3-month follow-up phone call. For all subjects, we tried to schedule the first 5 visits within 2 weeks as closely as possible.

Clinical assessments

Upon providing consent on Visit 1, subjects completed the Structured Clinical Interview for DSM-5 Disorders (SCID-5; First et al., 2015) to assess lifelong symptoms, current depression symptoms, and the presence of additional exclusionary conditions. Subjects completed the SCID-5 on Visit 1 only. Upon completion, subjects underwent the Montgomery-Åsberg Depression Rating Scale (MADRS) structured clinical interview (Montgomery & Asberg, 1979; Williams & Kobak, 2008) to assess depression symptoms specifically over the week preceding Visit 1. Of note, the MADRS was also administered on Visits 5, 6, and 7 to assess depression severity within the previous week (see Table 3.1). This interview served as our main clinical assessment of how depression severity changed over time.

Behavior tasks

Behavior-only tasks were conducted in the CNDS Laboratory, Perelman School of Medicine, University of Pennsylvania. Stimuli were presented on 60 x 34 cm monitor (1280 x 1024 resolution) with the Psychophysics Toolbox for Matlab (http://psychtoolbox.org/).

Gaze task

A Tobii X120 eye-tracker (Tobii Technology, Stockholm, Sweden) was used to record eye movements (sampling rate = 120 Hz). Subjects' chins were fixed to a free-standing chin rest during the task to ensure that they had limited head motion and maintained a constant distance from the monitor. After eye-tracking calibration, subjects were instructed to freely view the different images presented on the screen as if they were watching television or looking at pictures in a photo album (as in Kellough et al., 2008). We told subjects that we were tracking their eyes to make sure they were attending to the images on the screen.

As in Kellough et al. (2008), the task consisted of 20 trials, with each trial lasting 30 s. Each trial was either a neutral filler or a valence trial. During neutral filler trials, the 4 images were all neutral. During valence trials, there was a neutral, dysphoric, threatening, and happy image shown on the screen. The full task consisted of 8 neutral filler trials and 12 valence trials that were presented in a random order for each subject. As mentioned above, subjects completed 4 different versions of this task. See Figure 3.1 for trial structure and example stimuli.

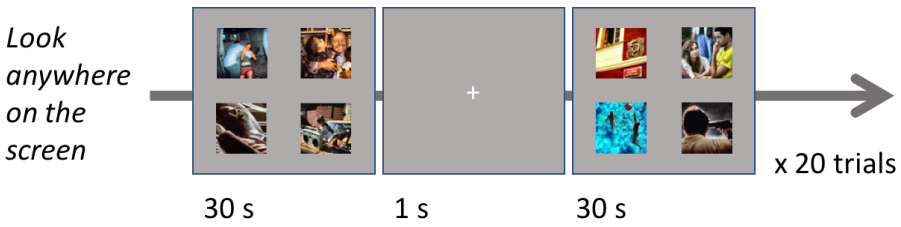


Figure 3.1: Trial structure for the gaze task. Subjects were instructed to freely view the images as if they were watching television. During each trial, 4 images were shown on the screen at once. Eye movements during the full 30-s trial were recorded to estimate sustained attention over time. *Note:* both of the example trials shown are valence trials, as there is a dysphoric, threatening, happy, and neutral image on the screen. Neutral filler trials were the same, except that all 4 images were neutral. The task consisted of 20 trials total: 12 valence trials and 8 neutral filler trials.

Note: Various obstacles (eye makeup, pupil size, glasses, etc.) impeded continuous eye

position estimates for some subjects. We adapted to the difference in measurements by normalizing our estimates by the number of gaze fixations that were recorded, as described subsequently (see Methods, Section 3.2.7.3).

Go/no-go task - behavior

Subjects completed 4 runs of this version of the go/no-go task. Each run contained 8 50-s blocks, where subjects were instructed to attend to a particular category (faces or scenes). As shown in Figure 3.2, there were 4 stimulus conditions in total. Thus, each of the 4 conditions was shown twice during each 8-block run. The condition order was randomized for each subject within each run, as shown in the top of Figure 3.5.

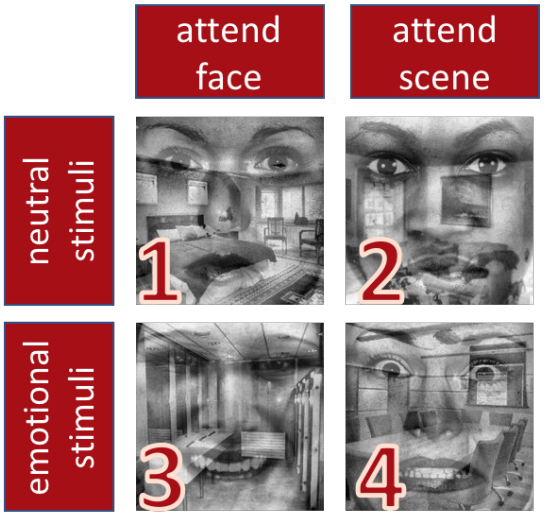


Figure 3.2: All stimulus conditions for the go/no-go task. In all conditions, subjects viewed a face and a scene mixture, while having to attend to one category and ignore the opposite category. The columns specify which category was attended (face or scene), and the rows specify if the faces were neutral or emotional. Altogether, the 4 conditions were: **(1)** attend neural faces; ignore neutral scenes, **(2)** attend neutral scenes; ignore neutral faces, **(3)** attend happy faces, ignore neutral scenes, **(4)** attend neutral scenes; ignore negative faces.

At the start of each block, subjects were given a cue that indicated the block type and *go*

stimulus. For instance, during a face block, the attended category was faces, and the specific *go* stimulus was either male or female faces. During a scene block, the *go* stimulus was either indoor or outdoor scenes. For instance, if the cue said *indoor scenes*, then subjects had to press their right index finger every time they saw the *go* indoor scenes (90% of trials), and had to refrain from pressing when they saw the *no-go* outdoor scenes (10% of trials). Additionally, while making indoor/outdoor judgements, subjects had to ignore the irrelevant category (e.g., faces). For each subject, the *go* scene category and *go* face category were the same across all visits. Assignment of categories was counterbalanced across subjects within each group. Figure 3.3 displays the specific trial structure. See deBettencourt et al. (2015) for additional task details.

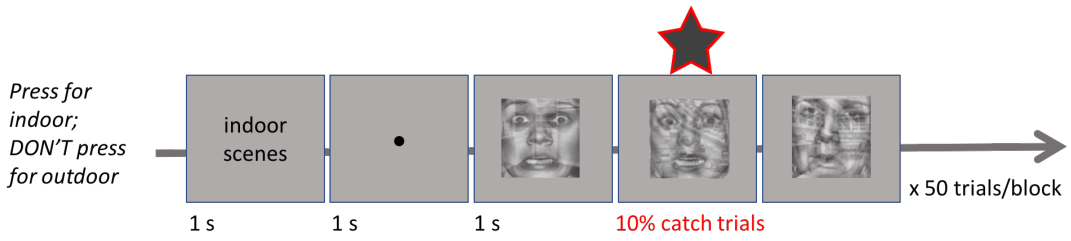


Figure 3.3: Example trial structure of the go/no-go task, for both the behavioral and scanner version. The *go* stimulus (indoor) and block category (scenes) were cued for 1 s. After that, each stimulus (a face and scene composite image) was displayed for 1 s. Subjects pressed their right index finger each time they saw the *go* stimulus (indoor scenes), and needed to refrain from pressing when they saw the *no-go* stimulus (outdoor scenes). Thus, this task required sustained attention throughout the block to properly inhibit the prepotent response during catch trials and continually attend to the correct image category.

Scanner tasks

Resting state

PsychoPy2 (Peirce et al., 2019) was used for this task. Subjects viewed a white cross-hair on a gray screen for 6 minutes. Before scanning, subjects were instructed to keep their eyes open

and to stay awake during the scan. Subjects completed this task during Visits 2 and 4 (Table 3.1).

Face-matching task

PsychoPy2 (Peirce et al., 2019) was used for this task. Subjects completed 2 5-minute runs of this task during Visits 2 and 4 (Table 3.1). The task involved matching images in 18-s blocks. There were 5 block types of different stimulus categories: happy faces, fearful faces, neutral faces, objects, and fixation. Each block type repeated 3 times within a run. Subjects were instructed to choose the bottom image that matched the top image by pressing their index fingers to indicate a left match or their middle fingers to indicate a right match (Chai et al., 2015). Figure 3.4 shows the trial structure in each block.

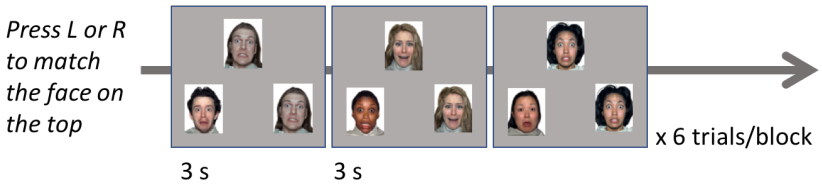


Figure 3.4: Trial structure for the face-matching task. Subjects pressed their index or middle fingers to match the top image to the left or right bottom image, respectively. There were 6 3-s trials within each block. All block types (happy faces, fearful faces, neutral faces, objects, and fixation) repeated 3 times within a run. Shown above is an example of a fearful face block.

Go/no-go task - scanner

Subjects completed a modified go/no-go task in the scanner that served as the neurofeedback task. Subjects completed 7, 9, or 8 runs on Visits 2, 3, and 4, respectively (Table 3.1). We chose the number of runs on each visit to maximize training time within a 2-hour scanning session. As in the behavioral version, each run contained 8 blocks. This time, however, the

first 4 blocks of every run contained neutral stimuli with constant opacity because they served as training data for the face versus scene classifier.

The second 4 blocks in every run (except run 1) were the rt-fMRI neurofeedback blocks. These blocks were all the same block type: the attended category was always neutral scenes and the distractor category was always negative faces. The difference in run conditions between the behavioral and scanner versions is shown in Figure 3.5. Importantly, subjects were instructed at the beginning of neurofeedback that the change in opacity was determined by their neural representations of faces versus scenes, instead of their pressing accuracy. By informing subjects that the feedback was based on their brains only, we hoped that subjects could learn from neurofeedback without being discouraged by poor performance.

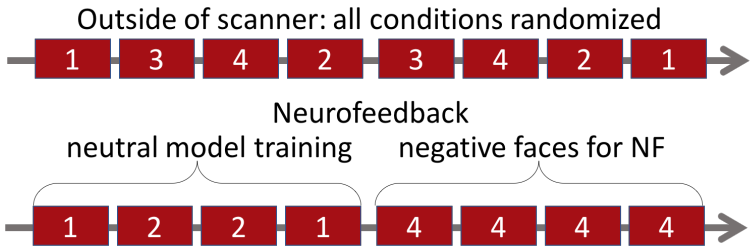


Figure 3.5: Behavioral and scanner runs of the go/no-go task. **Top:** During the behavioral version, all conditions were randomized and repeated twice within each run. **Bottom:** During neurofeedback (NF) runs in the scanner, the first 4 blocks served as model training, and included neutral images only (the order alternated between 1221, as shown, or 2112). Neurofeedback was delivered in the last 4 blocks of every run, when negative faces were always the distractors. See Figure 3.2 for example stimuli in each condition.

3.2.4 Data acquisition

All scanning was acquired with a 3T MRI scanner (Siemens Prisma), using a 64-channel head coil. Sequences were matched to deBettencourt et al. (2015) as much as possible. Thus, we collected a scout anatomical scan for anterior commissure-posterior commissure (ACPC)

automatic alignment from Siemens software. During the first scanning session, we collected a high-resolution magnetization-prepared rapid acquisition gradient-echo (MPRAGE) anatomical scan to construct the whole-brain mask used in real-time and for offline registration. To record as much of the brain as possible, the functional scans used a gradient-echo, echo-planar imaging sequence (2 s repetition time (TR), 28 ms echo time, 3 mm isotropic voxel size, 64 x 64 matrix, 192 mm field of view, 36 slices). At the end of each scanning session, a fieldmap scan was acquired for offline processing.

3.2.5 Real-time analyses

3.2.5.1 Preprocessing

After the MPRAGE scan, FSL (http://fsl.fmrib.ox.ac.uk/) was used to register the Montreal Neurological Institute (MNI) MNI152 standard-space T1-weighted average structure template (Grabner et al., 2006) to each subject's brain in functional space. We used this registered whole-brain ROI as the mask for each subject. During neurofeedback runs, each new DICOM was motion corrected to the previous time point following Siemens' custom motion correction. After that, the DICOM image was saved onto a local Linux machine in the scanning room. Then, we masked and flattened the data into a 2D-array. From there, the data were sent to the cloud server for further processing. See Table 3.2 for the resources used for each preprocessing step.

3.2.5.2 Cloud processing

Parallel to data collection, cloud processing was developed and implemented so that the analysis of DICOM images was no longer reliant upon local resources or proprietary software.

Index	Description	Performed on
1	Motion correction	Siemens' custom motion correction
2	ROI masking	Local Linux machine
→ Data sent to cloud →		
3	Smoothing	Cloud
4	High-pass filtering	Cloud
5	Z-scoring	Cloud
6	Classification	Cloud
← Data returned to local Linux machine ←		
7	Converting classification to opacity	Linux machine
8	Updating display	Linux machine

Table 3.2: Real-time processing steps.

We transferred all analysis to Python and executed the code on a Microsoft Azure virtual environment that was part of Penn Medicine's secure network. *NOTE:* Before cloud processing was completed, all real-time preprocessing and classification were performed identically to deBettencourt et al. (2019b). This was the case for HC subjects 1-2 and MDD subjects 1-6, who received Matlab-based neurofeedback. All other subjects received feedback from Python-based cloud processing, unless a specific technological issue (e.g., the internet being down) obstructed us from using the cloud. In that case, Matlab was used. However, for all offline analyses, all runs were retrained with Python to ensure that any differences between Matlab and Python did not affect our results.

Cloud processing involved sending the masked 2D-array of each volume of DICOM data to a virtual machine for further processing. Once the data arrived, images were spatially smoothed with a Gaussian kernel full-width half-maximum (FWHM) of 5 mm. Because the first 4 blocks were stable training periods, real-time processing only occurred in the last 4 blocks. Thus, during neurofeedback blocks, each 2-D array was high-pass filtered in real-time

(cutoff = 200 s, as in deBettencourt et al., 2019b), and z-scored using the mean and standard deviation from the previous 4 stable blocks (as in deBettencourt et al., 2015).

Next, the preprocessed data were used to predict the probability of scene versus face attention at every TR during neurofeedback. The current neurofeedback run used the model output that was trained from the most recent 8 stable blocks. For example, neurofeedback run 4 used the trained model that contained data from the 4 stable blocks from runs 2 and 3, (with the exception of run 2, which used training data from all 8 blocks of run 1). Scikit-learn's logistic regression function (Pedregosa et al., 2011) was used for training (with the parameters selected to best match Matlab processing: solver='saga', penalty='l2', max_iter=300). During neurofeedback, we solved for the scene versus face probability analytically and used the difference between scene and face probability as the output classification score. This score was saved as a text file and sent back to the local computer to influence the display on the following TR. See Figure 3.6 for a graphical delineation of processing steps.

3.2.5.3 Neurofeedback display

The local Matlab script controlling the display loaded a new text file each time a new file was detected. This score was converted to an opacity value for the neutral scene using the transfer function shown in Figure 3.7. Then, opacity was smoothed using a moving window with the previous 2 TRs to ensure the change in opacity was not abrupt (deBettencourt et al., 2015). This smoothed opacity value was set as the opacity for the following 3 trials (1.5 TRs) while the next TR was collected and preprocessed.

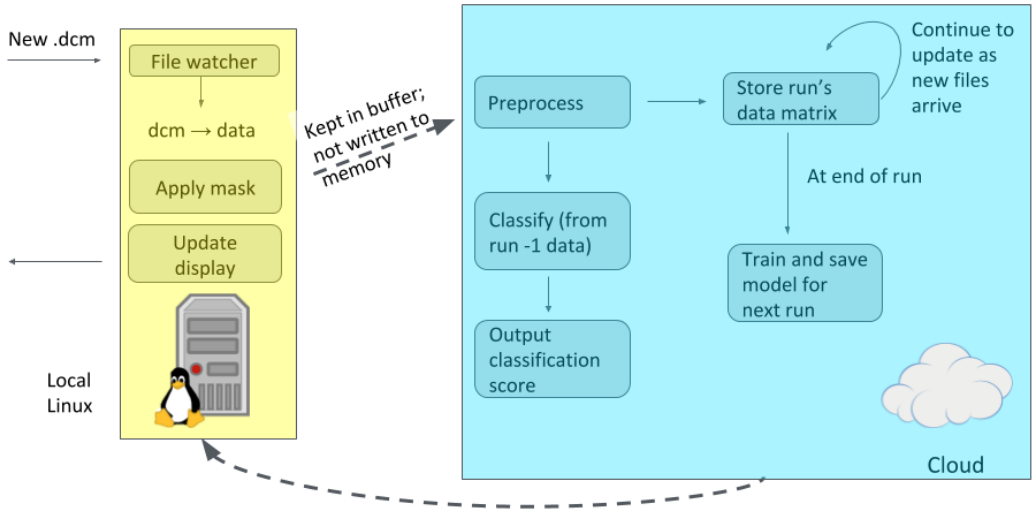


Figure 3.6: Processing configuration for the cloud setup. New DICOM images were sent to the local Linux machine, where they were masked and sent to the cloud computer. After they were preprocessed and scored with the cloud computer, the score was returned to the local Linux machine so that the display could be updated accordingly at the next TR.

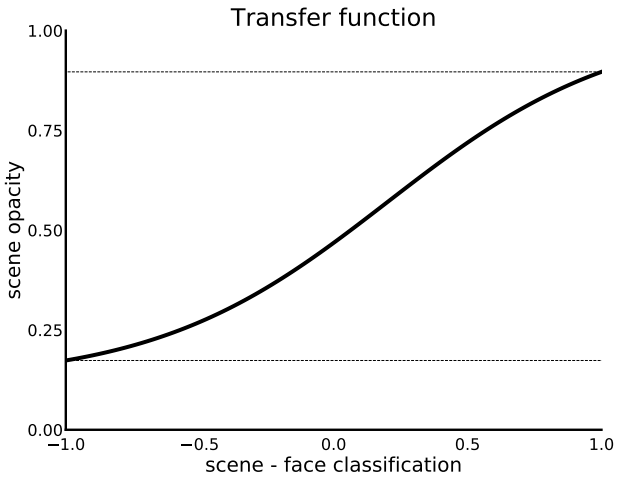


Figure 3.7: Transfer function to convert the neurofeedback output to scene opacity. The x-axis shows the scene — face classification score, while the y-axis shows the resulting opacity value for the scenes. We adjusted the logistic function from deBettencourt et al. (2015) by decreasing the gain to show an even mixture of faces and scenes more often. Scene opacity ranged from 17-90% so that subjects never saw *only* negative faces and could focus on neutral scenes to pull themselves out of negative attentional states.

3.2.6 Offline processing

After scanning, data were stored following the Brain Imaging Data Structure (BIDS) specification for further processing (Gorgolewski et al., 2016). Then, fMRIPrep (Esteban et al., 2019) was used for subsequent offline processing and image registration.

Results included in this section come from preprocessing performed using *fMRIPprep* 1.0.11 (Esteban et al., 2018b,a, RRID:SCR_016216), which is based on *Nipype* 1.1.6-dev (Gorgolewski et al., 2011, 2018, RRID:SCR_002502). The following description of preprocessing was generated with fMRIPrep.

Anatomical data preprocessing

Each T1w (T1-weighted) volume was corrected for INU (intensity non-uniformity) using `N4BiasFieldCorrection` (ANTs 2.1.0, Tustison et al., 2010) and skull-stripped using `antsBrainExtraction.sh` (ANTs 2.1.0), using the OASIS template. Brain surfaces were reconstructed using `recon-all` from FreeSurfer 6.0.0 (RRID:SCR_001847, Dale et al., 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (RRID:SCR_002438, Klein et al., 2017). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009, RRID:SCR_008796) was performed through nonlinear registration with the `antsRegistration` tool of ANTs 2.1.0 (RRID:SCR_004757, Avants et al., 2008), using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using `fast` (FSL 5.0.9,

RRID:SCR_002823, Zhang et al., 2001).

Functional data preprocessing

Functional data were slice time corrected using `3dTshift` from AFNI (AFNI 16.2.07 Cox, 1996) and motion corrected using `mcflirt` (FSL 5.0.9, Jenkinson et al., 2002). Distortion correction was performed using fieldmaps processed with `fugue` (FSL 5.0.9, Jenkinson, 2003). When there was no fieldmap scan due to ending a scan short (for one subject on one session), we used "fieldmap-less" distortion correction. This was performed by co-registering the functional image to the same-subject T1w image with intensity inverted (Wang et al., 2017; Huntenburg, 2014) constrained with an average fieldmap template (Treiber et al., 2016), implemented with `antsRegistration` (ANTs). After distortion correction, this was followed by co-registration to the corresponding T1w using boundary-based registration with 9 degrees of freedom, using `bbregister` (Freesurfer 6.0.0, Greve & Fischl, 2009). Motion correcting transformations, field distortion correcting warp, BOLD-to-T1w transformation and T1w-to-template MNI warp were concatenated and applied in a single step using `antsApplyTransforms` (ANTs 2.1.0) using Lanczos interpolation (Lanczos, 1964). Frame-wise displacement (Power et al., 2014) was calculated for each functional run using the implementation of *Nipype*.

Many internal operations of *fMRIPrep* use *Nilearn* 0.4.2 (Abraham et al., 2014, RRID:SCR_001362), principally within the BOLD-processing workflow. For more details of the pipeline, see the section corresponding to workflows in *fMRIPrep*'s documentation.

3.2.7 Primary analyses

3.2.7.1 Depression severity

MADRS scores were analyzed for changes over time. Specifically, we looked to see if depression severity changed between Visits 1 and 5 (pre- versus post-neurofeedback).

3.2.7.2 Go/no-go task - scanner

As mentioned above, all models were retrained after scanning to ensure that no differences between Matlab versus Python could affect results. We first preprocessed all data with the same real-time cloud pipeline. Then, we trained and tested the data with the same classification steps as those taken in real-time. Category separation was calculated by subtracting the probability of scene − face evidence at every TR. Because there were only two categories, the probability of one category was 1 − the probability of the other category. Thus, subtracting just changed the scale of scores from [0,1] to [-1,+1].

To understand how performance during neurofeedback differed by group and by neurofeedback session, we isolated the first three neurofeedback runs of the first session (Visit 2; Early NF) and the last three neurofeedback runs of the last session (Visit 4; Late NF). This way, we could compute pre- versus post-training differences.

Our first approach simply averaged scene − face classification across each run to understand changes over time. However, averaging over an entire time series did not capture the full distribution of attentional states. Instead of an average over time, we sought an estimate that better described the attentional states of subjects.

To quantify attentional states, we had to convert the continuous distribution of scene −

face classification into discrete bins. Because we used a logistic classifier for neurofeedback, classification values were most often near the extremes (± 1). To create attentional states with roughly equal samples, we adjusted the scene − face classification bins. The widest bins were in the middle to compensate for the smaller number of observations in the middle range, while the smallest bins were at the ± 1 boundaries. Figure 3.8 shows the adjusted bin sizes and their respective observation frequencies.

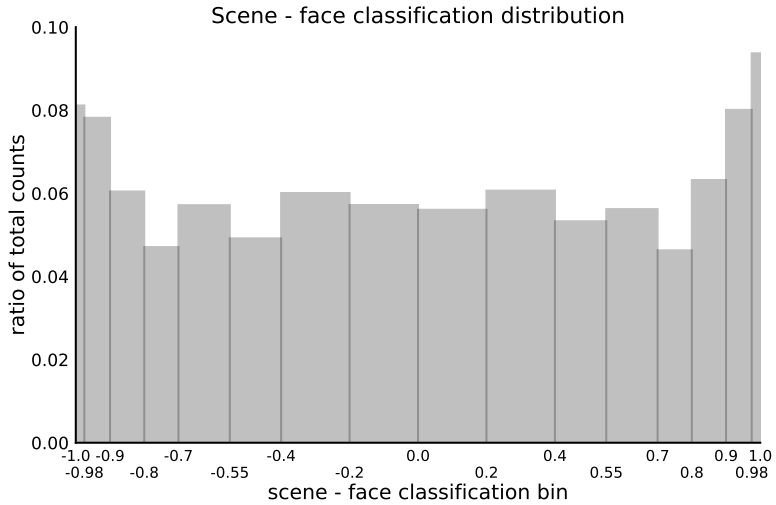


Figure 3.8: Histogram over all neurofeedback scene − face data to display our custom bins sizes. The x-tick labels show the actual bin ranges; bar height shows the proportion of frequency within that bin. Bins on the extremes [-1, -0.98] and [0.98, 1] incurred the largest frequency. To account for the difference in frequency, we made the bins towards the middle the widest. As a result, the distribution across all bins was roughly equal.

For each state, we quantified the persistence or *stickiness* within that particular attentional state. We operationalized this measure as the conditional probability that the scene − face evidence remained in the same bin range at a later time. We were interested in neural changes induced by a given display, so we focused on a 5-s HRF shift. As each TR was 2 s, we separately calculated results using 2 and 3 TR shifts, and averaged the results to estimate a 5-s shift.

Specifically, for each of the time delays, d (2 or 3 TRs), for a given attentional state bin, A, we calculated the persistence of that state by counting the number of times that the scene − face classification fell within A at time $t + d$, given it was in A at time t. We then divided this number by the total number of times the time series was in state A, shown in the equation below:

$$p(A_{t+d}|A) = \frac{\sum(state(t+d) = A)}{\sum(state(t) = A)} \tag{3.1}$$

We then averaged all conditional probabilities from the 2 and 3 TR shifts.

3.2.7.3 Gaze task

We recorded the (X,Y) positions from each eye during the 30-s trials. Following Kellough et al. (2008), we preprocessed the eye tracking data by defining eye positions in terms of *fixations*. In this way, we filtered the data such that fixations had to (1) be maintained for at least 100 ms and (2) remain within 1° of visual angle for this duration. After filtering all samples into a time series of fixations that met this criteria, each fixation was classified according to the attended emotional image. Because there was an unequal number of fixations for a given subject on a given trial, we normalized each of the metrics described next by the number of data points collected.

Initial orientation

To assess initial orientation differences by image category and group, for each trial, we recorded the image category upon which the subject first fixated. We calculated the probability of initial fixation for each day by taking the total number of times that the subject first fixated on

that image category divided by the total number of times that the subject fixated on any of the images. In this way, we did not count the trials in which no viable data were collected.

Total viewing ratio

To calculate the total viewing ratio of dysphoric images, for each trial, we totaled the number of fixations on dysphoric images divided by the total number of fixations recorded for that trial. We then averaged across all trials for each day.

Maintenance viewing ratio

To understand if participants had difficulty *disengaging* attention away from dysphoric stimuli, for each trial, we counted the number of fixations on the dysphoric image, given the subject was already fixated on the image. If the subject fixated on the dysphoric image, viewed other images, and then returned to the dysphoric image, we only considered the first continuous fixation in our analysis. We then divided this number of samples by the number of fixations recorded for that trial. Again, this normalization step corrected for the lack of data in some trials for some subjects. Finally, we averaged the maintenance ratios across all trials to get one ratio per day per subject.

3.2.8 Secondary analyses

3.2.8.1 Go/no-go task - behavior

To measure sustained attention with behavior, we estimated performance according to the nonparametric A′ (deBettencourt et al., 2015). This metric measured sensitivity to the *go* versus *no-go* categories through the hit and false alarm rates. To estimate differences specifically

in ignoring negative faces compared to neutral faces, we calculated the negative bias as

$$bias_{negative} = A'_{neutral} - A'_{negative} \tag{3.2}$$

Thus, negative biases greater than 0 indicate *worse* performance when ignoring negative faces, compared to neutral faces. This would suggest problems with disengaging attention away from negative stimuli.

Likewise, we calculated the difference in attention to positive versus neutral faces as

$$bias_{anti-positive} = A'_{neutral} - A'_{happy} \tag{3.3}$$

Thus, anti-positive biases greater than 0 indicate *worse* performance when attending to happy faces, compared to neutral faces, which would suggest difficulty in focusing attention on happy stimuli.

3.2.8.2 Face-matching task

After data were preprocessed and transformed into MNI space with fMRIprep, we extracted confounds for removal. For each run of the task, we extracted the 6 motion parameters along with frame-wise displacement. Processing included (1) removing the first 5 TRs of each run, (2) spatial smoothing with 6 FWHM using nilearn (Abraham et al., 2014), and (3) z-scoring each voxel over time. The two runs completed on each day were combined for further processing with `3dTproject` in AFNI (Cox & Hyde, 1997) which involved: high-pass filtering (cutoff = 100 s) and regressing out motion confounds. Although calculating beta coefficients for block types as in Chai et al. (2015) would be an appropriate way to analyze the data, we

did not calculate beta coefficients so that we could extract responses over time. We sought to replicate the findings in Sheline et al. (2009), specifically that amygdala activity continued to rise for MDD subjects towards the end of viewing sad faces.

Because we were interested in limbic response to negative faces, we focused on the left amygdala (LA). To derive the LA ROI, we repeated the following steps for each subject: (1) we registered all aparc+aseg labels to functional T1w space using Freesurfer's mri_convert and mri_label2vol functions (Dale et al., 1999), (2) we extracted the LA from the individual Freesurfer labels in native space, (3) as a final step, we converted the amygdala mask to MNI space using the ANTs-derived registration matrix from fMRIPrep (Avants et al., 2008). Finally, we created an aggregate LA mask by intersecting all of the amygdala masks from each subject.

To compare amygdala activity across groups, we averaged the time series from each block within the LA ROI. We then subtracted the resulting time series between the negative and neutral blocks. To consider the entire hemodynamic response, we included 5 TRs before the block start and 5 TRs after the block start.

3.2.8.3 Resting state

After data were preprocessed and registered to MNI space with fMRIPrep, we followed standard practices to further remove confounds (Yu et al., 2019; Satterthwaite et al., 2016). Namely, we extracted the following confounds that were calculated by fMRIPrep: the six motion parameters and global signals (CSF, WM, and GM). For each of these 9 confounds, we calculated the temporal derivative, quadratic signal, and the derivative of the quadratic signal, resulting in 36 total regressors. After extracting the preprocessed fMRIPrep runs in MNI

space, we first removed the first 4 TRs. Next, we removed confounds using `3dTproject` in AFNI with the following settings: polort = 2 to detrend the signal, passband filter between 0.01 - 0.08 Hz, and spatial smoothing with a 6 mm FWHM Gaussian filter. The final residual output from this preprocessing step was then loaded into Python for further analysis. *NOTE*: The first 2 healthy control subjects and the first 5 depressed subjects did not receive resting state scans. Thus, this analysis contains 7 fewer subjects total than the other analyses.

Node analysis

To understand if task-positive or task-negative nodes were more or less functionally-connected as in Yu et al. (2019), we extracted nodes from the Power atlas (Power et al., 2011), resampled to functional space (see the methods in Yu et al., 2019 regarding which systems were kept). As our sample size was limited, we focused our analyses on two main networks: one task-negative (DMN), and the other task-positive (FPN).

For each subject and each day, we loaded the residual data from AFNI and used nilearn's `NiftiLabelsMasker` to z-score and average voxel activity that belonged to the same nodes as marked in the Power atlas. Thus, each of the 227 nodes of interest had one time series associated with them for each run. We then correlated each node's time series with all other nodes. Following previous work (Yu et al., 2019; Uys et al., 2006), the within network connectivity, W_a for a given network ($a \in \{1,2,...11\}$) was calculated as:

$$W_a = \frac{\sum_{i,j \in a} A_{ij}}{N_a^2} \tag{3.4}$$

where i and j are the indices for all the nodes in the network a, N_a is the number of nodes in network a, and A_{ij} is the connectivity matrix.

To calculate pairwise connectivity between two networks, we used the same calculation as in Yu et al. (2019) to calculate pairwise connectivity PB_{a-b} between networks a and b (a $\neq$ b) as:

$$PB_{a-b} = \frac{\sum_{i \in a, j \in b} A_{ij}}{N_a N_b} \tag{3.5}$$

which summed all connections between nodes of each networks and normalized by the number of nodes in each network. Thus, each subject had one average connectivity score at each day that were later entered into statistical analysis.

ROI analysis

For this analysis, we again extracted the residual time series after running AFNI's `3dTproject`. Similar to the node analysis, we wanted to test if connectivity between limbic and frontal regions specifically were different before neurofeedback across groups. However, due to our limited sample size, instead of performing searchlight analyses as in Satterthwaite et al. (2016), we chose a priori ROIs to then average within region and calculate connectivity between regions. We focused on the LA to represent the limbic system, using the same ROI as described above.

For the LA ROI, we similarly extracted the average resting state time series and correlated activity with all the time series nodes within the FPN network (Power et al., 2011). Similar to the node analysis, we calculated the LA $\rightarrow$ FPN connectivity as:

$$PB_{LA-FPN} = \frac{\sum_{j \in FPN} A_j}{N_{FPN}} \tag{3.6}$$

3.3 Results

3.3.1 Primary analyses

3.3.1.1 Depression severity

As expected, depression severity decreased over time for depressed subjects (Figure 3.9). Depression scores decreased significantly from Visit 1 to Visit 5 (one-tailed $t(14) = 3.61$, $p = 0.0014$), from Visit 1 to Visit 6 (one-tailed $t(13) = 2.85$, $p = 0.0069$), and from Visit 1 to Visit 7 (one-tailed $t(11) = 3.13$, $p = 0.0048$).

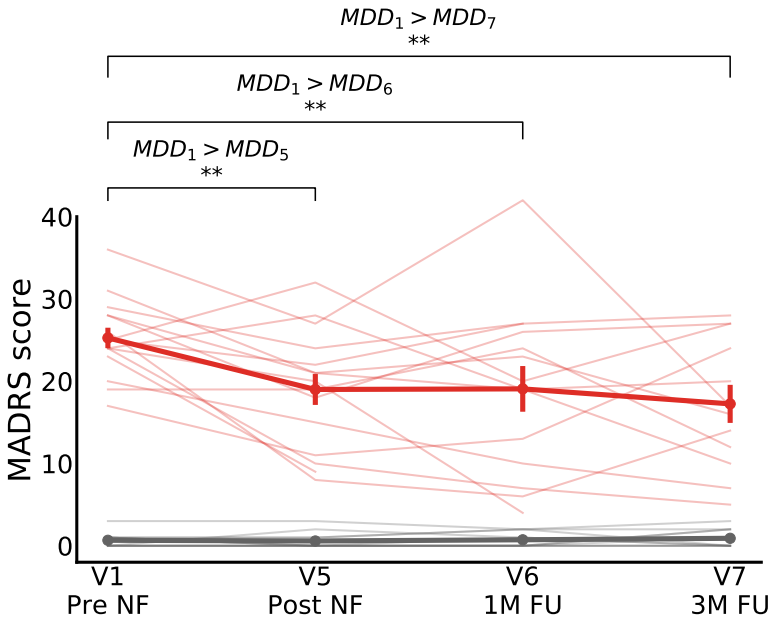


Figure 3.9: Depression severity decreased over time, as measured by the MADRS clinical interview. **Red** = MDD group; **black** = HC group. Thin lines represent individual subjects; thick lines represent group averages. Error bars represent ± 1 s.e.m. ** = $p < 0.01$

3.3.1.2 Go/no-go task - scanner

Neurofeedback performance changes over time

First, we examined the average of scene − face classification across entire runs. We did not observe classification differences between depressed and control subjects, either at the beginning of neurofeedback ($t(25) = 0.32$, $p = 0.76$) or at the end of neurofeedback ($t(25) = 0.34$, $p = 0.74$). On average, both groups showed a decrease in average scene − face classification from early to late neurofeedback (combined paired $t(26) = 2.61$, $p = 0.015$). Figure 3.10 displays these results.

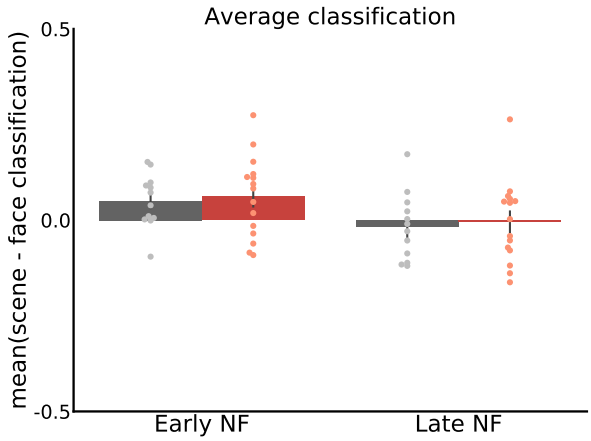


Figure 3.10: Changes in average classification scores. **Red** = MDD group; **black** = HC group. Circles represent individual subjects; bars represent group averages. Error bars represent ±1 s.e.m. There were no group differences during early or late neurofeedback.

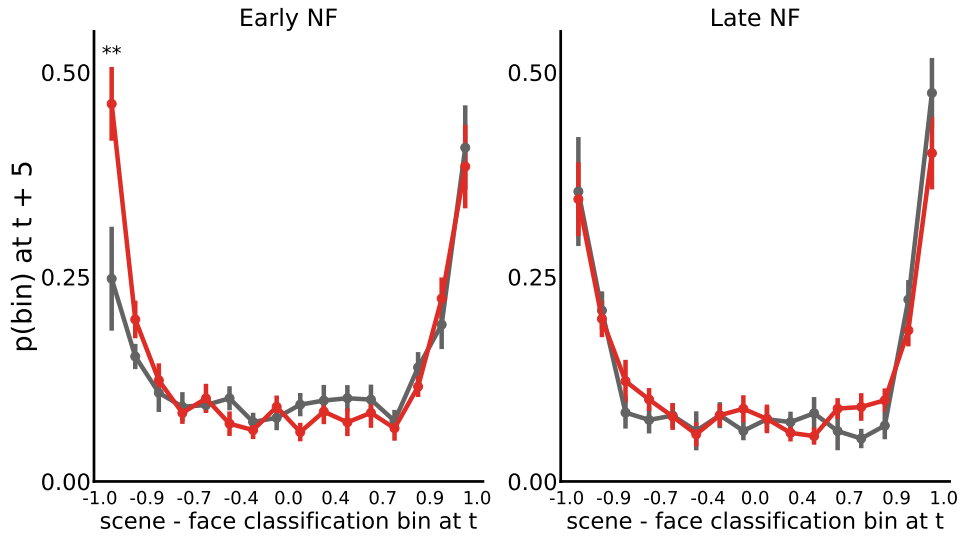
However, when considering scene − face classification in terms of discrete attentional *states*, we did observe group differences during early neurofeedback. Figure 3.11a shows the conditional probabilities that represent the persistence in each state during early neurofeedback **(left)** and late neurofeedback **(right)**. Initially, the largest difference between groups

was that depressed subjects were more likely than healthy controls to remain in the most negative state (one-tailed $t(25) = 2.80$, $p = 0.0049$). We will refer the probability of staying in the most negative bin as *sustained negative attention*.

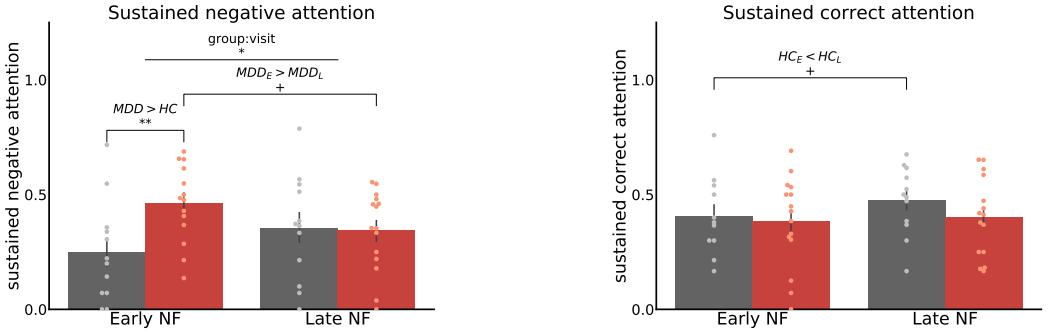
Changes in sustained negative attention from early to late training are shown in Figure 3.11b. At the end of neurofeedback training, MDD subjects showed marginally less sustained negative attention than at the start of neurofeedback training (paired one-tailed $t(14) = 1.67$, $p = 0.059$). Additionally, there was a significant interaction between group and visit ($\beta = -0.11$, $t = -2.042$, $p = 0.045$), such that the MDD group decreased in sustained negative attention from early to late training, while the HC increased their sustained negative attention.

At the end of neurofeedback training, the largest difference between groups was now in the exact opposite scene − face evidence bin, where attention was maximally maintained on neutral scenes. We will refer to this conditional probability of staying in this most correct bin as *sustained correct attention*. Although this state was where the groups differed the most at the end of training, the group difference was not significant (one-tailed $t(25) = -1.17$, $p = 0.13$).

Changes in sustained correct attention from early to late training are shown in Figure 3.11c. The increased group difference from early to late neurofeedback was caused by the control group specifically. The control group marginally improved their sustained correct attention (paired one-tailed $t(11) = -1.45$, $p = 0.089$), while the depressed subjects did not improve as much (paired one-tailed $t(14) = -0.24$, $p = 0.41$). We did not find a significant interaction between group and visit ($\beta = -0.025$, $t = -0.55$, $p = 0.58$).



(a) Conditional probability across all bins. **Red** = MDD group; **black** = HC group. Error bars represent ±1 s.e.m. **Left:** During early neurofeedback, the largest group difference was that MDD subjects got stuck in the most negative state by attending to negative faces (*sustained negative attention*). **Right:** At the end of neurofeedback, the largest group difference was that control subjects properly maintained attention on the neutral scenes (*sustained correct attention*), though this group difference was not significant. ** = p < 0.01



(b) **(c)**

Figure 3.11: **(a)** Sustained attention in all states during early and late neurofeedback. **(b)** Changes for sustained negative attention only (most negative bin). MDD subjects showed significantly more sustained negative attention during early, but not late, neurofeedback. Additionally, there was a significant interaction between group and visit. **(c)** Changes for sustained correct attention (most positive bin). HC subjects showed a trending improvement in sustained correct attention from early to late neurofeedback. **Red** = MDD group; **black** = HC group. Circles represent individual subjects; bars represent group averages. Error bars represent ±1 s.e.m. ** = p < 0.01; * = p < 0.05; + = p < 0.1

Relationship to MADRS

As shown in Figure 3.12, the reduction in sustained negative attention was related to the reduction in depression severity. There was a significant positive correlation for all subjects (Pearson r = 0.50, p = 0.0094) and a trending positive correlation when considering MDD subjects only (Pearson r = 0.48, p = 0.083).

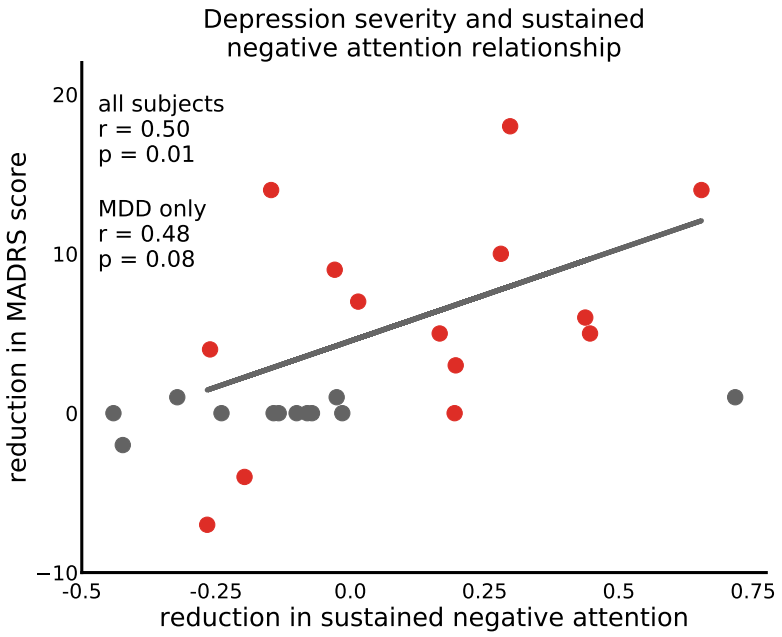


Figure 3.12: Reduction in MADRS symptom severity plotted as a function of reduction in sustained negative attention. **Red** = MDD group; **black** = HC group. Circles represent individual subjects. There was a trending positive relationship between the reduction of sustained negative attention and the reduction of depression severity when only considering MDD subjects; the black line shows the line of best fit for MDD subjects only.

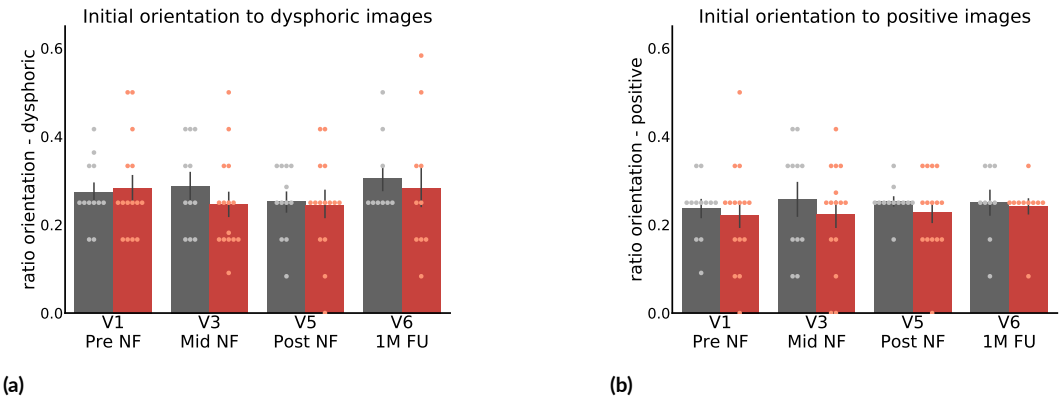


Figure 3.13: Initial orientation to **(a)** dysphoric and **(b)** positive images. No significant differences were found between groups at the onset of training, or on subsequent visits. **Red** = MDD group; **black** = HC group. Circles represent individual subjects; bars represent group averages. Error bars represent ± 1 s.e.m.

3.3.1.3 Gaze task

Initial orientation

As stated above, we thought groups may differ in initial orientation; for example, depressed subjects orienting to the happy images less often. However, we were not confident about group differences, as the first fixation was a separate attentional process from sustained attention.

We did not find initial group differences in orientations. On Visit 1 (Pre NF), depressed subjects did not orient to dysphoric images significantly more often (one-tailed $t(25) = 0.26$, $p = 0.40$; Figure 3.13a), nor did depressed subjects orient to positive images significantly less often (one-tailed $t(25) = -0.37$, $p = 0.36$; Figure 3.13b). There were no significant differences between groups on subsequent visits either.

Total viewing ratio

As the total time spent viewing an image is more reflective of sustained attention than initial orientation, we had more confidence that depressed subjects would spend more time viewing dysphoric images. On Visit 1, depressed subjects spent more total time viewing dysphoric images compared to control subjects on average, though this wasn't significant (one-tailed $t(25) = 1.28$, $p = 0.11$). Additionally, depressed subjects showed a trending decrease in the total amount of time they spent viewing dysphoric images from pre- to post-neurofeedback (paired one-tailed $t(14) = 1.44$, $p = 0.086$). Looking at Visits 1 and 5 only, we did not find a significant interaction between group and visit ($\beta = -0.021$, $t = -1.11$, $p = 0.28$). Figure 3.14a plots group performance for all visits.

Because depressed subjects decreased the time spent viewing dysphoric images, we examined the relationship between the decrease in total time spent viewing dysphoric images and the decrease in MADRS scores. There was a slight positive relationship including all subjects (Pearson $r = 0.29$, $p = 0.14$) and only MDD subjects (Pearson $r = 0.30$, $p = 0.28$), but neither correlation was significant.

Maintenance viewing ratio

We next examined the fraction of time that gaze was maintained on the dysphoric image in a continuous fixation. This estimate more closely captures the tendency to persist in a negative state. On Visit 1, depressed subjects maintained fixation for a longer amount of time compared to control subjects on average, though this was trending (one-tailed $t(25) = 1.69$, $p = 0.051$). From pre- to post-neurofeedback, depressed subjects significantly decreased their maintenance time on dysphoric images (paired one-tailed $t(14) = 1.85$, $p = 0.043$). Looking

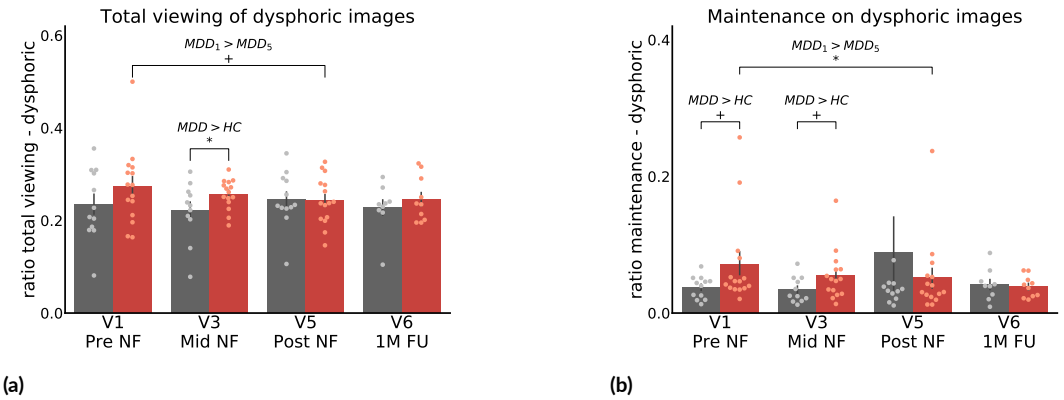


Figure 3.14: **(a)** Total viewing ratio for dysphoric images **(b)** Maintenance ratio on dysphoric images. MDD subjects improved in both metrics from pre- to post-neurofeedback, but initial group differences and MDD improvements were larger for maintenance on dysphoric images. *Note:* One outlier in the HC group only had one fixation during V5 and thus the maintenance ratio of 1 is not shown within the axis limits in (b). **Red** = MDD group; **black** = HC group. Circles represent individual subjects; bars represent group averages. Error bars represent ± 1 s.e.m. * = $p < 0.05$; + = $p < 0.1$

at Visits 1 and 5 only, we did not find a significant interaction between group and visit (β = -0.035, t = -1.29, p = 0.20). Figure 3.14b shows group performance over all visits.

As the depressed subjects reduced maintenance on dysphoric images and depression severity after neurofeedback, we then examined the relationship between these improvements. There was a slight positive relationship including all subjects (Pearson r = 0.11, p = 0.58), but no relationship upon considering depressed subjects only (Pearson r = -0.033, p = 0.91).

3.3.2 Secondary analyses

3.3.2.1 Go/no-go task - behavior

Given that depressed subjects showed difficulty in ignoring negative faces during neurofeedback, we looked to see if there were any biases present in behavior. Our behavioral metric was the nonparametric measure A′. As explained above, if depressed subjects showed an A′ negative bias, we would see worse A′ scores when they were ignoring negative faces compared

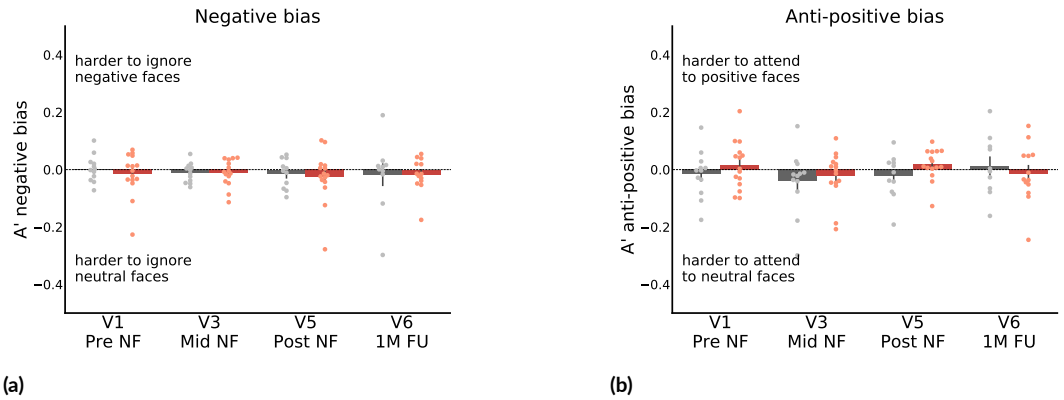


Figure 3.15: A′ biases for **(a)** negative and **(b)** anti-positive images. No significant group differences were shown for either attentional bias condition at any visit. **Red** = MDD group; **black** = HC group. Circles represent individual subjects; bars represent group averages. Error bars represent ±1 s.e.m.

to when they were ignoring neutral faces. As shown in Figure 3.15a, we found no group differences initially ($t(25) = 0.78$, $p = 0.44$), nor at any point during the course of training for the A′ negative bias.

We also checked to see if depressed subjects showed an A′ anti-positive bias, or had a more difficult time attending to positive faces, as compared to neutral faces. If this had been the case, we would expect A′ to be larger when subjects were attending neutral faces compared to positive faces. We did not find any group differences in A′ anti-positive bias during Visit 1 ($t(25) = -0.94$, $p = 0.36$), nor at any stage during training (Figure 3.15b).

3.3.2.2 Face-matching task

As described above, we expected depressed subjects to show initial biases when viewing negative faces. We hypothesized that difficulty disengaging attention from negative faces would result in larger and more prolonged responses in the LA ROI. If we found this initial difference in processing negative images, and our neurofeedback manipulation successfully re-

duced negative attentional bias, we expected the difference between groups to decrease after neurofeedback.

To estimate if there was increased processing that was specific to negative faces, we subtracted the average LA activity for neutral face blocks from that of the negative face blocks. First, we looked to see if average activity over the entire block differed by group before neurofeedback. We shifted the average amygdala signal by 2 TRs (4 seconds) to account for the HRF delay. Then, we averaged activity over the entire block (18 s, or 9 TRs). On average, depressed subjects did not show larger activity to negative versus neutral faces before neurofeedback (one-tailed $t(25) = 0.96$, $p = 0.17$).

Next, instead of considering the average activity over an entire block, we evaluated if the time series showed group differences. Figure 3.16a plots the average time series for each group for negative − neutral blocks before neurofeedback. As shown, the response to negative faces in the MDD group increases over time. Eventually, the group difference grows to a significant level 1 TR (unshifted) into the next block (one-tailed $t(25) = 1.95$, $p = 0.032$).

As shown in Figure 3.16b, after neurofeedback, depressed subjects did not show an increased LA response for negative faces 1 TR into the next block (one-tailed $t(25) = 0.35$, $p = 0.36$). From pre- to post-neurofeedback, depressed subjects showed a trending decrease in their response at this TR (paired one-tailed $t(14) = 1.57$, $p = 0.069$), but control subjects did not (paired one-tailed $t(11) = -0.034$, $p = 0.49$). We did not find a significant interaction between group and visit for this specific TR response ($\beta = -0.21$, $t = -1.19$, $p = 0.24$).

Even though depressed subjects decreased their LA response to negative faces from before to after neurofeedback, we did not find a significant relationship between decreases in LA activity and reduction in MADRS scores with all subjects (Pearson $r = 0.086$, $p = 0.67$), nor

only depressed subjects (Pearson r = -0.047, p = 0.87).

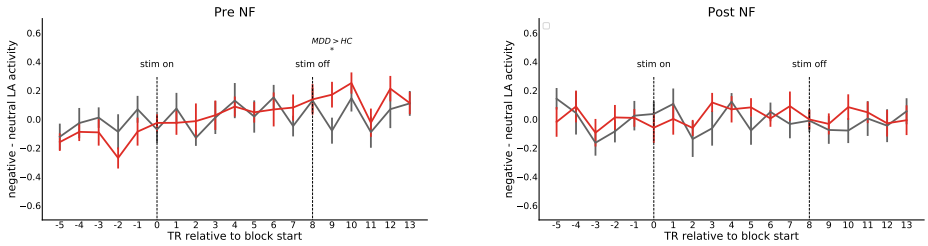


Figure 3.16: Left amygdala response to negative — neutral face blocks. **(a)** Pre NF responses **(b)** Post NF responses. MDD subjects showed increased LA signal during negative face blocks before neurofeedback, but not after. **Red** = MDD group; **black** = HC group. Lines represent group averages. Error bars represent ±1 s.e.m. * = p < 0.05

3.3.2.3 Resting state

Based on previous literature, we expected to find initial group differences in task-negative (DMN) and task-positive (FPN) networks, along with connectivity between the LA and FPN. Again, our results are limited in sample size, as our first 7 subjects (2 HC and 5 MDD) did not have resting-state scans.

Node analysis

Figure 3.17 shows results for pre- and post-neurofeedback differences in node connectivity. As expected, we found trending, but not significant, larger DMN connectivity for MDD compared to HC before neurofeedback (one-tailed $t(18) = 1.42$, $p = 0.086$), but not after neurofeedback (one-tailed $t(18) = 0.37$, $p = 0.36$). Depressed subjects specifically showed a trending decrease in DMN connectivity from pre- to post-neurofeedback (paired one-tailed $t(9) = 1.46$, $p = 0.089$; Figure 3.17a). We did not see a linear relationship between the reduction in DMN connectivity and the reduction in MADRS scores for either all subjects

(Pearson r = -0.26, p = 0.26) or only MDD subjects (Pearson r = -0.45, p = 0.19). Additionally, these negative linear correlations were the opposite of our predicted direction, as we expected reduced DMN within-network connectivity to be associated with the reduction in depression severity.

Contrary to our expectation, we found a trend for MDD subjects to have larger FPN within-network connectivity before neurofeedback ($t(18) = 1.68$, $p = 0.11$; Figure 3.17b). After neurofeedback, depressed subjects still showed slightly larger average FPN connectivity, though there was a smaller group difference ($t(18) = 0.89$, $p = 0.39$). Depressed subjects did not show a significant decrease in FPN connectivity from pre- to post-neurofeedback (paired one-tailed $t(9) = 1.050$, $p = 0.16$).

Also against our expectation, depressed subjects showed a trend towards more negative DMN-FPN connectivity before neurofeedback ($t(18) = -1.77$, $p = 0.094$; Figure 3.17c). After neurofeedback, the difference between groups was weaker ($t(18) = 0.093$, $p = 0.93$). We did not find a significant interaction between group and visit ($\beta = 0.0002$, $t = 1.14$, $p = 0.26$).

ROI analysis

Against our expectations, we found (nonsignificantly) less negative FPN-LA connectivity in depressed, compared to control, subjects before neurofeedback ($t(18) = 1.84$, $p = 0.082$). After neurofeedback, there was a weaker group difference ($t(18) = -0.29$, $p = 0.78$). As shown in Figure 3.18, this change from pre- to post-neurofeedback appears to be driven by changes within the control group.

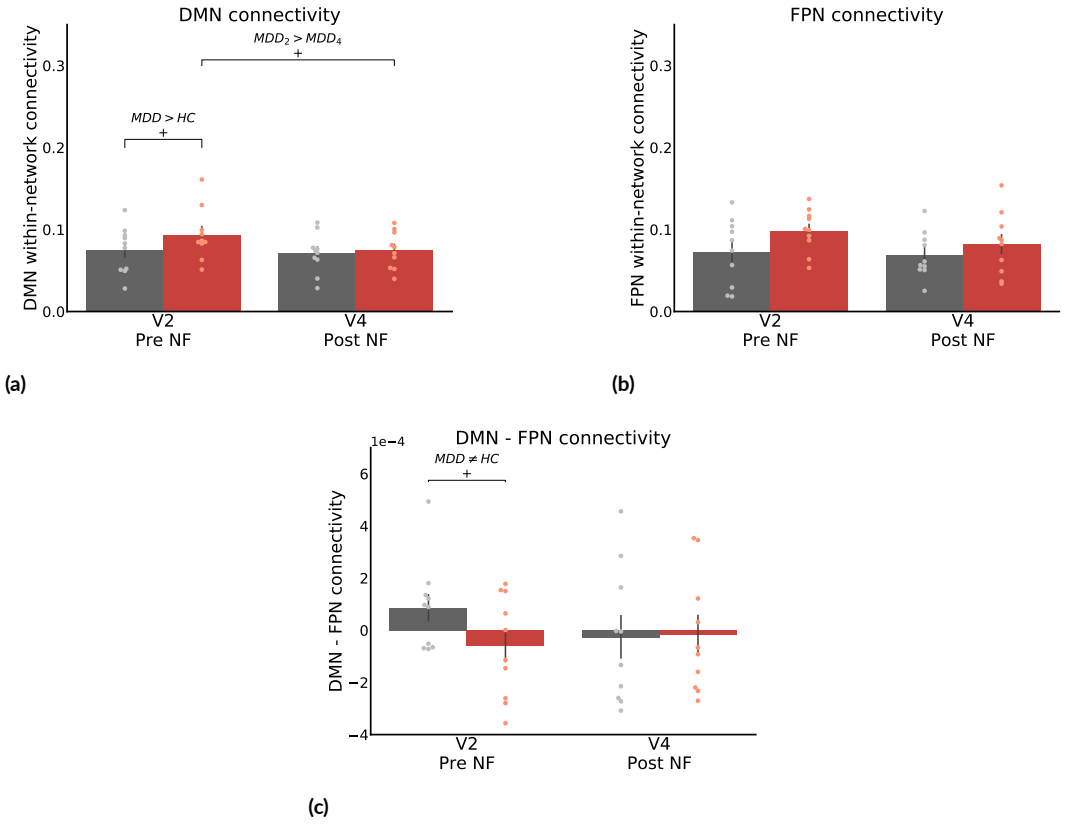


Figure 3.17: Within-network connectivity for **(a)** default mode network (DMN) and **(b)** fronto-parietal network (FPN). **(c)** Between network connectivity between DMN and FPN. **(a)**: MDD subjects showed trending larger DMN activity before neurofeedback, but not after. **(b)**: No significant group differences were observed in FPN connectivity. **(c)**: MDD subjects showed trending more negative DMN-FPN connectivity before neurofeedback, but not after. **Red** = MDD group; **black** = HC group. Circles represent individual subjects; bars represent group averages. Error bars represent ± 1 s.e.m. + = p < 0.1

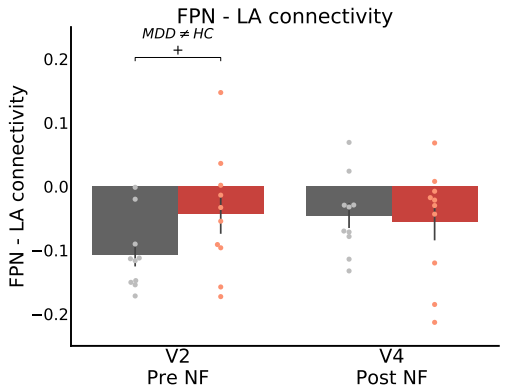


Figure 3.18: Connectivity between fronto-parietal network (FPN) and left amygdala (LA). Before neurofeedback, MDD subjects showed slightly less negative FPN-LA connectivity, opposite to prior hypotheses and findings. There were smaller group differences after neurofeedback, but this change appears to be driven by the HC group. **Red** = MDD group; **black** = HC group. Circles represent individual subjects; bars represent group averages. Error bars represent ± 1 s.e.m. + = p < 0.1

3.4 Discussion

Overall, the tasks that were best suited to capture differences in negative disengagement ultimately yielded the most robust initial group differences. Specifically, our sustained negative attention measure during neurofeedback showed the strongest statistical difference in initial group performance. Out of the three eye-tracking measures that we computed for the gaze task, maintenance time on the dysphoric images yielded the largest statistical initial group difference. Regarding initial group differences for the face-matching task, the LA response to negative faces did not differ when averaged over all time points, but was different only at the end of the block.

These results suggest that the negative attentional biases found in depressed subjects are most evident when subjects must inhibit their attention to negative stimuli. In addition, the presence of group differences for the go/no-go task neurally — but not behaviorally — implies that neural measures may be more sensitive than behavioral ones to capture attention

differences. This underscores the value of using rt-fMRI training to reduce negative attentional bias, as our task successfully demonstrated.

Compared to most depression studies in rt-fMRI literature, this study is unique in its design and analysis methods. First, we used neurofeedback to train attention directly, as opposed to regulating mean activity within an ROI (e.g., Linden et al., 2012) or connectivity between regions (e.g., Yamada et al., 2017). Additionally, most of the previous rt-fMRI studies trained depressed subjects to increase neural responses when recalling happy autobiographical memories (Young et al., 2014; Yuan et al., 2014; Young et al., 2017; Zotev et al., 2016). Regulating positive emotions has yielded robust benefits, as Mehler et al. (2018) even found unintentional clinical benefits for the control group, who were instructed to imagine relaxing scenes while regulating scene-specific ROIs. Thus, the benefits of upregulating one's response to positive emotions may be robust, despite behavioral training not yielding consistent effects. Our study explored the less-common approach of using negative, instead of positive, stimuli for training.

Instead of reinforcing neural responses to positive stimuli, we used neurofeedback to train attention away from negative stimuli. Our study differs from the small number of other rt-fMRI studies using negative stimuli in that we regulated a decoded cognitive state, as opposed to mean ROI activity. For instance, Hamilton et al. (2016) trained 10 depressed subjects to reduce salience network activity in response to negative scenes. In another example, MacDuffie et al. (2018) had subjects recall negative memories while using a strategy from cognitive behavioral therapy (CBT). During CBT application, a neurofeedback signal trained subjects to decrease anterior cingulate cortex (ACC) activity. Both studies yielded promising clinical benefits specific to real-time training, in the form of decreased negative self-descriptions

(Hamilton et al., 2016) and increased use of the trained CBT strategy after neurofeedback (MacDuffie et al., 2018). Although the studies described above trained depressed subjects to decrease their neural response to negative stimuli, our study is unique in its use of closed-loop attention training.

Our future goals for this treatment program involve (1) proving that individualized neurofeedback was necessary to achieve clinical benefits, and (2) transferring the neural trace of attention to another modality to adapt this closed-loop design for a less-costly method (e.g., EEG, eye-tracking, or reaction time, as in Keynan et al., 2019; Sanchez-Lopez et al., 2019b; deBettencourt et al., 2019a, respectively). Regarding the first aim, we plan to include a follow-up control group of depressed subjects. These depressed subjects would received sham neurofeedback, either by receiving feedback that is yoked to a previous depressed participant's brain (deBettencourt et al., 2015) or determined by a ROI deemed irrelevant to the subject's current attentional state (Young et al., 2017). In this design, if the control group of depressed subjects do not show the same benefits, we can be more certain of the benefits of attention-based individualized neurofeedback.

Regarding the second goal of transferring our neural signal from fMRI to another method (as demonstrated with amygdala activity in Meir-Hasson et al., 2014; Keynan et al., 2019), we first must understand which areas of the brain best determine clinical benefits. In the analyses included in Appendix A, we replicated the analysis done in deBettencourt et al. (2015) to understand how different networks contributing to the neurofeedback signal relate to different outcomes. Specifically, we correlated the relative involvement of the perceptual and attentional networks in the neurofeedback signal with the various results detailed in the above sections.

We did not find any significant relationships between either network's contributions to neurofeedback and changes in sustained negative attention. However, we found a trending positive relationship (all subjects) and a significant positive relationship (MDD subjects only) between the increase in attentional network contribution and the improvement in sustained correct attention. This difference in results between our two sustained attention measurements suggests that there may be different neural mechanisms underlying the reduction in sustained negative attention versus the increase of sustained correct attention.

The former reduction in sustained negative attention reflects an improvement in inhibiting attention away from negative faces. Thus, changes to this type of attention may involve regions in the brain related to emotional regulation, such as connectivity between the amygdala and frontal regions (Satterthwaite et al., 2016). As the latter improvement in sustained correct attention demands maintaining attention on neutral scenes, one would expect improvements in this domain to require increased attentional network involvement. See Appendix A for more details on this analysis and results.

Finally, we plan to use the rt-cloud software in future iterations while exploring other training methods to optimize neurofeedback (e.g., training depressed subjects to focus on happy faces). The open-source cloud processing will allow other institutions to deploy this software in a way that is efficient and reproducible. Ultimately, we hope to have both the real-time processing and neurofeedback display controlled via a web server to minimize the amount of dependency on the local scanning suite. With multiple variants of the task able to be deployed simultaneously, we can determine the best approach for the optimal clinical trial.

Acknowledgements

This work was supported by funding from the Intel Corporation to KAN and NTB, National Institutes of Health (T32MH065214) to ACM, and the University of Pennsylvania Endowment.

4

Real-time neurofeedback to alter interpretations using naturalistic stimuli

Perception is reality.

Lee Atwater

4.1 Introduction

Compared to healthy subjects, depressed subjects are more likely to assume a negative interpretation of ambiguous scenarios, especially for those that use self-referential prompts (Everaert et al., 2017). Past research has associated increased depression severity with the persistence of these negative interpretations even after learning that the scenarios ended positively (Everaert et al., 2018).

To reduce this negative bias and encourage positive interpretations, researchers have used cognitive bias modification for interpretation (CBM-I) training (Mathews & Mackintosh, 2000). During positive CBM-I training, depressed subjects are shown scenarios that differ in interpretation depending on the final word, e.g., *The new people you meet will find you (boring/friendly).* During training, the final word is revealed to disambiguate the meaning (e.g., *friendly*) so that subjects become more likely to expect the positive outcome (Joormann et al., 2015).

A meta-analysis collapsing across anxious and depressed individuals found that CBM-I training led to a larger reduction in negative biases compared to attention bias modification (ABM) training (Hallion & Ruscio, 2011). One explanation for the increased efficacy is that CBM-I training uses self-relevant stimuli that are inherently more realistic and relatable than negative faces or single words used in ABM training.

As with other training paradigms, however, results are mixed in regards to whether CBM-I training actually improves depressive symptoms (Jones & Sharpe, 2017). One explanation for the mixed clinical results is that the training uses a one-size-fits-all approach. In other words, all subjects are trained in the same way, regardless of momentary lapses in attention,

effort, or belief in one of the interpretations.

4.1.1 Current study

In the present study, we explored the feasibility of using real-time neurofeedback training to bias healthy subjects toward a specific interpretation of a story. Instead of explicitly revealing the correct interpretation as done in CBM-I, we used real-time fMRI (rt-fMRI) training to reinforce neural activity corresponding to a target interpretation. In our study, instead of one interpretation being positive and one negative, both of the two possible story interpretations were negative and differed in which characters were to blame for an unhappy marriage. Neither interpretation was "correct", in that there was no ground truth in the ambiguous story. Each subject's target interpretation was instead determined by random assignment before training. Ultimately, our goal was to induce different story interpretations according to random group assignment.

We based our experiment on prior work showing that neural responses to a story vary depending on how that story was interpreted (Yeshurun et al., 2017). In this study, Yeshurun et al. (2017) explicitly instructed subjects to adopt one of two different interpretations before listening to the ambiguous story in the fMRI scanner. This manipulation ensured that the two groups of subjects would interpret the story in different ways, allowing the authors to look for neural signatures of interpretation that were shared within each group. Indeed, the authors successfully identified neural regions that accurately predicted the interpretation group of held-out subjects. Within these regions, the parts in the story that varied the most in meaning based on interpretations were more likely to yield higher neural classification accuracy. For further details regarding the results, see Yeshurun et al. (2017).

In our current study, we designed an offline classifier using the same data published in Yeshurun et al. (2017). For detailed methods regarding how we built this classifier, see Appendix B. Then, while real-time subjects listened to the same story in the fMRI scanner, we used the offline classifier to decode which interpretations the subjects were most likely thinking about in a given moment, and then provided intermittent feedback to push the subjects toward their assigned interpretations.

After listening to the story, subjects answered questions to assess if neurofeedback successfully biased story interpretations. Thus, the critical comparison was between each subject's actual interpretation at the end of neurofeedback and the target interpretation determined by random group assignment. If neurofeedback was successful, subjects in each of the assigned groups would be more likely to adopt the target interpretation of that group, causing interpretations between assigned groups to differ. Since the success of neurofeedback depended on both (1) our ability to decode interpretations based on neural activity, and (2) the subjects' own ability to understand and manipulate their neural responses, our analyses described below assess success within each of these domains.

As outlined above, our primary goal was to alter final interpretations between the assigned groups. As this study represents the first iteration of this project, we have not yet included a control group in our design. However, for each of the experiments described below, we look at group-level variance in neurofeedback as a preliminary control measure.

4.2 Experiment 1

In Experiment 1, we tested an initial version of our neurofeedback task and successfully modified story interpretations across assigned groups. However, we also found that subjects

within an assigned group received very similar neurofeedback scores. As such, it is unlikely, that individualized neurofeedback produced this result.

4.2.1 Methods

4.2.1.1 Participants

Nineteen subjects consented to participate in this study. Two subjects did not understand the task and were excluded from analyses. Data from the remaining 17 subjects were retained for further analysis (n female = 10, n left-handed = 1, mean age = 22.4 years). Subjects received monetary compensation for their participation, including an additional bonus based on their neurofeedback performance ($20 maximum). The study was approved by Princeton University's Institutional Review Board.

4.2.1.2 Stimuli

All subjects listened to a 12-minute adapted version of "Pretty Mouth and Green My Eyes" by J. D. Salinger, read by a professional actor. The same stimulus was used in Yeshurun et al. (2017). The audio began with 18 s of music and ended in silence. See Yeshurun et al. (2017) for further details on the audio used.

The story begins when Arthur calls his friend, Lee, after returning home late at night without his wife, Joanie. Arthur is concerned about Joanie's whereabouts. Lee is at home in bed with a woman, although the narrator purposefully leaves her identity ambiguous. Yeshurun et al. (2017) created two interpretation groups by briefing subjects on this woman's identity *before* they heard the story. The two interpretation groups were:

1. **cheating**: Joanie is the woman in Lee's bed. In this case, Joanie is cheating on her husband, Arthur, with Lee.

2. **paranoid**: Lee's girlfriend, Rose, is the woman in Lee's bed. In this case, Arthur is paranoid that Joanie is cheating on him and is bothering Lee late at night.

In our present study, subjects were also randomly assigned to an interpretation group (cheating n = 9, paranoid n = 8), but we did not tell subjects which interpretation we assigned as their target interpretation. Instead, we told all subjects about the two possible interpretations before they listened to the story. The subjects' task was to figure out which of the two interpretations was correct through neurofeedback.

We presented a visual stimulus to deliver neurofeedback (Figure 4.1). While listening to the story, subjects viewed a rectangle on the screen. For most of the story, the rectangle was gray, indicating that the subject should simply continue listening. Neurofeedback was estimated based on brain activity during specific points of the story, which we refer to as *recording stations*. During these recording stations, the rectangle turned red to signal that brain activity was being *recorded*. We did this so that subjects were aware of when we were analyzing their brain activity. This served to help subjects interpret each neurofeedback score according to their previous mental state, which is an advantage of using intermittent over continuous neurofeedback (Oblak et al., 2017).

After each recording station, the classifier preprocessed their neural response and estimated the probability that the subject belonged to either interpretation group. The probability for the assigned group was directly used as the neurofeedback score. The rectangle then displayed the neurofeedback score; a horizontal line through the center (chance classification) indicated the threshold to earn any extra money and the filled area corresponded to the score. If sub-

jects scored above threshold, the score was shown in green with the monetary reward underneath (e.g., a subject assigned to the cheating group would receive $0.67 for classification p(cheating) = 0.67). Otherwise, the score was shown in gray with $0.00 below the rectangle.

By thresholding the value needed to earn rewards, subjects were not rewarded unless their neural activity was more similar to the assigned interpretation group than that of the other interpretation group. When subjects scored under threshold, the proportion shown by the gray rectangle allowed subjects to still get a continuous estimate of their score. Altogether, subjects completed 4 runs of neurofeedback.

To assess final interpretations after all runs, subjects answered the same questions from Yeshurun et al. (2017). The 39 questions contained 27 comprehension questions (e.g., *What was the girl doing when the phone rang?*) and 12 questions that directly probed interpretation (e.g., *Did you think Joanie was cheating on Arthur?*). Additionally, subjects provided numerical ratings (spanning 1-5) on opinions, such as how much they empathized with the characters, enjoyed the story, thought neurofeedback helped, etc.

4.2.1.3 Procedure

On the first visit, subjects consented to the experiment and underwent initial structural and functional scans for offline registration. Because our offline classifier was trained in standard Montreal Neurological Institute (MNI) space, registering each subject's brain to MNI space allowed us to apply the classifier in real-time. For registration details, see Methods, Sections 4.2.1.5 and 4.2.1.7.

On the second visit, (at least one day later, mean delay = 3.1 days), subjects returned for neurofeedback training. Before training began, subjects were randomized in a double-blind

fashion into the cheating or paranoid interpretation group. Group assignment was done via a Python script that saved assignment as a text file to be used later during neurofeedback. This way, both the experimenter and participant were blind to group assignment.

Note: We emphasized the following in our instructions:

- Your mission is to figure out the truth behind a phone conversation and solve the mystery
- The neurofeedback scores will reflect how well you're interpreting the story; a higher score means you are closer to the correct interpretation

Then, subjects listened to the story 4 times with neurofeedback. Once scanning was completed, subjects answered the story questions. Figure 4.1 illustrates the full experimental design and the neurofeedback stimuli.

4.2.1.4 Data acquisition

All scanning was acquired with a 3T MRI scanner (Siemens Skyra) and a 64-channel head coil. Sequences were matched to Yeshurun et al. (2017) as much as possible. Both scan sessions began with a Siemens scout scan for automatic alignment to the anterior commissure-posterior commissure (ACPC) axis. On Visit 1, we then collected a high-resolution, T1-weighted magnetization-prepared rapid acquisition gradient-echo (MPRAGE) anatomical scan (repetition time (TR) = 2,300 ms, TE = 3.08 ms, flip angle = 9°, resolution = 0.86 x 0.86 x 0.9 mm, FOV = 220 mm^2) to register each subject's brain to standard MNI space.

To increase brain coverage during scanning, the functional scans used a T2*weighted echo-planar imaging sequence (1.5 s TR, 28 ms echo time, flip angle =64° x 3 x 4 mm voxel size,

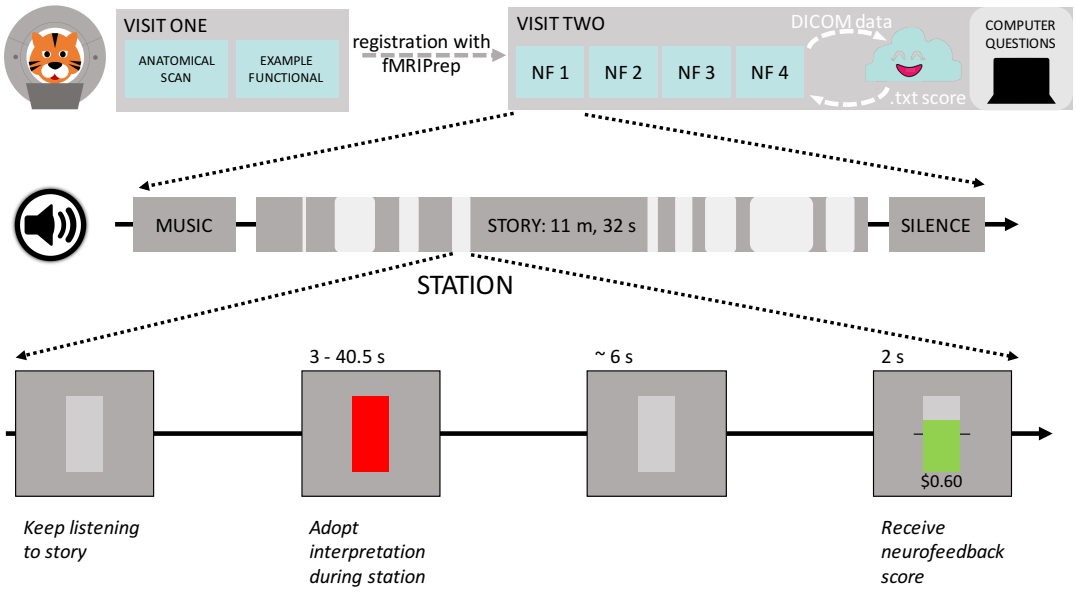


Figure 4.1: Design for the experiment. On Visit 1, a high-resolution anatomical scan and a short functional scan were acquired for registration to to MNI space with fMRIPrep. At least one day later, subjects returned for 4 runs of story listening with neurofeedback training. Each run was about 12 minutes long, and had the same audio stimulus used in Yeshurun et al. (2017). Each run included 9 *recording stations* when brain activity was analyzed. Stations varied in length between 3 - 40.5 seconds. After each station, the neurofeedback score was displayed and subjects received a reward if their neural classification was consistent with their group assignment. Afterwards, subjects answered questions outside of the scanner to assess their final interpretation of the story.

64 x 64 matrix, 192 mm field of view, 27 slices, 0-mm gap between slices, interleaved slice-acquisition).

No fieldmap scans were collected, as we were trying to best match offline and real-time preprocessing, and we would not be able to do so if the offline data included distortion correction but real-time data did not.

4.2.1.5 Image registration

To register all subjects to MNI space (between Visits 1 and 2), we used *fMRIPprep* 1.2.3 (Esteban et al., 2018b; Esteban et al., 2018a; RRID:SCR_016216), which is based on *Nipype* 1.1.6-dev (Gorgolewski et al., 2011; Gorgolewski et al., 2018; RRID:SCR_002502). The following description of preprocessing was generated with fMRIPrep.

Anatomical data preprocessing

> The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) using `N4BiasFieldCorrection` (ANTs 2.2.0; Tustison et al., 2010), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped using `antsBrainExtraction.sh` (ANTs 2.2.0), using OASIS as target template. Brain surfaces were reconstructed using `recon-all` (FreeSurfer 6.0.1, RRID:SCR_001847, Dale et al., 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (RRID:SCR_002438, Klein et al., 2017). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009, RRID:SCR_008796) was performed through nonlinear registration with `antsRegistration` (ANTs 2.2.0, RRID:SCR_004757, Avants

et al., 2008), using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using `fast` (FSL 5.0.9, RRID:SCR_002823, Zhang et al., 2001).

Functional data preprocessing

For the 1 BOLD run per subject, the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. The BOLD reference was then co-registered to the T1w reference using `bbregister` (FreeSurfer) which implements boundary-based registration (Greve & Fischl, 2009). Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) were estimated before any spatiotemporal filtering using `mcflirt` (FSL 5.0.9, Jenkinson et al., 2002). The BOLD time series was resampled onto the original, native space by applying a single, composite transform to correct for head-motion. The BOLD time series was then resampled to MNI152NLin2009cAsym standard space. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*.

All resamplings were performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices and co-registrations to anatomical and template spaces). Gridded (volumetric) resamplings were performed using `antsApplyTransforms` (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964).

Many internal operations of *fMRIPrep* use *Nilearn* 0.4.2 (Abraham et al., 2014, RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in *fMRIPrep*'s documentation.

4.2.1.6 Cloud configuration

The goal of our rt-cloud project was to minimize processing done by a local Linux machine in the scanner control room, and thus, decrease individual computing requirements for new researchers. As part of rt-cloud development, this study served as a use case to create a generic software framework for real-time cloud processing. Once DICOM files arrived at the local Linux machine, they were kept in memory as bytes and immediately sent to the cloud computer. On the cloud server, the BOLD data were then registered to MNI space, preprocessed, and scored for a neurofeedback value. Finally, a text file was returned to the local Linux machine to update the display. Control of the rt-fMRI pipeline was accessed via a website that was accessible to any computer (with the necessary security permissions). Figure 4.2 illustrates both the delineation between local vs. cloud processing, as well as the real-time processing steps.

4.2.1.7 Real-time pipeline

BOLD data were acquired in the subject's native space, but the classifier was trained on data in standard MNI space. Therefore, we had to transform each incoming BOLD volume to MNI space in real-time. To do this, we combined previous offline registration steps from fMRIPrep with real-time registration of each new BOLD volume. In real-time, we used `mcflirt` (FSL 5.0.9, Jenkinson et al., 2002) to register each incoming BOLD volume with

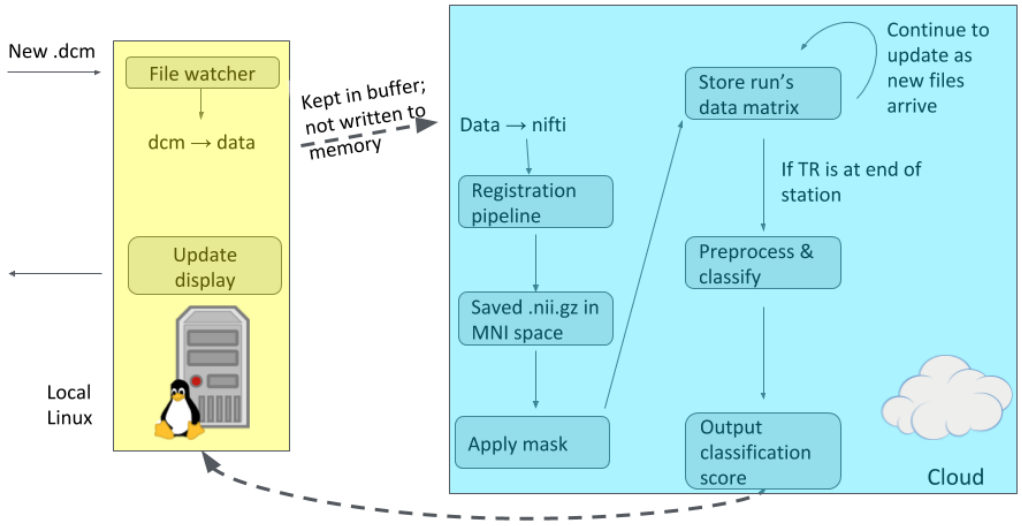


Figure 4.2: Cloud processing configuration. Data was sent as bytes to the remote cloud server, and a text file containing the neurofeedback score was returned to the local computer. Specific processing steps are also shown.

the example functional image acquired on Visit 1. We combined this transformation matrix with the two transformation matrices calculated offline with fMRIPrep: (1) the transformation from functional → T1w space, and (2) the transformation from T1w space → MNI space. All transformations were concatenated and performed in a single step using `antsApplyTransforms` (Avants et al., 2008). Details of this pipeline performed on the cloud server are shown in Table 4.1.

4.2.1.8 Real-time classification

We trained 9 separate offline classifiers corresponding to the 9 *recording stations*, or moments in the story when we analyzed brain activity. See Appendix B for details on classifier construction. During real-time runs, each station classifier predicted the likelihood that the subject

Step	Description	Software used
1	Convert from DICOM bytes to save as a NIfTI image	nilearn (Abraham et al., 2014)
2	Calculate transformation matrix between current TR and reference image	`mcflirt` FSL 5.0.9 (Jenkinson et al., 2002)
3	Convert transformation matrix .mat → .txt	c3d_affine_tool Convert3D (Yushkevich et al., 2006)
4	Transform current TR to MNI space by concatenating all transformations	antsApplyTransforms ANTs 2.1.0 (Avants et al., 2008)
5	Read and mask NIfTI data in MNI space, store in memory	nibabel (Brett et al., 2019)
If current TR is the last TR in a station:		
6	Z-score data within voxels from the beginning of the story to that TR (and remove voxels that were outside of the brain by setting activity to 0 if value was < 100 or std < 0.001)	custom Python code
7	Subtract the average signal from the offline data set	custom Python code
8	Classify station and generate NF score	Scikit-learn (Pedregosa et al., 2011)
9	Save score as text file	rt-cloud software
10	Send text file back to local Linux to update display	rt-cloud software

Table 4.1: List of preprocessing steps completed on the cloud server for real-time generation of neurofeedback scores.

was interpreting the story in line with the cheating or paranoid interpretation. We generated this probability estimate using scikit-learn's predict_proba function (Pedregosa et al., 2011). The neurofeedback score displayed to each subject was:

$$score_{final} = \begin{cases} p_{cheating}, & \text{if assigned group} = cheating \\ 1 - p_{cheating}, & \text{if assigned group} = paranoid \end{cases} \tag{4.1}$$

Thus, the higher the neurofeedback score, the more similarly the real-time subject's neural response matched those of previous subjects in the same assigned interpretation group.

4.2.1.9 Story comprehension and interpretation scores

We analyzed the 39 story questions by dividing them into comprehension (27 questions) and interpretation-specific questions (12 questions) as done in Yeshurun et al. (2017). We did not analyze one of the interpretation-specific questions that asked for Lee's girlfriend's name, since we informed subjects that Lee had a girlfriend named Rose prior to the experiment. All subjects answered this question correctly, regardless of their interpretation.

To score the interpretation-specific questions, we assigned a +1 score to each answer corresponding to the cheating interpretation, and a -1 score to each answer corresponding to the paranoid interpretation. To calculate a final score, we totaled all interpretation-specific scores and divided by the total number of questions. Thus, a +1 indicates answering all questions consistent with a cheating interpretation, and a -1 indicates answering all questions consistent with a paranoid interpretation.

To collapse across assigned groups and compare "correctness" of final interpretations (i.e., consistency with group assignment versus inconsistency with group assignment), we rescaled

the interpretation-specific scores. To do this, we multiplied the scores for all subjects assigned to the paranoid group by -1. In this framework, a +1 score indicates answering all questions consistent with one's group assignment (correct), and a -1 score indicates answering all questions inconsistent with one's group assignment (incorrect).

4.2.1.10 Empathy Ratings

After answering the story questions, subjects rated how much they empathized with Arthur, Lee, Joanie, and "the girl" (whomever they presumed was the mysterious woman in Lee's bed). We expected those who assumed the cheating interpretation to empathize with Arthur, and not with Lee and Joanie. Likewise, we expected those who believed the paranoid interpretation to feel more empathy for Lee and Joanie, instead of Arthur. To compute one score that captured the empathy bias, we subtracted the difference between each subject's empathy ratings for Arthur and Lee. When computing "correct empathy" scores, we again multiplied the empathy difference for all subjects assigned to the paranoid group by -1. Thus, positive "correct empathy" scores indicate that the subject empathized with Arthur and Lee in a manner that was consistent with the assigned interpretation group.

4.2.2 Results

If neurofeedback successfully modified interpretations, then we expected the randomly assigned groups to show final interpretations consistent with their group assignment. Additionally, if the neurofeedback was specific to the individual subject, we expected neurofeedback scores to vary on an individual level.

4.2.2.1 Story comprehension and interpretation

All subjects understood the story as evidenced by near 100 percent accuracy on comprehension questions (Figure 4.3). The two groups did not differ in comprehension ($t(15) = -0.42$, $p = 0.68$).

The groups significantly differed in how they answered interpretation-specific questions. As shown in Figure 4.4, subjects in the paranoid group were more likely to select the paranoid interpretation, and subjects assigned to the cheating group were more likely to select the cheating interpretation. Thus, interpretation scores significantly differed between the assigned groups (one-tailed $t(15) = -1.95$, $p = 0.035$).

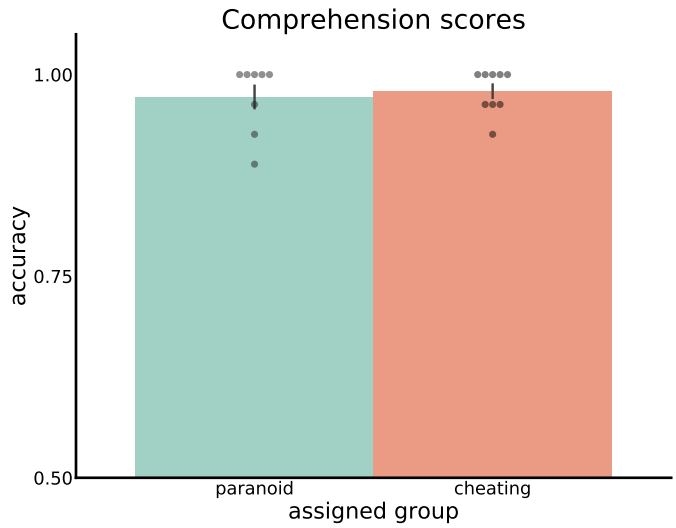


Figure 4.3: All subjects understood the story. Each dot represents a subject in either the paranoid group **(green)** or cheating group **(peach)**. Error bars represent ± 1 s.e.m.

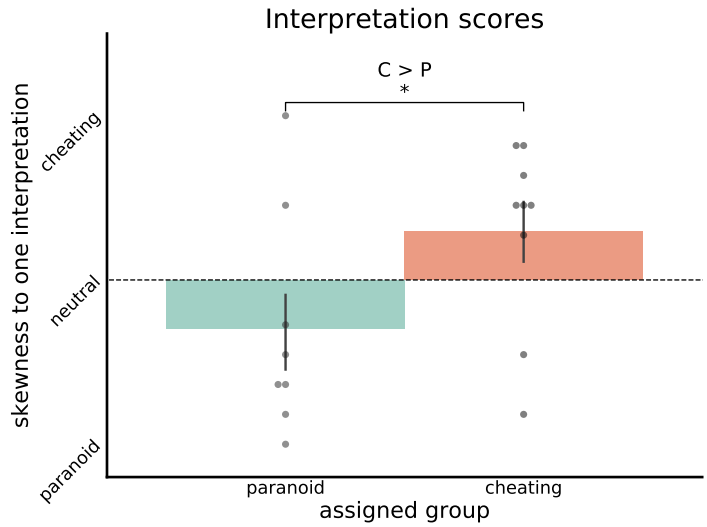
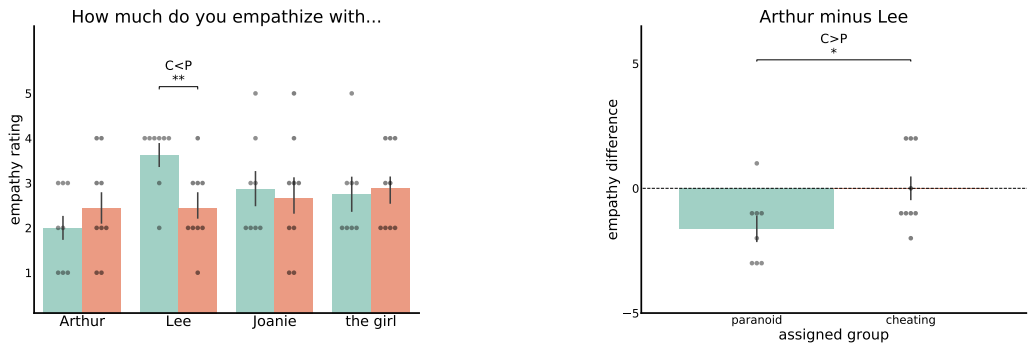


Figure 4.4: Interpretation scores differed significantly by assigned group. Each dot represents a subject in either the paranoid group **(green)** or cheating group **(peach)**. Error bars represent ± 1 s.e.m. * = $p < 0.05$

4.2.2.2 Empathy ratings

As predicted, subjects randomly assigned to the paranoid group reported significantly more empathy for Lee than those assigned to the cheating group (one-tailed $t(15) = -2.96$, $p = 0.0049$). There were no significant group differences in empathy for the other characters (all other $p \geq 0.20$). The empathy difference (Arthur $-$ Lee) was also significantly different between groups (one-tailed $t(15) = 2.23$, $p = 0.021$). Figure 4.5 plots the empathy ratings by group.

As both the explicit interpretation questions and empathy ratings probed final beliefs, we examined whether these two measures were related. There was a nonsignificant positive correlation between empathy rating (Arthur $-$ Lee) and interpretation score (Pearson $r = 0.33$, $p = 0.19$; Figure 4.6). Thus, we cannot conclude that subjects answered the explicit interpre-



(a) **(b)**

Figure 4.5: Empathy ratings for **(a)** all characters and **(b)** Arthur — Lee only, separated by assigned group (**green** = paranoid group; **peach** = cheating group). **(a)**: Empathy for Lee differed significantly based on group assignment. **(b)**: Arthur — Lee empathy also differed significantly by assigned group. Circles represent individual subjects. Error bars denote ± 1 s.e.m. ** = $p < 0.01$; * = $p < 0.05$

tation questions and empathy ratings consistently.

4.2.2.3 Classification

The purpose of providing real-time neurofeedback was to give subjects individualized scores based on their own neural responses to the story. Thus, if scores did not vary across subjects, then neurofeedback would not have been necessary.

To understand how neurofeedback pushed subjects to the assigned interpretation, we averaged the classification probability outputs within each group and plotted the average by run in Figure 4.7. The two groups showed very similar neural trajectories throughout all 4 runs. In other words, the classifier predicted very similar probabilities that subjects were thinking of the cheating or paranoid interpretation at a given point in the story. This was unexpected because we successfully modified self-reported interpretations across the assigned groups.

To understand how similar classification outputs could still lead to behavioral differences,

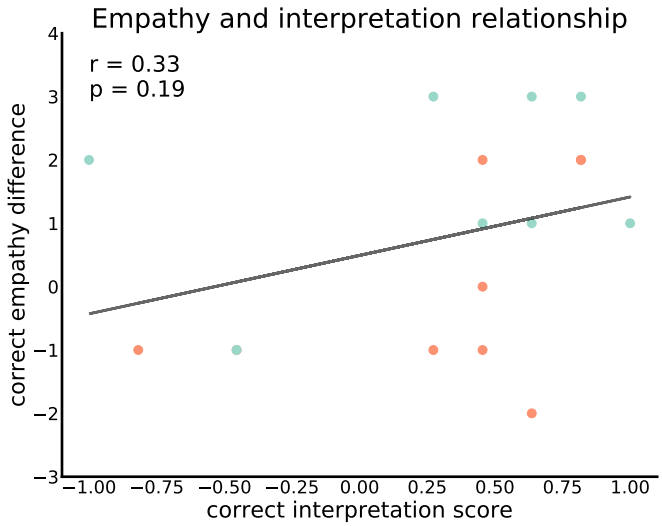


Figure 4.6: Correlation between interpretation score and Arthur — Lee empathy difference, rescaled by "correctness". The black line shows the line of best fit. Dots represent individual subjects; colors denote randomly assigned groups (**green** = paranoid group; **peach** = cheating group).

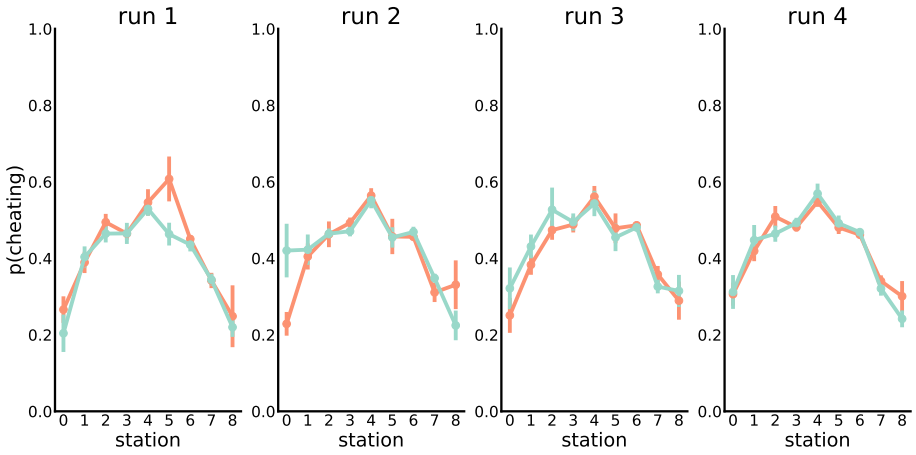


Figure 4.7: Real-time neurofeedback probability outputs that subjects were adopting the cheating interpretation. As shown, classification scores followed a consistent trajectory for all subjects. Color denotes assigned group (**green** = paranoid group; **peach** = cheating group). Error bars represent ± 1 s.e.m.

we plotted the neurofeedback score subjects received during neurofeedback instead of classification probabilities. Figure 4.8 shows that the similar neural patterns (and presumably the interpretation trajectories) shared across subjects led to differences in when each group received rewards. The paranoid group received larger rewards in the beginning and end of the story, while the cheating group received larger rewards in the middle of the story.

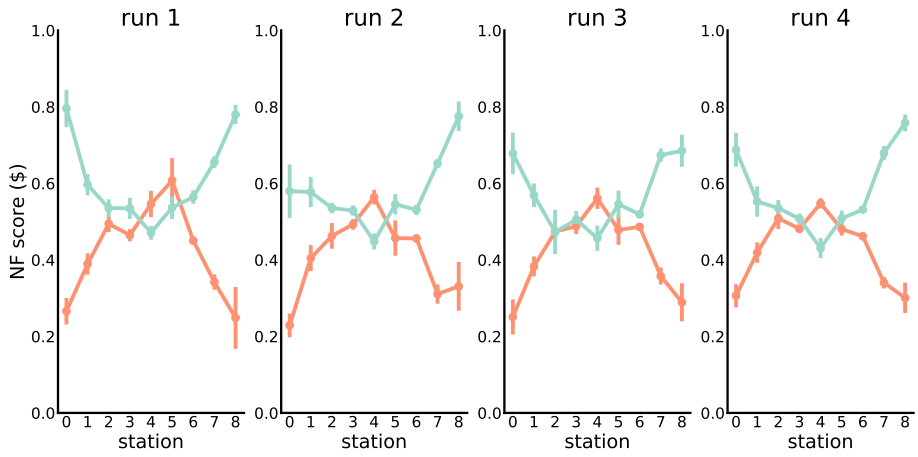


Figure 4.8: Neurofeedback scores delivered to subjects during real-time neurofeedback. As shown, the timing of rewards was opposite across groups. This could have led to successful differences in interpretations, despite similar neural classifications. Colors denote assigned group (**green** = paranoid group; **peach** = cheating group). Error bars represent ± 1 s.e.m.

We realized that the story's narrative structure could be biasing our real-time subjects in a way that was different from subjects who had been told an interpretation beforehand, as was the case for the subjects used to train the classifier. For example, at the end of the story, Arthur tells Lee that Joanie has returned home. For subjects who weren't told what to think beforehand, this most likely biased listeners to think that Arthur was paranoid. Thus, when everyone adopted the paranoid interpretation during that recording station (as indicated by the low cheating probability in station 7) the paranoid group was rewarded and the cheating

group was not. Therefore, the groups were likely swayed to their assigned interpretation based on the feedback score, not unique patterns of brain activity.

Since all subjects within a group were received very similar feedback scores, we hypothesize that if we had switched one subject's neurofeedback scores for another subject's in the same assigned group, this would have yielded similar behavioral effects. Thus, we do not think that individualized neurofeedback was necessary in generating differences in interpretations between assigned groups.

4.2.3 Discussion

A gold standard in rt-fMRI analysis is demonstrating that the effects achieved were due specifically to regulating one's own brain via neurofeedback (Sorger et al., 2019). A yoked control group that receives neurofeedback based on other subjects' brain activity can provide this proof (deBettencourt et al., 2015). While we did not include a yoked control group, the lack of individual variation within group allows us to predict with confidence that yoked subjects in each group would have been similarly driven to the assigned interpretation. It does not appear that individualized neurofeedback was necessary to produce the observed group differences.

While designing our real-time study, we had assumed based on Yeshurun et al. (2017) that there would be two distinct time courses for the two interpretations throughout the story. However, discriminability between interpretations became less robust when subjects were not told what to believe beforehand. When subjects were allowed to form their own interpretations, they were naturally pulled to each interpretation at different points in the story. This was most problematic at the end of the story, when the classifier was the most accurate

in predicting group labels in the offline data (see Appendix B, Figure B.3). As mentioned above, Arthur's revelation that Joanie returned home would have induced stark differences in subjects already believing one of the two contexts (as in Yeshurun et al., 2017), but pushed uncertain real-time subjects to the more-likely interpretation that Arthur was paranoid.

We aimed to correct several of Experiment 1's flaws in Experiment 2, discussed next.

4.3 Experiment 2

In Experiment 2, our goal was to change the neurofeedback such that scores were different for subjects within the same group. This way, if we were able to replicate the same behavioral effects as in Experiment 1, we could better demonstrate the potential of using rt-fMRI for interpretation training. We recognize that we will ultimately need a yoked control group to fully prove that individualized rt-fMRI was necessary to modify story interpretations. However, before recruiting this cohort, we first aimed to prove that our paradigm was still able to change story interpretations when each subject received entirely different neurofeedback scores. Additionally, to assess if we were accurately classifying real-time interpretations, we had subjects report their current beliefs preceding each recording station. This gave us a ground truth to assess our decoding ability. Table 4.2 summarizes the problems from Experiment 1 and how we solved them in Experiment 2.

4.3.1 Methods

4.3.1.1 Participants

Twenty-two new subjects consented to participate in this experiment, but 2 subjects did not return for their second visit. Thus, 20 subjects were included in the analyses (n female = 12,

Experiment 1 Problem	Experiment 2 Solution
Stations were strongly biased to one interpretation, meaning that there was little variation across subjects	We normalized classification probabilities based on the mean and standard deviation in Experiment 1
The two last stations, in particular, were very biased toward the paranoid interpretation	We removed the last two stations
We had no way to verify if subjects were completing the task or if classification was accurate	We explicitly asked subjects for their current beliefs before recording stations
Subjects were told that there was one true interpretation	We told subjects that the story was purposefully ambiguous and they should learn according to neurofeedback only

Table 4.2: List of changes made when designing Experiment 2 based on the problems in Experiment 1.

n left-handed = 2, mean age = 20.5 years). Of the 20 subjects, 10 were randomly assigned to the cheating group and 10 were randomly assigned to the paranoid group. Group assignment was double-blind to the experimenter and participant, as described in Experiment 1. Group assignment only changed the final neurofeedback score presented after each station.

4.3.1.2 Stimuli

All subjects listened to the same audio file and saw the same visual stimulus as described in Experiment 1. This time, however, subjects received a *probe* prior to each station, asking them to choose an interpretation "lens" for the the upcoming station. Subjects were free to choose either interpretation at any time, as long as they were trying to maximize their scores. Subjects pressed their index or middle fingers to indicate their choice of either the cheating or paranoid interpretation (left/right position of choices was counterbalanced across subjects).

A couple of seconds after the probe, the *recording station* began. Neurofeedback scores and monetary rewards were displayed the same as in Experiment 1. See Figure 4.9 for example stimulus displays.

After completing neurofeedback runs, subjects answered the same questions as in Experiment 1 as well as a survey to record perception and strategies.

4.3.1.3 Procedure

Subjects completed the same experimental procedure as in Experiment 1. On Visit 1, subjects consented to the experiment and underwent initial structural and functional scans for offline registration.

On Visit 2 (at least one day later, mean delay = 4.6 days), subjects returned to complete 4 runs of neurofeedback training (1 subject only completed 2 runs due to technical problems).

Note: This time, we emphasized the following in our instructions before scanning began, to make sure that subjects were attending to neurofeedback throughout the entire experiment:

- The story was written to be purposefully ambiguous, instead of there being one true interpretation.
- You should try to interpret the story based on your neurofeedback, not what you think was intended by the author.
- Even if you think that the correct interpretation was revealed in the story, characters might be lying. Keep using neurofeedback to guide your interpretations.

- The neurofeedback scores reflect how clearly and correctly your brain is interpreting the story. Low scores can indicate either (1) noisy signals (e.g., poor focus) and/or (2) incorrect interpretations.

- The neurofeedback score only represents the neural information recorded during the station. We keep track of your responses to the probes for analysis purposes, but the neurofeedback score you see is only based on your brain activity.

The full set of instructions used in Experiment 2 is included in Appendix B, Figure B.5.

After scanning, subjects completed the story questions and survey. Figure 4.9 illustrates the full experimental design and the neurofeedback display.

4.3.1.4 Data acquisition

We used the same acquisition parameters as Experiment 1.

4.3.1.5 Classification

We learned from Experiment 1 that real-time classification probabilities were likely to be pushed along similar trajectories, as subjects tended to share a common interpretation of the story as it progressed. To introduce more variability between subjects and remove station biases, we adjusted classification outputs using data from Experiment 1.

For each subject in Experiment 2, at a given station *st*, we applied the following formula to transform the classifier's cheating prediction probability $p(c)$ to a neurofeedback score:

$$score_{cheating} = \frac{p(c)_{st} - \mu_{st}}{3\sigma_{st}} + 0.5 \tag{4.2}$$

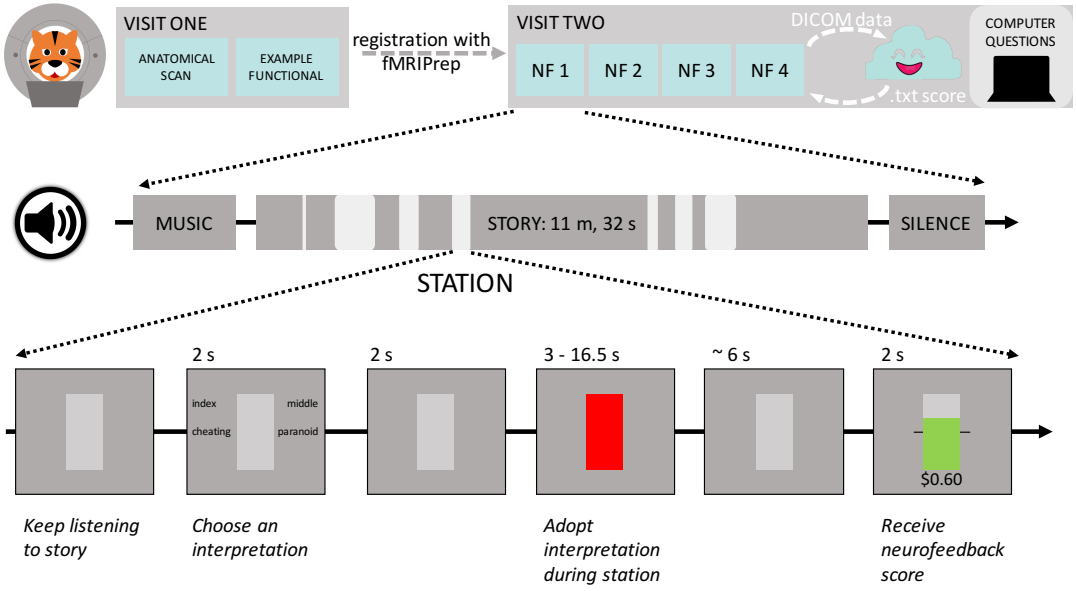


Figure 4.9: Design for the experiment. Subjects came in for an initial scanning session to get their anatomical and functional data registered to MNI space with fMRIPrep. At least one day later, subjects came back for 4 runs of neurofeedback. Each neurofeedback run was about 12 minutes long, and had the same audio stimulus used in Yeshurun et al. (2017). Each run included 7 *recording stations*, or moments when brain activity was analyzed. Stations varied in length between 3 - 16.5 seconds. Before each station, subjects answered a *probe* to report which interpretation they would adopt in the upcoming station. After each station, the neurofeedback score informed subjects of how consistent their interpretation was with their assigned group, solely based on brain activity (not their behavioral response to the probe). Afterwards, subjects answered questions outside of the scanner to assess their final interpretation of the story.

where μ_{st} and σ_{st} were that station's mean and standard deviation, respectively, collapsed across all subjects and runs in Experiment 1. We included a scaling factor of 3 so that the z-scored differences would scale between [-0.5,0.5]. Thus, by adding 0.5 to this ratio, we could set the range to be [0,1]. In case scores were much larger or much lower than station means, we added the thresholds:

$$score_{cheating} = \begin{cases} 0, & \text{if } score_{cheating} < 0 \\ 1, & \text{if } score_{cheating} > 1 \end{cases} \tag{4.3}$$

Finally, we generated a score based on group assignment with:

$$score_{final} = \begin{cases} score_{cheating}, & \text{if assigned group} = cheating \\ 1 - score_{cheating}, & \text{if assigned group} = paranoid \end{cases} \tag{4.4}$$

In this framework, subjects scored at the 0.5 threshold if their classification was equal to the mean of subjects in Experiment 1. To earn any reward, subjects had to score *above* that station's mean, in the assigned direction. Subjects would not receive any reward if classification was lower than that station's mean. This happened when (1) classification predicted the non-assigned interpretation group, *or* (2) classification predicted the assigned group, but the probability was lower than the mean. Because of this ambiguity, we informed subjects that their scores not only reflected their interpretation, but also the noisiness of their brain, so they should be focused in order to solve the task and maximize reward. Additionally, we switched from a support vector machine (SVM) classifier in Experiment 1 to a logistic classifier in Experiment 2 to ensure accurate prediction probabilities (sklearn solver='lbfgs',C=1).

4.3.2 Results

4.3.2.1 Story comprehension and interpretation scores

Comprehension accuracy indicated that all subjects understood the story (Figure 4.10), and there was no significant difference between assigned groups ($t(18) = 1.41$, $p = 0.17$).

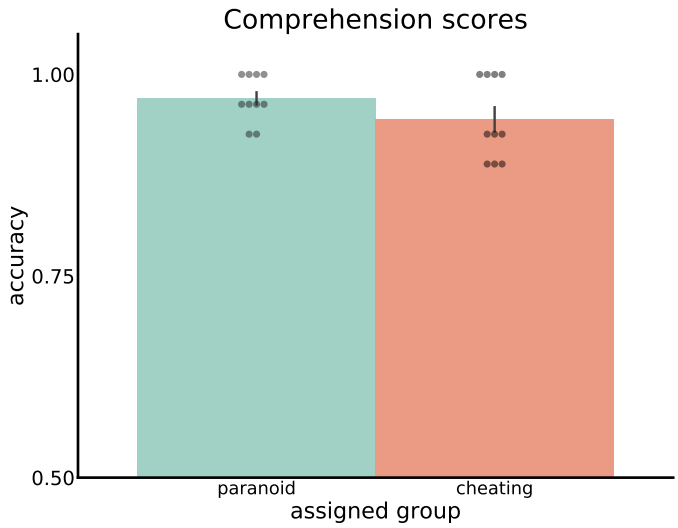


Figure 4.10: All subjects understood the story. Each dot represents a subject in either the paranoid group **(green)** or cheating group **(peach)**. Error bars represent ± 1 s.e.m.

Our neurofeedback manipulation did not push subjects to their assigned interpretation, as indicated by the interpretation scores. Figure 4.11 shows the interpretation scores by assigned group. Neither group showed significant bias toward their assigned target (all $p > 0.20$). Further, the groups did not differ in interpretation scores (one-tailed $t(18) = 0.062$, $p = 0.48$).

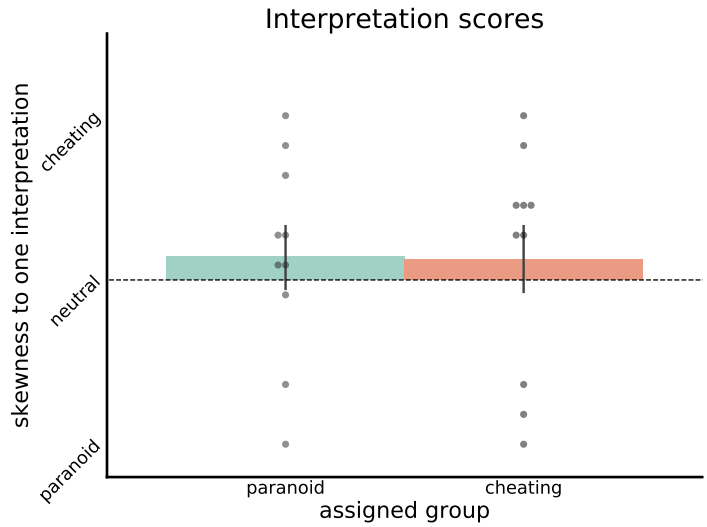
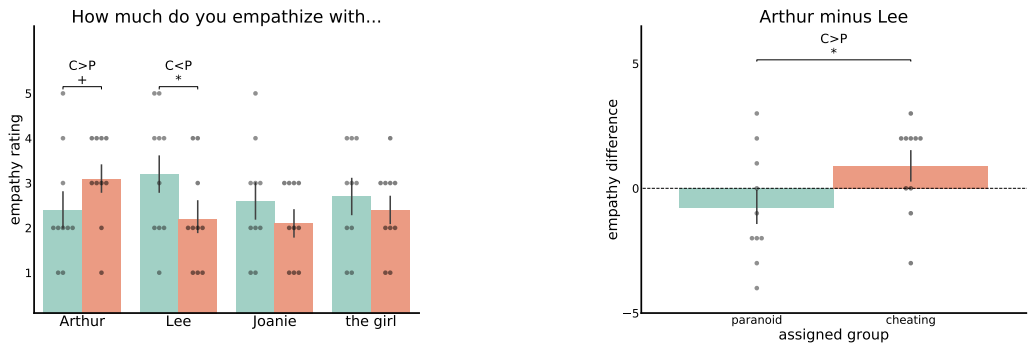


Figure 4.11: Interpretations were not modified by neurofeedback. Neither group was significantly pushed toward their assigned interpretation (**green** = paranoid group, **peach** = cheating group). Error bars represent ± 1 s.e.m.

4.3.2.2 Empathy ratings

The empathy ratings for some of the characters *did* differ by group in the "correct" direction of the assigned interpretations. There was a trend for subjects in the cheating group to have more empathy for Arthur than subjects in the paranoid group (one-tailed $t(18) = 1.38$, $p = 0.093$). Additionally, subjects in the paranoid group had more empathy for Lee than the subjects in the cheating group (one-tailed $t(18) = -1.76$, $p = 0.048$). The Arthur − Lee empathy difference was significantly different between the assigned groups, with those assigned to the cheating condition having more empathy for Arthur than Lee (one-tailed $t(18) = 1.84$, $p = 0.041$). Figure 4.12 plots the empathy ratings by assigned group.

Next, we examined if empathy ratings and interpretation scores were related. There was a significant positive correlation between the Arthur − Lee empathy difference and interpre-



(a) **(b)**

Figure 4.12: Empathy ratings for **(a)** all characters and **(b)** Arthur — Lee only, separated by assigned group (**green** = paranoid group; **peach** = cheating group). **(a)**: Empathy ratings differed significantly in the predicted direction for Lee based on group assignment. **(b)**: Arthur — Lee empathy also differed significantly by assigned group. Circles represent individual subjects. Error bars denote ± 1 s.e.m. * = $p < 0.05$; + = $p < 0.1$

tation scores (Pearson $r = 0.51$, $p = 0.021$; Figure 4.13). In other words, the more subjects empathized with Arthur over Lee, the more likely they were to endorse the cheating interpretation in response to the interpretation questions. Unlike in Experiment 1, this time empathy ratings and explicit interpretation responses were related.

4.3.2.3 Probe responses

Figure 4.14 shows the probe responses chosen, divided by assigned interpretation groups. While the average choices diverged between groups in the assigned or "correct" direction by the end of the last run, there were no significant differences between groups when correcting for multiple comparisons via Bonferroni correction and treating each station as a separate null hypothesis.

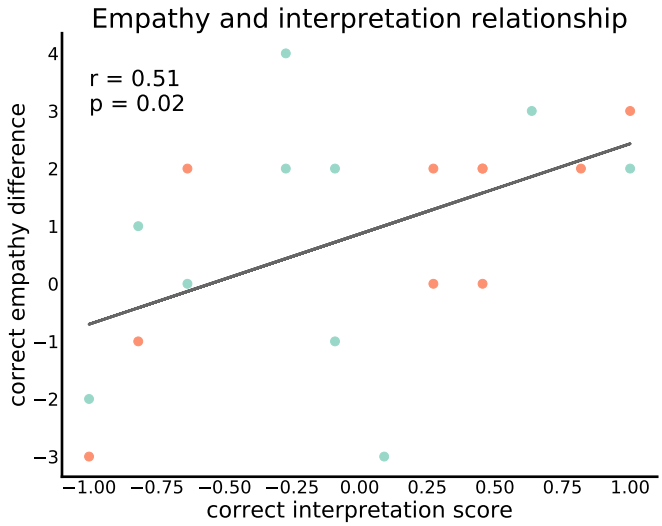


Figure 4.13: Responses to interpretation questions as a function of the empathy difference between Arthur and Lee, rescaled in terms of the "correct" beliefs, or how consistent that subject's scores were with group assignment. The black line shows the line of best fit. Colors denote randomly assigned groups (**green** = paranoid group; **peach** = cheating group).

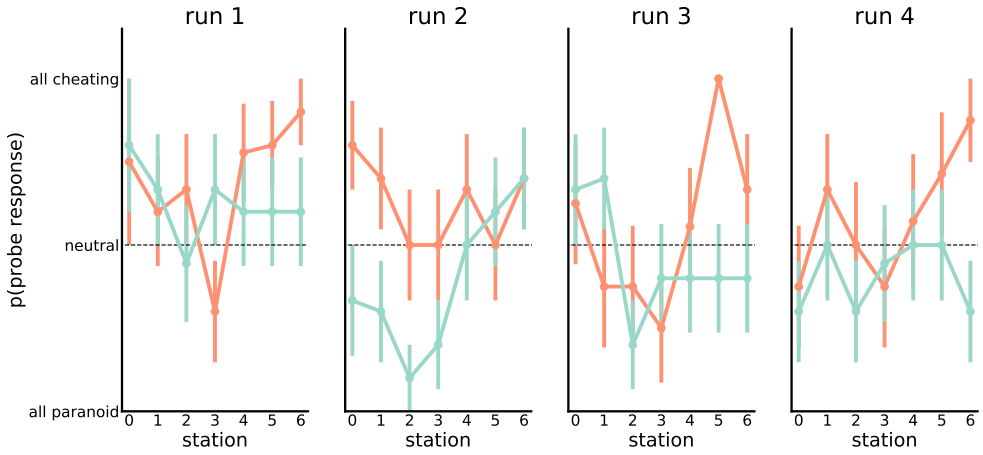


Figure 4.14: Average probe responses during neurofeedback, separated by assigned group (**green** = paranoid group; **peach** = cheating group). Responses diverged at the end of the last run, though not to a level of significance. Error bars represent ±1 s.e.m.

4.3.2.4 Classifier checks

To successfully use rt-fMRI to push subjects to the assigned interpretation, the following conditions had to be met:

1. Neurofeedback scores within group had to be variable, meaning that the subjects within each group were not receiving similar scores. While we know that this is not sufficient in proving the necessity of rt-fMRI in changing behavior, variance within groups must first be demonstrated before collecting future cohorts.

2. We had to accurately decode which interpretation subjects chose at each station during neurofeedback runs.

3. Subjects had to properly learn from the neurofeedback signal. Learning in this task involved listening to the story, attending to the neurofeedback scores, and subsequently altering one's neural response at each station.

First, we checked that there was sufficient variability in neurofeedback scores. If this was the case, we should not see correlated scores within and across groups. Figure 4.15 shows the neurofeedback scores, averaged by group. Unlike in Experiment 1 (Figure 4.8), rewards were not strongly anti-correlated between assigned groups. In the first run, there was a nonsignificant negative correlation between assigned groups (Pearson r = -0.41, p = 0.36). In the last run, the rewards were significantly positively correlated (Pearson r = 0.77, p = 0.043). Thus, if anything, the rewards on average became more similar between assigned groups from the first to the last run. Additionally, the implementations in Experiment 2 increased within-group variance. We increased the average within-group standard error of the mean from an average of 0.016 in Experiment 1 to 0.059 in Experiment 2.

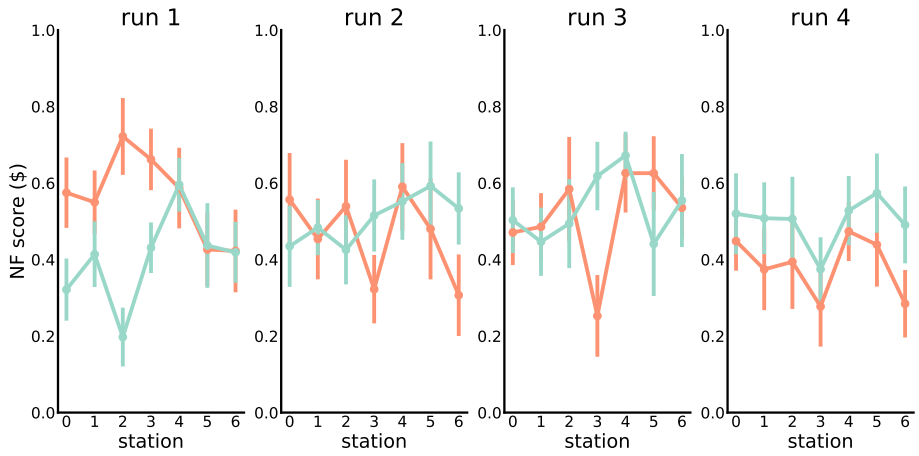


Figure 4.15: Neurofeedback scores delivered to subjects during real-time neurofeedback, divided by assigned group (**green** = paranoid group; **peach** = cheating group). Unlike in Experiment 1, neurofeedback scores were not anti-correlated between assigned group and showed variation between runs. Error bars represent ± 1 s.e.m.

To assess decoding accuracy, we analyzed how the probe responses before each station aligned with the corresponding classification output. If we were properly decoding interpretations in real-time, we would see higher cheating probabilities when subjects reported thinking of the cheating compared to paranoid interpretation. For each subject, we separated all classification probabilities by probe response. We averaged the cheating probability over all the times the subject chose the cheating interpretation and then averaged over all the times the subject chose the paranoid interpretation. We estimated *classifier accuracy* for each subject by subtracting the average cheating probability for cheating − paranoid probe responses. Thus, more positive (e.g., more accurate) differences indicate high cheating probability following cheating responses and low cheating probability (or high paranoid probability) following paranoid responses.

On average, there was a trend for larger cheating probability when subjects endorsed the cheating interpretation compared to the paranoid interpretation (one-tailed $t(19) = 1.40$, p

= 0.088); 11/20 of the subjects showed this pattern. Results are shown in Figure 4.16. The subject-level variation implied that some subjects received more accurate neurofeedback than others. *Note:* Here, we used actual classification probabilities, instead of adjusted scores.

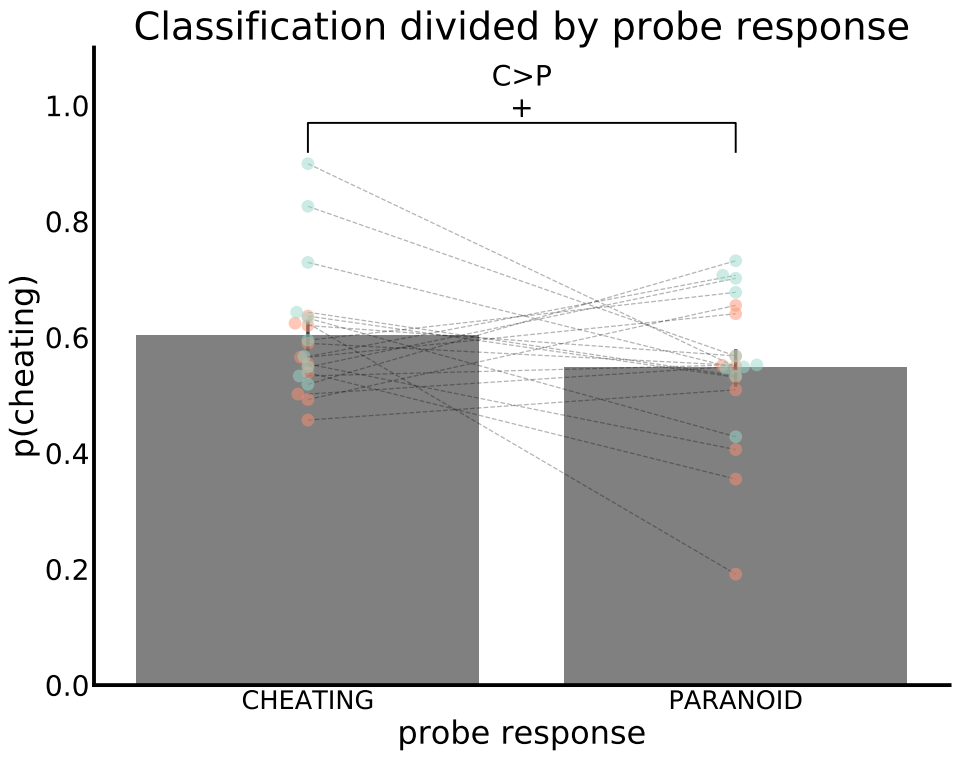


Figure 4.16: Cheating probability was higher when subjects endorsed the cheating interpretation compared to the paranoid interpretation, although this difference did not reach statistical significance. *Note:* the x-axis represents the probe responses at each station, *not* group. Lines connect each subject's average probability for each response. Circle colors denote group (**green** = paranoid; **peach** = cheating). Error bars represent ± 1 s.e.m. + = p < 0.1

If subjects were able to learn from neurofeedback, we would expect that more accurate neurofeedback scores would lead to stronger beliefs in the assigned or "correct" interpretation. Indeed, there was a significant positive relationship between interpretation scores and classifier accuracy (Pearson r = 0.56, p = 0.010), suggesting that subjects with the most accu-

rate neurofeedback were able to learn from the neurofeedback and were pushed toward the assigned interpretation (Figure 4.17).

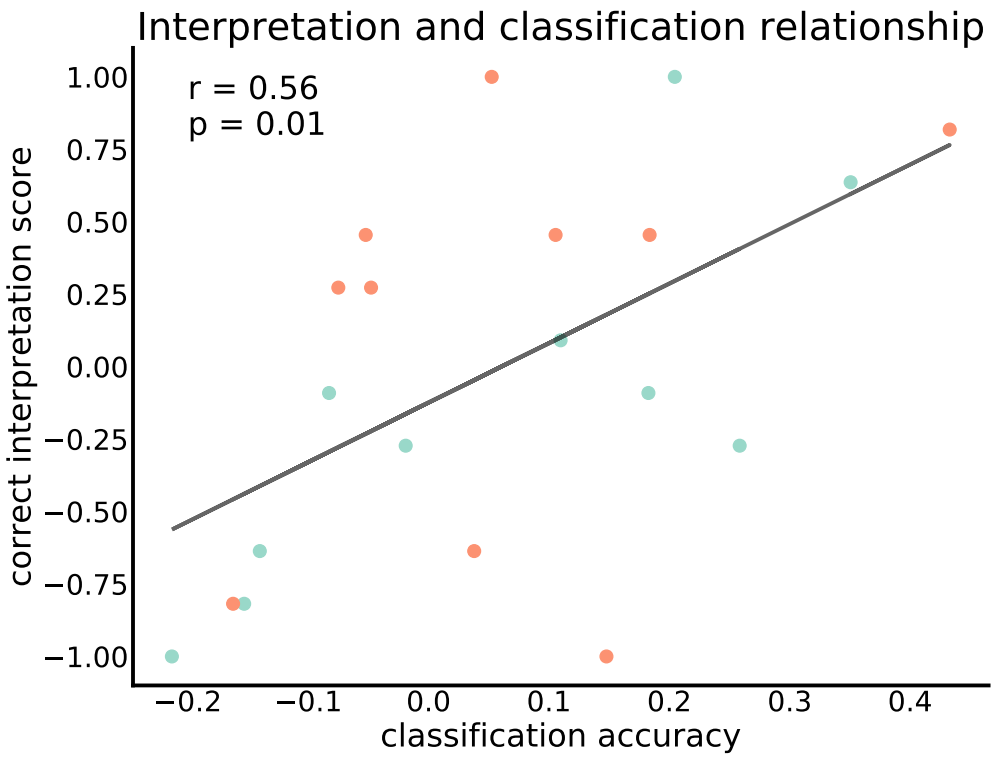


Figure 4.17: Correct interpretation plotted as a function of classifier accuracy. Classification accuracy and correct interpretation scores were positively correlated, suggesting that subjects with accurate neurofeedback were able to learn to adopt the assigned interpretation. The black line indicates the line of best fit. Individual dots represent each subject, with the color indicating group (**green** = paranoid; **peach** = cheating).

4.3.2.5 Dividing subjects by classification accuracy

Given that some subjects received more accurate neurofeedback, we assessed if classification accuracy influenced the behavioral outcomes described above. For these analyses, we median split each assigned group based on classifier accuracy. Thus, 5 subjects from the cheating

group and 5 subjects from the paranoid group were in the best group and 5 from each group were in the worst group. Next, we assessed behavioral measures divided by the best and worst classifier accuracy, instead of assigned group.

Story comprehension and interpretation scores

Comprehension scores did not vary significantly between classifier accuracy group ($t(18) = -1.41$, $p = 0.18$), shown in Figure 4.18.

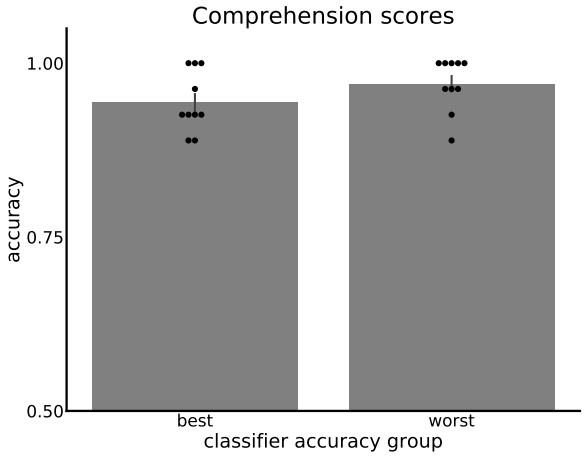


Figure 4.18: Comprehension scores split by classifier accuracy. Comprehension scores did not vary significantly between groups.

Unlike story comprehension, interpretation scores were significantly higher for the best compared to worst classifier accuracy group (one-tailed $t(18) = 2.44$, $p = 0.013$), which was expected based on the significant positive correlation between classification accuracy and correct interpretation. Figure 4.19 shows the results.

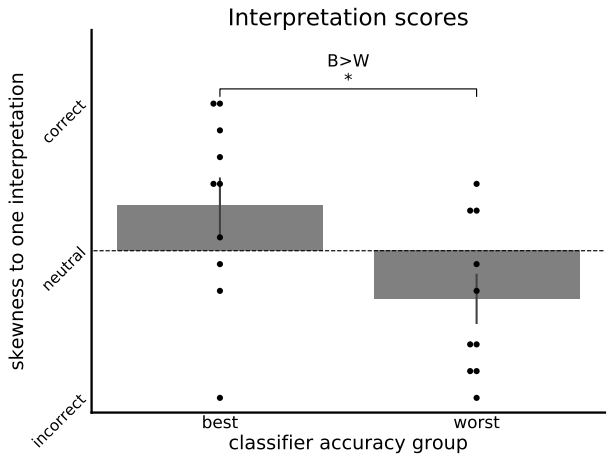


Figure 4.19: Interpretation scores split by classifier accuracy. Here, the interpretation score is adjusted for assigned group such that positive values indicate correct responses. Correct interpretation was significantly larger for the subjects with more accurate classification. **B** = best; **W** = worst. * = $p < 0.05$

Empathy ratings

Arthur minus Lee empathy difference did not vary significantly depending on classifier accuracy (one-tailed $t(18) = 0.77$, $p = 0.23$; Figure 4.20). The absence of an overall mean difference is consistent with the fact that we were able to successfully modify empathy scores across all subjects, so dividing the groups should not yield strong differences. Still, subjects with the best classification had a higher mean and smaller variance than those with the worst classification. The difference in distributions between the best and worst classifier groups may be informative for how effective our manipulation was in modifying empathy scores.

Probe responses

As shown in Figure 4.21, choosing the correct interpretation during probes was heavily influenced by the accuracy of the classifier. In the final neurofeedback run, subjects with the

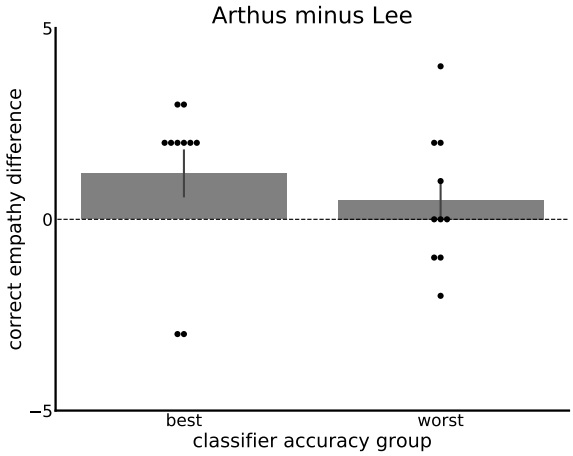


Figure 4.20: Empathy difference for Arthur - Lee, split by classifier accuracy. Here, the empathy difference is adjusted for assigned group such that positive empathy differences indicate the correct direction. There was no significant difference between the best and worst classifier groups in terms of modifying empathy.

best classification were significantly more likely than subjects with the worst classification to choose the correct assigned interpretation for 3 of the 7 stations (one-tailed t-test values were all below $p < 0.01$ significance, after Bonferroni correction for multiple comparisons).

4.3.3 Discussion

The modifications that we made in Experiment 2 allowed us to test if more sensitive neurofeedback would still be successful in changing story interpretations. Overall, the most robust difference between groups was found in empathy ratings, with the cheating group empathizing more with Arthur and paranoid group empathizing more with Lee. Only when we split the subjects by best versus worst classifier accuracy did we see behavioral differences in interpretation scores and probe responses.

At this time, it is unclear what caused the difference in classification accuracy across subjects. One possibility is that some subjects showed poor classification because they were un-

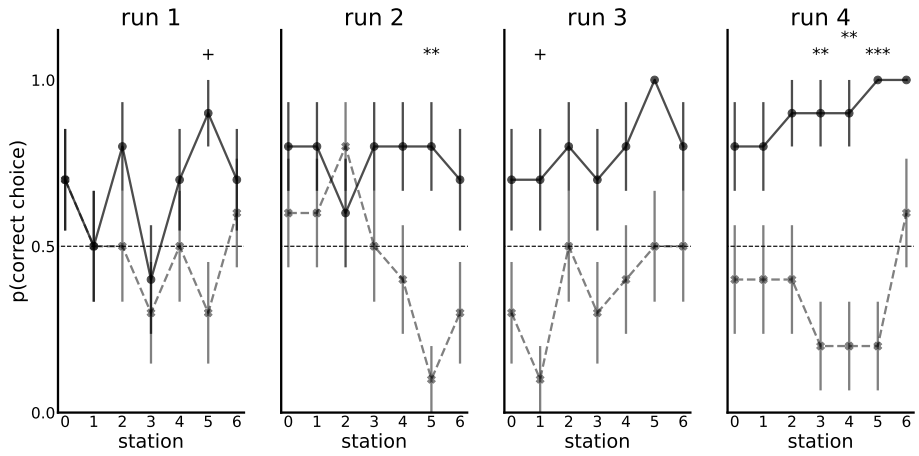


Figure 4.21: Probe choices split by classification accuracy (**solid lines** = best classification group; **dashed lines** = worst classification group). Subjects with the most accurate classification were more likely to choose probes consistent with their assigned interpretation by the end of training. Error bars represent ± 1 s.e.m. Significance was corrected for multiple comparisons using Bonferroni correction. ** = $p < 0.01$; * = $p < 0.05$; + = $p < 0.1$

able to clearly shift their story interpretations (e.g., due to differences in attention span, cognitive control capacity, working memory capacity, etc.). That is, the classifier might be accurately registering a failure to adopt different interpretations. If this is the case, perhaps screening subjects beforehand with various behavioral tasks could help exclude subjects who will ultimately not perform well.

Another possible explanation for poor classification (other than individual differences in cognitive ability) is that our analysis pipeline may be insufficiently sensitive to real differences in cognitive states. If poor classification was due to limitations of our pipeline (e.g., registration inaccuracy, ROI selection, classifier parameter choices), then our efforts would be best directed at improving the pipeline.

4.4 Discussion

With regards to changing final story interpretations according to randomly assigned groups, both Experiments 1 and 2 succeeded in their own way. Experiment 1 led to robust differences in how groups responded to interpretation-based questions at the end of scanning. However, we think that this result was due to feedback scores alone that were insensitive to small neural fluctuations. Since real-time subjects tended to be biased in the same way at the same points in the story, their feedback scores were determined mostly by group assignment instead of their own unique neural responses.

Experiment 2 was able to assess how accurately the classifier decoded interpretations in real-time. Behaviorally, our manipulation modified empathy ratings across groups for Lee and the Arthur − Lee difference. We believe we improved neurofeedback by increasing variability in neurofeedback scores within group. However, we still cannot conclude that individualized neurofeedback was necessary for these results without recruiting a separate yoked control group.

For the future yoked cohort, we cannot yoke to a single subject as is typical in other real-time neurofeedback experiments (e.g., in deBettencourt et al., 2015). Subjects will be choosing different interpretations throughout the story, so the same neurofeedback score will have a different meaning depending on their current state. To account for this difference, we will randomly sample yoked scores from those in the same group who have chosen the same response. For example, if a yoked subject is assigned to the paranoid group and chooses the cheating interpretation for run 1, station 3, we will randomly select a neurofeedback score from all paranoid subjects in Experiment 2 who also chose the cheating interpretation for

run 1, station 3. As a result, subjects will receive neurofeedback that they can interpret in the same way as the actual real-time subjects. Only upon running this control can we then conclude whether or not this rt-fMRI paradigm is able to alter thoughts through individualized neurofeedback.

A puzzling result is how some subjects in Experiment 2 showed accurate neural decoding based on self-reported interpretations, while others did not. Individual differences in responsiveness to neurofeedback has been noted in previous studies. It is estimated that up to 30% of subjects are 'non-responders', meaning they cannot regulate their neural activity with neurofeedback (Allison & Neuper, 2010; Sorger et al., 2019; Oblak et al., 2019). However, Oblak et al. (2019) warns that the reliability of the neurofeedback signal itself is a more pressing issue. With the given results from Experiments 1 and 2, we cannot rule out which one of these was to blame for poor results in some of the subjects.

On a broader level, this work presents the first attempt to use rt-fMRI to alter thoughts related to naturalistic stimuli. Naturalistic stimuli have the advantage of producing robust neural responses and, thus, generating more similar responses across people. This improves across-subject analyses, such as intersubject correlation (ISC; Hasson et al., 2004), and likely helped our effort to train the offline, between-subject classifier. This feature, however, may simultaneously be a limiting factor in our attempt to change individual signals. Naturalistic stimuli evoke large shared responses and may not be well-suited to find measurable and malleable individual differences. The consistent responses across people and runs in Experiment 1 showed how the story yielded robust neural signatures across people. More work must be done to find the balance between achieving high decoding accuracy across subjects and high sensitivity to individual differences.

As we continue to test different iterations of this task, our rt-cloud software system will facilitate this work by allowing us to perform identical real-time experiments across multiple locations. Through efficient data collection, we can quickly adjust different parameters and optimize our design for future clinical usage.

Acknowledgements

This work was supported by funding from the Intel Corporation to KAN and NTB and National Institutes of Health (T32MH065214) to ACM.

5

Conclusion and future directions

It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.

Albert Einstein

5.1 Lessons learned

In each of the previous chapters, we considered how rt-fMRI can be used to alter thoughts. First, we explored how to change negative attentional states common to depression with rt-fMRI neurofeedback. The review in Chapter 2 identified one attentional process — disengaging attention away from negative stimuli — that yielded robust group differences between depressed and healthy participants; shown through eye movements (Sanchez et al., 2013), key presses (Gotlib & Joormann, 2010), and neural responses (Sheline et al., 2009). Surveying previous literature allowed us to determine suitable approaches for our own rt-fMRI research.

In Chapter 3, we recruited our own cohort of depressed and healthy participants for an rt-fMRI study designed to train attention away from negative faces. Our results supported our hypothesis that disengagement is a key problem for depressed subjects. Compared to healthy control subjects, depressed subjects showed increased sustained negative attention, as measured through our neurofeedback task (and eye movements in a separate task) before training. Additionally, after neurofeedback, depressed subjects reduced their sustained negative attention, as well as their depression severity. Finally, the reduction of sustained negative attention during neurofeedback was positively related to the reduction in depression severity. This suggests that rt-fMRI training may have led to these improvements, though another control group is needed for further confirmation.

In Chapter 4, we explored the possibility of using rt-fMRI to alter story beliefs in two different experiments. In Experiment 1, we significantly changed story interpretations between our randomly assigned groups. However, we observed that neurofeedback within groups

was largely similar, and thus not sensitive to an individual's neural activity. We improved on our design in Experiment 2 by adjusting neurofeedback scores so they varied between participants. This time, we no longer managed to change story interpretations, though we significantly changed empathy across groups for Lee and Arthur – Lee. Furthermore, we *did* change story interpretations for the subjects with the most accurate neural classification. Thus, we believe that individualized neurofeedback shaped interpretations successfully when the signal was accurate, though a follow-up control group is needed for verification.

5.2 Future directions

While each of our rt-fMRI studies succeeded in changing thoughts — in the form of attention to negative faces or beliefs surrounding a story — the role of rt-fMRI in inducing these effects remains uncertain. To verify its role, we will include a control group in each of our studies.

Regarding our depression study outlined in Chapter 3, we will need to include a depressed control group that receives another kind of neurofeedback, either based on activity from an irrelevant region (e.g., auditory cortex), or a previous subject's brain activity. If the depressed subjects in this control group gain similar benefits as our original subjects, then we will conclude that individualized rt-fMRI was not critical in our study. However, if the new control group does not gain the same benefits, we can conclude with more confidence that the neurofeedback was effective by forcing subjects out of negative states, particularly at moments when they were stuck.

For our interpretation study described in Chapter 4, we will include a control group that receives neurofeedback based on a previous subject's neural response. We will always yoke to

subjects in the same assigned group. However, we will not yoke to a single previous subject, as probe responses can differ at each station. To account for this, we will randomly select neurofeedback scores from subjects in the same group who chose the same interpretation at a particular station. In this way, we can control for the different choices subjects made in real-time. If the yoked control group shows similar interpretation group effects as our experimental subjects in Experiment 2, we will conclude that the neurofeedback did not need to be individualized to a subject's neural response. However, if this yoked group does not show interpretation differences across assigned interpretations, we can more confidently claim that neurofeedback provided to subjects' neural response at a specific moment in the story led to interpretation differences.

Future iterations of both experiments will be facilitated by our rt-cloud software. We can efficiently reproduce our real-time processing pipelines for multiple users and institutions, as the software is largely independent of local resources.

5.3 Final conclusion

This thesis has explored the potential of using rt-fMRI neurofeedback to change thought trajectories over time. While the two projects are currently unable to prove the causal role of rt-fMRI without follow-up control groups, each experiment has yielded preliminary success in altering thoughts. The current results showcase the ability of using computational tools to decode complex cognitive states and deliver prompt neurofeedback. It is our hope that our projects contribute to the rt-fMRI community through the unique designs, scientific findings, and open-source rt-cloud processing pipelines.

A

Supplementary Materials: Chapter 3

A.1 Correlation with specific networks

Although we used a whole-brain ROI to train attention in real-time, we wanted to understand how the contributions of specific networks – perceptual versus attentional – affected subsequent learning. To do this, we replicated the methods from deBettencourt et al. (2015). Namely, we first downloaded ROIs for the attentional and perceptual networks from Neu-

rosynth (Yarkoni et al., 2011). Similar to deBettencourt et al. (2015), we used functional and anatomical definitions to create each network ROI. For functional masks, we queried Neurosynth for 'attention' and 'faces' or 'places' for the attentional and perceptual network, respectively (statistical threshold $z \geq 2.3$). We then used anatomical masks to ensure that the attentional network included only frontal and parietal lobes, and that the perceptual network only included the occipital and temporal lobes (using the respective MNI atlas labels in FSL).

Next, we retrained and tested the new classifiers using the same data that were recorded in real-time, but this time we masked the data first with each network ROI. Thus, the new neurofeedback scores represented what the neurofeedback would have been had that ROI been used in real-time. Finally, we correlated the original neurofeedback scores with each of the scores from the ROIs to understand the relative contribution that each network made in determining the neurofeedback score. To understand how network contributions impacted learning, we then correlated the relative contributions with behavioral and neural outputs.

A.1.1 Results

All of our results are summarized in Table A.1.

First, we focused on replicating the results from deBettencourt et al. (2015) as closely as possible. Thus, the two metrics we correlated were (1) average network correlation over *all neurofeedback runs*, and (2) average go/no go - behavior task $\Delta A'$ over *all possible conditions* (e.g., when ignoring sad faces, attending happy faces, etc.). No relationship between network correlation and behavioral improvement was found for either network, considering all subjects and MDD subjects only (first row in Table A.1; Figure A.1).

If *average* correlations between networks did not relate to behavioral learning, we wanted

Correlation	Performance measure	All: Perceptual	MDD: Perceptual	All: Attentional	MDD: Attentional
Average	$\Delta A'$	r = 0.32 p = 0.10	r = 0.14 p = 0.63	r = -0.23 p = 0.25	r = -0.49 p = 0.066
Δ	$\Delta A'$	r = 0.45 p = 0.019	r = 0.14 p = 0.63	r = 0.38 p = 0.048	r = 0.49 p = 0.063
Δ	reduction in sustained negative attention	r = -0.17 p = 0.42	r = 0.13 p = 0.66	r = -0.13 p = 0.53	r = 0.020 p = 0.95
Δ	increase in sustained correct attention	r = 0.10 p = 0.62	r = 0.10 p = 0.72	r = 0.38 p = 0.057	r = 0.60 p = 0.018

Table A.1: All Pearson r correlation results for network contributions. **All** = all subjects included; **MDD** = only MDD subjects included; **yellow** = $p < 0.05$

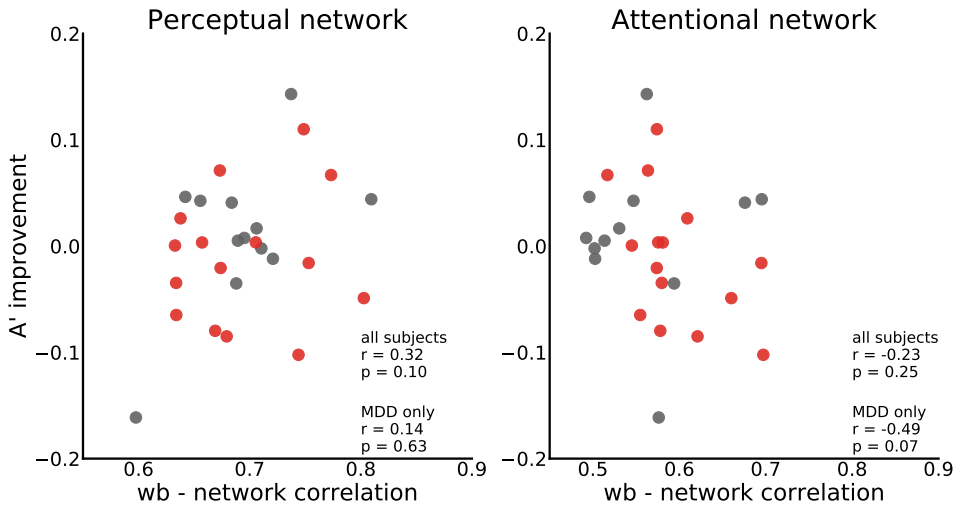


Figure A.1: Correlation results that sought to replicate that of deBettencourt et al. (2015) as closely as possible. No significant correlations were found from averaged network correlations over the course of the three days of neurofeedback. **wb** = whole-brain ROI. Circle color indicates group (**red** = MDD group; **black** = HC group).

to see if the *change* of network contributions did relate to improvements in behavior. Of note, the neurofeedback in deBettencourt et al. (2015) included one day of neurofeedback, so averaging over all runs may have been more fitting in that design. Because our study encompassed multiple days of training, pre- versus post-neurofeedback metrics may be more descriptive of learning, as opposed to averaging across all days. To do this analysis, we took the first 3 runs of the first neurofeedback session (early NF) and the last 3 runs of the last neurofeedback session (late NF), as we had done for our other classification analyses (see Methods, Section 3.2.7.2).

When including all subjects, the contributions of both networks were positively correlated with behavioral improvements. However, when considering only depressed subjects, only the attention network contribution changes were marginally positively correlated to behavioral changes (second row in Table A.1; Figure A.2).

Next, we assessed if change in network correlations were related to changes in sustained attention during neurofeedback, both for negative and correct attentional states. There were no significant relationships between changes in network contribution and changes in sustained negative attention (third row in Table A.1; Figure A.3).

However, improvements in sustained correct attention *were* related to increases in attentional, but not perceptual, network contribution (fourth row in Table A.1; Figure A.4). The positive relationship was marginal when considering all subjects and significant when considering MDD subjects only. Thus, this relationship seems to be driven by the depressed subjects.

The difference in results between sustained negative attention and sustained correct attention suggests that there may be different mechanisms that produce improvements within

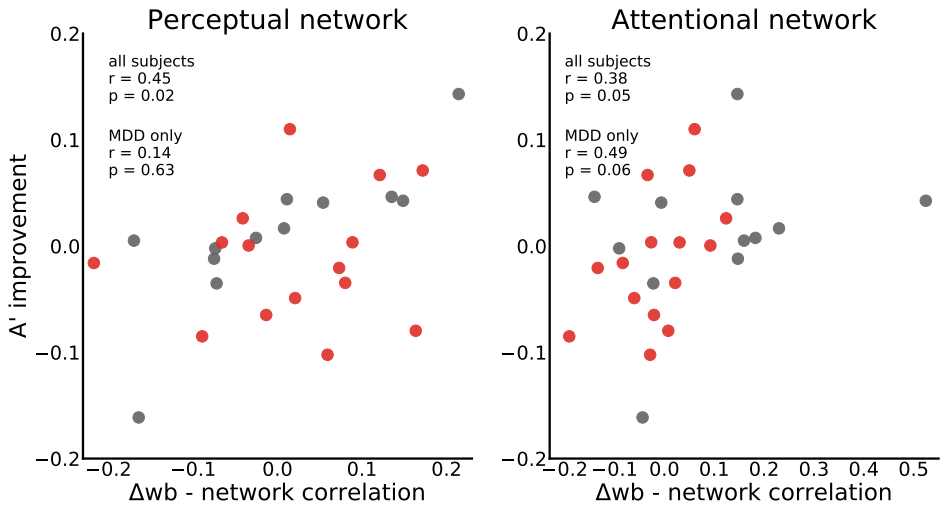


Figure A.2: Behavioral changes as a function of change in network correlations. It appears that the HC group is driving the correlation between increased perceptual network contribution and A′ improvement, while the MDD group is driving the correlation between increased attentional network contribution and A′ improvement. **wb** = whole-brain ROI. Circle color indicates group (**red** = MDD group; **black** = HC group).

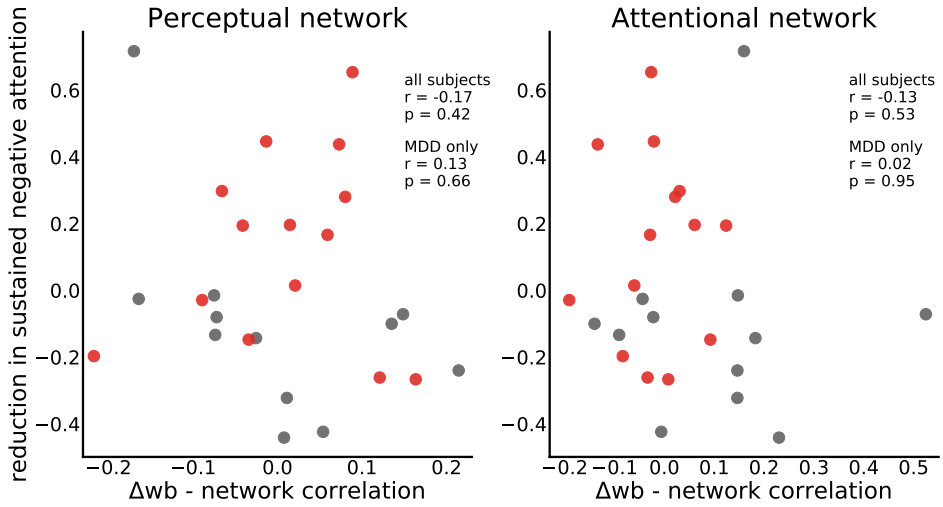


Figure A.3: Reduction of sustained negative attention as a function of change in network correlations. We did not see any significant relationships. **wb** = whole-brain ROI. Circle color indicates group (**red** = MDD group; **black** = HC group).

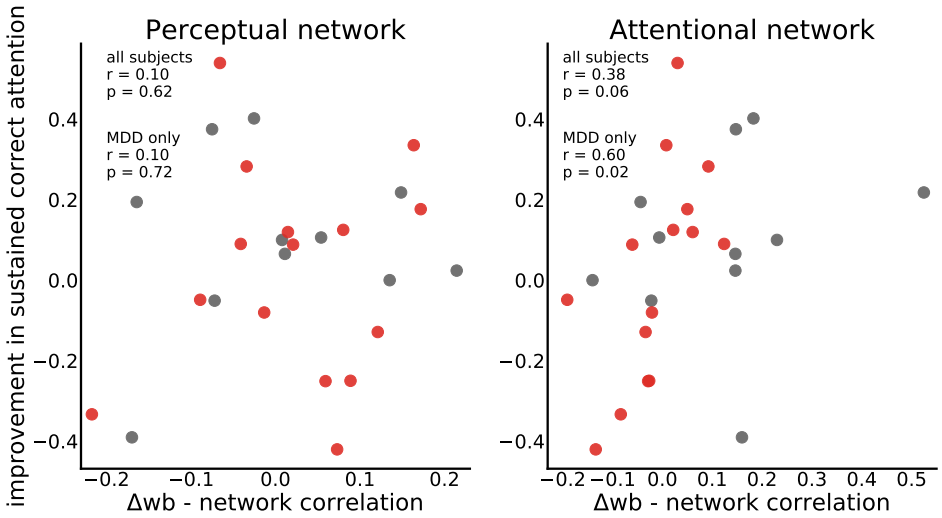


Figure A.4: Improvement in correct sustained attention as a function of change in network correlations. Here, the increase in attention network contribution from early to late NF was positively correlated with the improvement in sustained correct attention. **wb** = whole-brain ROI. Circle color indicates group (**red** = MDD group; **black** = HC group).

each domain. Reductions in sustained negative attention involve inhibiting attention to negative stimuli, while increases in sustained correct attention involve properly maintaining attention on neutral stimuli. Thus, the attentional network may be key to ensure improvements in sustained correct attention, while reductions in sustained negative attention may rely on the other neural markers (e.g., network activity of other regions in the DMN, as in Sheline et al., 2009), activity within localized regions (e.g., amygdala, as in Fales et al., 2008), or changes in connectivity between regions (e.g., amygdala to frontal connectivity, as in Satterthwaite et al., 2016).

B

Supplementary Materials: Chapter 4

B.1 Designing the classifier

Before being able to provide neurofeedback based on story interpretations, we had to construct an accurate classifier based on previously collected data. In order to decide on decoding methods, we optimized neural decoding in the following ways: (1) how we preprocessed the data entered into classifier training, and (2) which time points of the story would be analyzed

in real-time. We outline the process for making each decision in the following sections.

B.1.1 Neural data

We used previously-collected data from Yeshurun et al. (2017), where subjects were explicitly instructed to adopt one of two different interpretations before listening to an ambiguous story in the fMRI scanner. This manipulation ensured that the two groups of subjects would interpret the story in different ways, allowing the authors to look for neural signatures of interpretation that were shared within each group. The data set included 38 subjects: 19 subjects who were told that Joanie was cheating on Arthur, and 19 subjects who were told that Arthur was paranoid. *Note:* The data published in Yeshurun et al. (2017) included 20 subjects per group, but 2 of the subjects had variable start times, so we only included the data from subjects that were perfectly synced to minimize the use of processing techniques that would not be available in real-time. For more details regarding this data set, see Yeshurun et al. (2017). We preprocessed the neural data from the remaining 38 subjects in the same way as detailed above in Chapter 4 (see Methods, Section 4.2.1.5), using *fMRIPprep* 1.2.3 (Esteban et al., 2018b; Esteban et al., 2018a; RRID:SCR_016216).

B.1.2 Classification approach

We designed our offline analysis method based on the insights provided by Chen et al. (2015). The authors analyzed the same data in Yeshurun et al. (2017), but this time used a novel shared response model (SRM) approach in two different steps to improve classification accuracy. Usually, SRM is used when subjects listen or view naturalistic stimuli (e.g., movies, audio) in order to extract the common signal to best represent the stimulus-dependent evoked response

(Chen et al., 2015; Vodrahalli et al., 2016).

As described in Chen et al. (2015), a subject's time series is represented as $X_i \in \mathbb{R}^{v \times d}$, where v is the number of voxels within the ROI and d is the number of TRs in the time series, for subjects $i = 1 : m$.

Using SRM, each subject's time series is broken down as a linear combination of shared and residual components, or

$$X_i = W_i S + E_i \tag{B.1}$$

In the above notation, $S \in \mathbb{R}^{k \times d}$ is the time series of activity in shared space, and is the same for all subjects. $W_i \in \mathbb{R}^{v \times k}$ is the subject-specific feature matrix that transforms voxels into shared k-dimensional space. Both terms are found through optimizing the model such that each subject's residual term, E_i, is minimized. The reduced dimension, k, is chosen by the experimenter before training the model.

Thus, one can determine the *residual response*, which represents responses unique to the individual via subtraction of the original and shared responses:

$$E_i = X_i - W_i S \tag{B.2}$$

While the same principles of SRM outlined above are relevant for our use, there is a critical difference. Because the data set in Yeshurun et al. (2017) included two groups who listened to the same narrative, we were not interested in the signal that was shared amongst *all* subjects. Rather, we wanted to extract the signal that was common *within each group*, as the shared signals within each of the two groups would likely reflect interpretations. For this reason,

Chen et al. (2015) used SRM in two stages.

In the first SRM step, the authors fit one SRM, S_1, in k_1 dimensions and included all subjects. Then, the authors kept only the residual signal, E_1, for each subject. This had the effect of removing any signal that was common to the two interpretation groups (and thus was not informative about interpretation). Then, the authors removed 1 subject per group for testing. They divided the remaining subjects by group for the next SRM. In the second SRM step, each group was fit to its own SRM in k_2 dimensions (e.g., one paranoid group model, S_{2p}, and one cheating group model, S_{2c}). This time, the authors kept only the shared responses, $W_2S_2 \in \mathbb{R}^{v \times d}$, and removed the residual term. The shared component of each subject's signal served as classifier training data. This classifier then predicted the group label of the 2 held-out subjects.

Using this two-stage SRM procedure, average prediction accuracy increased from 72% (no SRM) to 82% with the best-fitting dimensions set for each SRM, k_1, and k_2, respectively. See Chen et al. (2015) for additional mathematical details on SRM analysis and BrainIAK (https://brainiak.org/) for details on implementation. We adapted this approach for our offline analysis, though we modified it slightly for real-time application.

B.1.3 Decision 1: how to preprocess the data

To determine which parameters resulted in the most accurate classification, we ran a bootstrap analysis with the offline data set. The parameters tested were: ROI, choice of high-pass filter, dimensions for SRM (k_1 and k_2), and whether or not to remove the average response over all subjects. For each parameter value considered, we ran a bootstrap analysis 1,000 times. In each bootstrap iteration, we randomly selected 2 subjects (1 per group) as

testing data, and 36 subjects (18 per group) as training data. We kept the number of subjects in each group even by sampling within group. Details regarding the order of these steps are shown in Figure B.1. Specific options tested for each parameter are listed next.

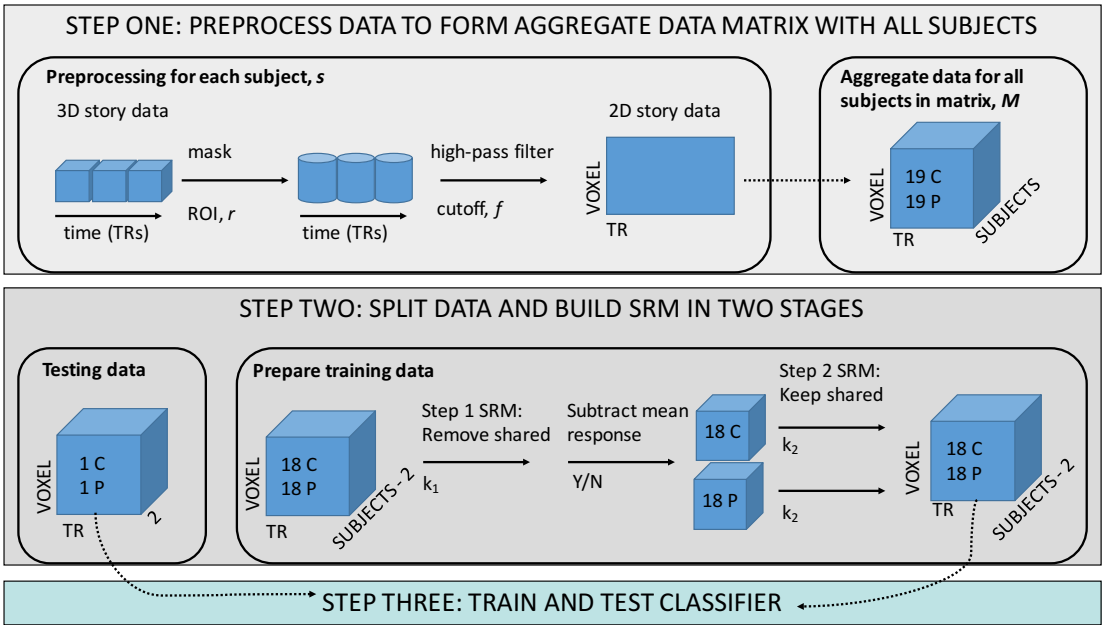


Figure B.1: Steps taken for each of our bootstrap iterations. Before splitting the data into training and testing data, we completed certain preprocessing steps for all subjects: we masked the story time series given a specific ROI, we high-pass filtered the data based on a particular cutoff (or no cutoff) frequency, and we flattened this time series into a 2D matrix. Next, we randomly selected 2 subjects (1 per group) to be left out for testing. For the remaining training data, we randomly selected 18 subjects per group (with replacement) so that groups were still balanced. Then, we trained an SRM on all training data using k_1 dimensions and removed the shared signal. Next, we subtracted the mean response (if this step was included). We then trained separate SRMs using k_2 dimensions and kept only the shared signal. Finally, we trained the classifier on the resulting residuals (in voxel space). We then used the classifier to predict group labels for the 2 held-out subjects.

B.1.3.1 Regions of interest

We considered three ROIs for this study. The list of ROIs and the steps taken to create each ROI is shown in Table B.1.

Index	Name	Method	N voxels
0	Default Mode Network (DMN)	Yeo et al. (2011): - resampled to functional space	3,757 voxels
1	Theory of Mind (TOM)	Neurosynth (Yarkoni et al., 2011): - used the association test only - thresholded at FDR-corrected $p \leq 0.01$ significance - resampled to functional space - masked by average brain mask across all subjects	2,414 voxels
2	TOM clusters	Yeshurun et al. (2017), Supplementary section (Table 1): - Extracted significant clusters: left TPJ, precuneus, posterior cingulate, VMPFC, DMPFC, rIFG, lIFG, lSTS, and left temporal pole - converted to MNI space - created 10-mm cube around each cluster - resampled to functional space - masked by average brain mask across all subjects	240 voxels

Table B.1: All ROIs considered in our bootstrap analysis, with a description of how we created each mask. The index on the left corresponds to the ROI key shown in Figure B.2.

B.1.3.2 High-pass filtering

Because the story lasted about 12 minutes, we had no a priori knowledge of what filtering cutoff would be optimal in removing noise. Thus, we tested 3 different frequency options for high-pass filtering: (1) no filtering, (2) high-pass filtering with a cutoff of 337.75 s as was done in Yeshurun et al. (2017), and (3) high-pass filtering with a cutoff of 140 s as is commonly done in fMRI studies. These options are represented in Figure B.2 as indices 0, 1, and 2, respectively.

B.1.3.3 Dividing the data into training and testing data

Once the data were preprocessed (first 2 TRs removed, high-pass filtered, masked), we z-scored each voxel's time series. Then we randomly selected 1 subject per group to remove from the subsequent processing steps and hold out for testing. The purpose of this step was to mimic real-time data collection as accurately as possible. To build the training data matrix, we randomly selected 18 subjects per group (with replacement). This way, the number of subjects per group in our data set would remain balanced.

B.1.3.4 Step 1 SRM

As described above, we executed an initial step of SRM and removed the component of the signal that was shared amongst **all training subjects**. The purpose of this step was to remove signal that captured low level processing that would be common across all subjects. We tested the benefits of this step by modifying the number of dimensions in the SRM, k_1. Note: we additionally tested not performing this step at all with $k_1 = 0$.

B.1.3.5 Subtracting the group signal

Next, we averaged each voxel's time series across all subjects (in the training data set only). We then subtracted this mean signal from each subject's voxel activity. This step was done in Yeshurun et al. (2017), as it accomplishes the same goal as the first SRM step by removing shared signal across groups. We tested the benefit of this step by varying whether or not it was included.

B.1.3.6 Step 2 SRM

Next, we divided the training data by interpretation group. In each group, we trained a separate SRM with k_2 dimensions. This time, we kept only the shared component for each subject, as the signal likely related to story interpretation. We then trained a classifier on this cleaner signal.

B.1.3.7 Training and testing

After all preprocessing steps were completed and we had our final training data matrix, we then tested the model on the 2 subjects that were held out.

We created two different types of classifiers:

1. TR classifier: We trained a separate classifier for each TR in the story. This allowed us to find which TRs were the most accurate in predicting group label.

2. Spatiotemporal classifier: We flattened each voxel x time matrix into a 1-D vector as done in Chen et al. (2015). This allowed us to test how spatial and temporal information together improved classification.

We used Scikit-learn's (Pedregosa et al., 2011) SVM classifier (kernel='linear', probability=True) to predict the group label of the 2 left-out subjects. By comparing classification accuracy for both classifier types, we learned how each classifier performed and the best moments to classify the signal.

B.1.3.8 Choosing the preprocessing method

The results of our bootstrap analysis is shown in Figure B.2. The results are ranked by the the mean accuracy across the TR and spatiotemporal classifiers over all bootstrap iterations. Thus, the classifier shown in the left-most or 0 index is the classifier with the overall highest accuracy. This classifier corresponded to the following preprocessing steps: ROI = large TOM mask, 337.73 s cutoff high-pass filter, average signal removed, $k_1 = 0$ (no step 1 SRM), and $k_2 = 25$.

We originally chose this classifier, but later decided to remove the high-pass filtering step. Upon further analysis, we noticed a spike in noise at the high-pass filter cutoff when we simulated real-time processing, as opposed to processing data only at the end of the run. We kept all other parameters the same. Thus, our final processing parameters were: ROI = large TOM mask, no high-pass filter, average signal removed, $k_1 = 0$ (no step 1 SRM), and $k_2 = 25$. This parameter combination yielded the highest TR classifier accuracy. The performance of this classifier is shown in the red box in Figure B.2. Compared to all other classifiers that did not include filtering, this classifier had the second highest overall accuracy.

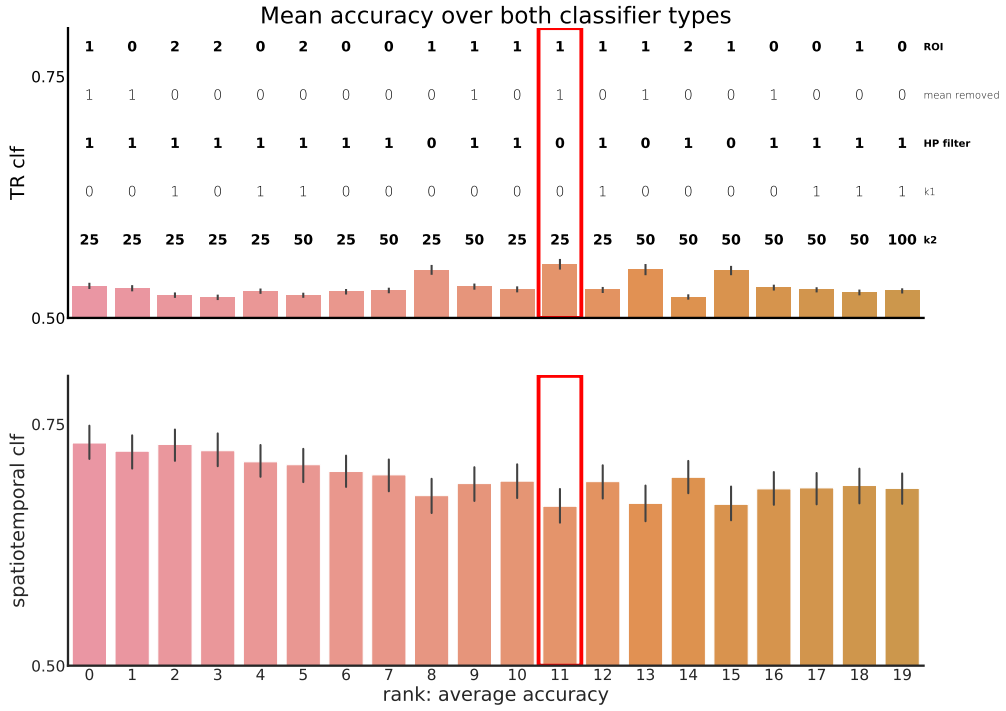


Figure B.2: Results from the classifier search. Shown are the top 20 performing parameters for classification ranked by the **average** accuracy across TR and spatiotemporal classifier performance. Best to worst is ordered from left to right. The red box marks which parameter set was ultimately chosen: ROI = large TOM mask, average signal removed, no high-pass filter, k_1 = 0, k_2 = 25. While we initially chose the top ranked classifier, we noticed high-pass filtering induced noise at the cutoff frequency. Amongst the other classifiers that didn't include high-pass filtering (third row = 0), we chose the second-highest performing classifier on average, and the top performing TR classifier out of all options. Error bars show 95% CI over bootstrap iterations.

B.1.4 Decision 2: which time points in the story to use for neurofeedback

To decide when to provide neurofeedback, we also used behavioral ratings from Yeshurun et al. (2017) to identify maximally different points in the story depending on interpretation. Each of the story's 179 segments were rated in terms of how much they differed in beliefs, emotions, and intentions depending on the context. Yeshurun et al. (2017) found that the segments that were rated as the most different were associated with more accurate group classification. This result is a good sanity check, as some parts of the story had nothing to do with the cheating versus paranoid interpretations (such as when the narrator described a chair).

To combine all the ratings into one score, we first z-scored belief, emotion, and intention ratings across raters. Then, we calculated *segment scores* by taking the mean of all ratings at each segment. Thus, larger segment scores meant that those time points were rated as differing more in the story. In Figure B.3, the segment scores are plotted in blue and the average individual TR accuracy (from our chosen classifier) is plotted in black. The Pearson correlation between the TR accuracy and segment scores was r = 0.21, p = 71E-6, meaning that the more a segment was rated as differing between contexts, the better the TR classifier performed. As mentioned above, this relationship was reported in Yeshurun et al. (2017), though the analysis methods differed.

To determine the best time points for neurofeedback, we required that included TRs must have: (1) individual TR classifier accuracy ≥ 0.55, (2) segment score ≥ 0, and (3) the first 2 conditions satisfied for at least 2 consecutive TRs. Once these TRs were extracted, individual spatiotemporal classifiers, or *stations*, were trained and tested using the same bootstrap analysis with the chosen preprocessing parameters. Out of the stations we tested, we chose final stations with the highest accuracy and even distribution throughout the story. Accuracy re-

sults for the chosen stations are shown in red in Figure B.3. See Figure B.4 for a comparison of classifier accuracy across TR, spatiotemporal, and station classifiers.

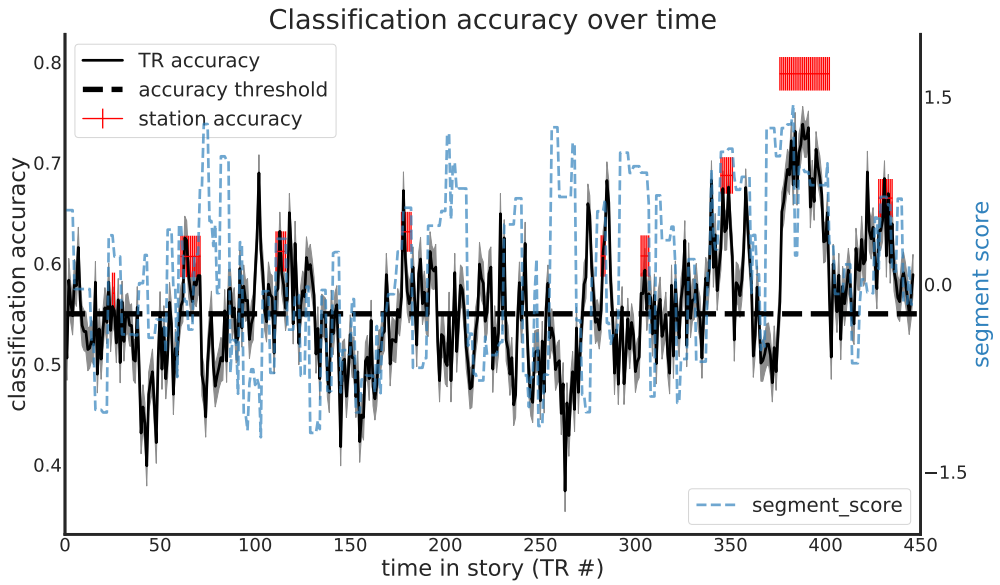


Figure B.3: TR and spatiotemporal station accuracy shown over time. The x-axis shows the TR number, while the left y-axis shows the classification accuracy. The **solid black line** represents the individual TR classifier accuracy, with the shaded gray 95% CI over all bootstrap iterations. The **dashed black line** indicates the minimum TR accuracy needed to include a given TR in a station. The **red lines** show chosen stations, with their height indicating classification accuracy, with 95% CI over all bootstrap iterations. The **blue line** shows the segment score ratings (right y-axis). The positive correlation between the TR classification accuracy and segment scores implies that points in the story that were rated to differ in interpretation generally yielded more accurate classification.

B.2 Instructions for Experiment 2

Next, we include the instructions that were used in Experiment 2 in Figure B.5.

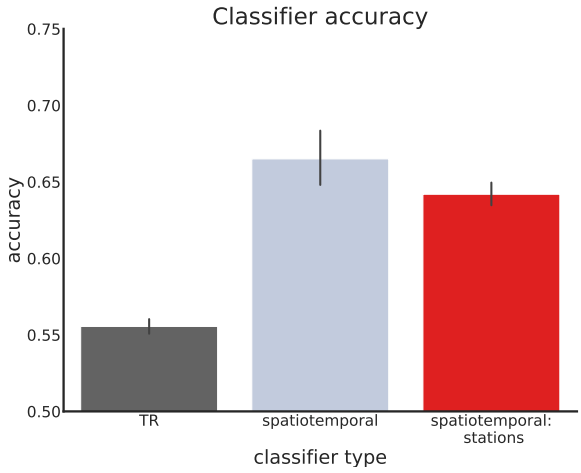
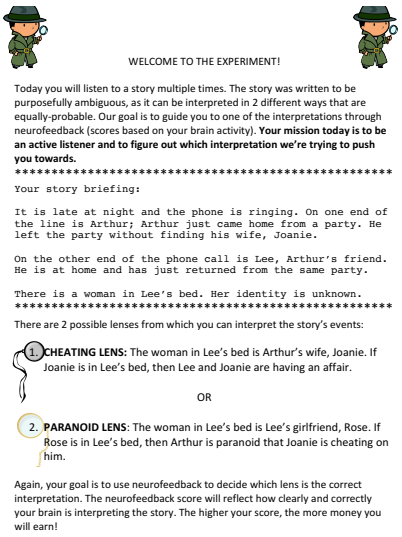


Figure B.4: Classifier results from the winning parameters: ROI = large TOM mask, average signal removed, no high-pass filtering, $k_1 = 0$, $k_2 = 25$. The different bars represent different classifier types. In black, the TR classifier accuracy was the least accurate, as it is averaged over all TRs in the story, which included uninformative TRs. The spatiotemporal classifier over the entire story was the most accurate, though isolating specific spatiotemporal stations still led to accurate, and less variable, results. Error bars represent 95% CI over all bootstrap iterations.



WELCOME TO THE EXPERIMENT!

Today you will listen to a story multiple times. The story was written to be purposefully ambiguous, as it can be interpreted in 2 different ways that are equally-probable. Our goal is to guide you to one of the interpretations through neurofeedback (scores based on your brain activity). **Your mission today is to be an active listener and to figure out which interpretation we're trying to push you towards.**

****

Your story briefing:

It is late at night and the phone is ringing. On one end of the line is Arthur; Arthur just came home from a party. He left the party without finding his wife, Joanie.

On the other end of the phone call is Lee, Arthur's friend. He is at home and has just returned from the same party.

There is a woman in Lee's bed. Her identity is unknown.

****

There are 2 possible lenses from which you can interpret the story's events:

1. **CHEATING LENS:** The woman in Lee's bed is Arthur's wife, Joanie. If Joanie is in Lee's bed, then Lee and Joanie are having an affair.

OR

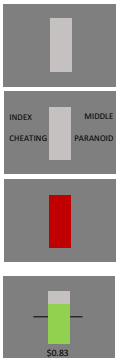
2. **PARANOID LENS**: The woman in Lee's bed is Lee's girlfriend, Rose. If Rose is in Lee's bed, then Arthur is paranoid that Joanie is cheating on him.

Again, your goal is to use neurofeedback to decide which lens is the correct interpretation. The neurofeedback score will reflect how clearly and correctly your brain is interpreting the story. The higher your score, the more money you will earn!

(a)

Task summary: As you listen to the story, you will pass through *recording stations* when we will be analyzing your brain. **Before** you pass through each recording station, you will be prompted to decide on an interpretation lens (cheating or paranoid) for that station. **After** the recording station ends, you will get a neurofeedback score based on *your brain activity during the recording station*.

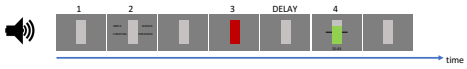
As the story plays, you will see the following screens:



1. **Listening only:** Keep listening and paying close attention to the story.
2. **Interpretation probe:** Decide on an interpretation, (cheating or paranoid), for the upcoming recording station. **This means you will continue listening to the story through this lens as you pass through the next station.** Press your index or middle finger to choose one. (The keypress order may be flipped from what's on the left.)
3. **Recording station:** The rectangle will turn red for recording. This means that we will be providing you with neurofeedback based on your brain activity from this time period. Adopt your chosen interpretation as strongly as possible as you continue to listen to the story. **Remember, you should be listening to the story through your chosen lens as if it were the correct and only interpretation of the story.**
4. **Neurofeedback:** You will receive a score based on your brain activity during the previous recording station. Higher scores indicate that you're clearly interpreting the story correctly. Lower scores can indicate either (1) noisy brain signals (like if you weren't super focused on the story) and/or (2) incorrect interpretations. If you score above the center line, you will see a green bar denoting your score with a monetary reward at the bottom. If you score below the center line, you will see a grey bar with $0. The largest reward possible is $1.

(b)

Putting everything together, the stations will look something like this:



You'll encounter multiple recording stations each time as you listen to the story. Thus, you'll have the opportunity to continue adjusting your interpretation to maximize your reward.

Here are some tips you can use to increase your reward:

- Your brain signal in the scanner takes about 6 s to be detected due to the physics of MRI. As shown above, **we will give you feedback about 6 seconds after the station ends based on your brain activity from the previous station only**.
- Science is never perfect, so your scores may be noisy and not always 100% accurate. Thus, just because you get a high score or low score for a particular station, **do not assume that you're definitely correct or definitely incorrect and stop using neurofeedback**. Keep using each score as a clue in your hunt to figure out the correct interpretation.
- We will give you a reward after each run. We will calculate the average bonus per run by averaging your reward per station (from $0 - $1.00) and multiplying that by 5. You will listen to the story 4 times, so you can earn up to $5 in each run, or **an extra $20 total today**.
- Even if you think that the story has directly revealed which interpretation is correct, characters might be lying. **Your job is to try to interpret the story according to what the neurofeedback is telling you, not to try to figure out which interpretation was intended by the author.**
- Even though this task has multiple parts, ultimately please listen to the story the whole time and try your best.

(c)

Lastly, to make your time worthwhile, we need to ensure good quality data. Please keep these pointers in mind:

- It is really important that you get comfortable beforehand so that you will be able to stay focused. Make sure before we start that you have gone to the bathroom and gotten comfortable in the scanner.
- Take rests between each run if you feel yourself getting sleepy or losing focus.
- Because we're doing real-time neurofeedback, it is extra important that you **try your best to avoid moving any part of your body, but especially your head, as much as possible**. We try to correct for motion afterwards, but the neurofeedback needs good data in the scanner too! To do this, make sure you're not crossing your arms or legs and again get comfortable beforehand.
- The ear buds can be finicky and may cut out if you move your head or jam the foam in too far into your ear. So please let me know asap if you can't hear well and try not to move particularly when the audio is playing!

To summarize, stay focused and keep listening throughout the story to succeed on your mission! ☻

Do you have any questions? If so, please ask!!!



(d)

Figure B.5: Instructions that span pages 1-4 shown in **a**, **b**, **c**, **d**, respectively.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8.

Abreu, R., Leal, A., & Figueiredo, P. (2018). EEG-Informed fMRI: A review of data analysis methods. *Front. Hum. Neurosci.*, 12, 29.

Allison, B. Z. & Neuper, C. (2010). Could anyone use a BCI? In D. S. Tan & A. Nijholt (Eds.), *Brain-Computer Interfaces: Applying our Minds to Human-Computer Interaction* (pp. 35–54). London: Springer London.

Armony, J. L. & Dolan, R. J. (2002). Modulation of spatial attention by fear-conditioned stimuli: an event-related fMRI study. *Neuropsychologia*, 40(7), 817–826.

Armstrong, T. & Olatunji, B. O. (2012). Eye tracking of attention in the affective disorders: a meta-analytic review and synthesis. *Clin. Psychol. Rev.*, 32(8), 704–723.

Avants, B., Epstein, C., Grossman, M., & Gee, J. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1), 26–41.

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. *J. Abnorm. Psychol.*, 124(3), 463–475.

Blamire, A. M., Ogawa, S., Ugurbil, K., Rothman, D., McCarthy, G., Ellermann, J. M., Hyder, F., Rattner, Z., & Shulman, R. G. (1992). Dynamic mapping of the human visual cortex by high-speed magnetic resonance imaging. *Proc. Natl. Acad. Sci. U. S. A.*, 89(22), 11069–11073.

Brett, M., Markiewicz, C. J., Hanke, M., Côté, M.-A., Cipollini, B., McCarthy, P., Cheng, C. P., Halchenko, Y. O., Cottaar, M., Ghosh, S., Larson, E., Wassermann, D., Gerhard, S.,

Lee, G. R., Wang, H.-T., Kastman, E., Rokem, A., Madison, C., Morency, F. C., Moloney, B., Goncalves, M., Riddell, C., Burns, C., Millman, J., Gramfort, A., Leppäkangas, J., Markello, R., van den Bosch, J. J., Vincent, R. D., Braun, H., Subramaniam, K., Jarecka, D., Gorgolewski, K. J., Raamana, P. R., Nichols, B. N., Baker, E. M., Hayashi, S., Pinsard, B., Haselgrove, C., Hymers, M., Esteban, O., Koudoro, S., Oosterhof, N. N., Amirbekian, B., Nimmo-Smith, I., Nguyen, L., Reddigari, S., St-Jean, S., Panfilov, E., Garyfallidis, E., Varoquaux, G., Kaczmarzyk, J., Legarreta, J. H., Hahn, K. S., Hinds, O. P., Fauber, B., Poline, J.-B., Stutters, J., Jordan, K., Cieslak, M., Moreno, M. E., Haenel, V., Schwartz, Y., Thirion, B., Papadopoulos Orfanos, D., Pérez-García, F., Solovey, I., Gonzalez, I., Palasubramaniam, J., Lecher, J., Leinweber, K., Raktivan, K., Fischer, P., Gervais, P., Gadde, S., Ballinger, T., Roos, T., Reddam, V. R., & freec84 (2019). nipy/nibabel: 3.0.0.

Browning, M., Holmes, E. A., Charles, M., Cowen, P. J., & Harmer, C. J. (2012). Using attentional bias modification as a cognitive vaccine against depression. *Biological Psychiatry*, 72, 572–579.

Calkins, A. W., McMorran, K. E., Siegle, G. J., & Otto, M. W. (2015). The effects of computerized cognitive control training on community adults with depressed mood. *Behav. Cogn. Psychother.*, 43(5), 578–589.

Caseras, X., Garner, M., Bradley, B. P., & Mogg, K. (2007). Biases in visual orienting to negative and positive scenes in dysphoria: An eye movement study. *J. Abnorm. Psychol.*, 116(3), 491–497.

Chai, X. J., Hirshfeld-Becker, D., Biederman, J., Uchida, M., Doehrmann, O., Leonard, J. A., Salvatore, J., Kenworthy, T., Brown, A., Kagan, E., de Los Angeles, C., Whitfield-Gabrieli, S., & Gabrieli, J. D. E. (2015). Functional and structural brain correlates of risk for major depression in children with familial depression. *Neuroimage Clin*, 8, 398–407.

Chechko, N., Augustin, M., Zvyagintsev, M., Schneider, F., Habel, U., & Kellermann, T. (2013). Brain circuitries involved in emotional interference task in major depression disorder. *J. Affect. Disord.*, 149(1-3), 136–145.

Chen, P.-H. c., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., & Ramadge, P. J. (2015). A Reduced-Dimension fMRI shared response model. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28* (pp. 460–468). Curran Associates, Inc.

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annu. Rev. Psychol.*, 62, 73–101.

Cisler, J. M. & Koster, E. H. W. (2010). Mechanisms of attentional biases towards threat in anxiety disorders: An integrative review. *Clin. Psychol. Rev.*, 30(2), 203–216.

Cohen, J. D., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., Norman, K. A., Pillow, J., Ramadge, P. J., Turk-Browne, N. B., & Willke, T. L. (2017). Computational approaches to fMRI analysis. *Nat. Neurosci.*, 20(3), 304–313.

Cohen, N., Margulies, D. S., Ashkenazi, S., Schaefer, A., Taubert, M., Henik, A., Villringer, A., & Okon-Singer, H. (2016). Using executive control training to suppress amygdala reactivity to aversive information. *Neuroimage*, 125, 1022–1031.

Cooper, J. A., Gorlick, M. A., Denny, T., Worthy, D. A., Beevers, C. G., & Maddox, W. T. (2014). Training attention improves decision making in individuals with elevated self-reported depressive symptoms. *Cogn. Affect. Behav. Neurosci.*, 14(2), 729–741.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.*, 29(3), 162–173.

Cox, R. W. & Hyde, J. S. (1997). Software tools for analysis and visualization of fmri data. *NMR in Biomedicine*, 10(4-5), 171–178.

Cox, R. W., Jesmanowicz, A., & Hyde, J. S. (1995). Real-time functional magnetic resonance imaging. *Magn. Reson. Med.*, 33(2), 230–236.

Cristea, I. A., Kok, R. N., & Cuijpers, P. (2015). Efficacy of cognitive bias modification interventions in anxiety and depression: meta-analysis. *Br. J. Psychiatry*, 206(1), 7–16.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2), 179–194.

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. *Nat. Neurosci.*, 18(3), 1–9.

deBettencourt, M. T., Keene, P. A., Awh, E., & Vogel, E. K. (2019a). Real-time triggering reveals concurrent lapses of attention and working memory. *Nat Hum Behav*, 3(8), 808–816.

deBettencourt, M. T., Turk-Browne, N. B., & Norman, K. A. (2019b). Neurofeedback helps to reveal a relationship between context reinstatement and memory retrieval. *Neuroimage*, 200, 292–301.

Detre, G. J., Natarajan, A., Gershman, S. J., & Norman, K. A. (2013). Moderate levels of activation lead to forgetting in the think/no-think paradigm. *Neuropsychologia*, 51, 2371–2388.

Disner, S. G., Beevers, C. G., Haigh, E. A. P., & Beck, A. T. (2011). Neural mechanisms of the cognitive model of depression. *Nat. Rev. Neurosci.*, 12(8), 467–477.

Duque, A. & Vazquez, C. (2018). A failure to show the efficacy of a dot-probe attentional training in dysphoria: Evidence from an eye-tracking study. *J. Clin. Psychol.*

Einstein, A., Calaprice, A., & Dyson, F. (2010). *The Ultimate Quotable Einstein*. Princeton University Press.

Eizenman, M., Yu, L. H., Grupp, L., Eizenman, E., Ellenbogen, M., Gemar, M., & Levitan, R. D. (2003). A naturalistic visual scanning approach to assess selective attention in major depressive disorder. *Psychiatry Res.*, 118(2), 117–128.

Erickson, K., Drevets, W. C., Clark, L., Cannon, D. M., Bain, E. E., Zarate, Jr, C. A., Charney, D. S., & Sahakian, B. J. (2005). Mood-congruent bias in affective go/no-go performance of unmedicated patients with major depressive disorder. *Am. J. Psychiatry*, 162(11), 2171–2173.

Esteban, O., Blair, R., Markiewicz, C. J., Berleant, S. L., Moodie, C., Ma, F., Isik, A. I., Erramuzpe, A., Kent, James D. andGoncalves, M., DuPre, E., Sitek, K. R., Gomez, D. E. P., Lurie, D. J., Ye, Z., Poldrack, R. A., & Gorgolewski, K. J. (2018a). Fmriprep 1.2.3. *Software*.

Esteban, O., Markiewicz, C., Blair, R. W., Moodie, C., Isik, A. I., Erramuzpe Aliaga, A., Kent, J., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S., Wright, J., Durnez, J., Poldrack, R., & Gorgolewski, K. J. (2018b). Fmriprep: a robust preprocessing pipeline for functional MRI. *bioRxiv*.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods*, 16(1), 111–116.

Everaert, J., Bronstein, M. V., Cannon, T. D., & Joormann, J. (2018). Looking through tinted glasses: Depression and social anxiety are related to both interpretation biases and inflexible negative interpretations. *Clin. Psychol. Sci.*, 6(4), 517–528.

Everaert, J., Podina, I. R., & Koster, E. H. W. (2017). A comprehensive meta-analysis of interpretation biases in depression. *Clin. Psychol. Rev.*, 58, 33–48.

Fales, C. L., Barch, D. M., Rundle, M. M., Mintun, M. A., Snyder, A. Z., Cohen, J. D., Mathews, J., & Sheline, Y. I. (2008). Altered emotional interference processing in affective and cognitive-control brain circuitry in major depression. *Biol. Psychiatry*, 63(4), 377–384.

Ferrari, G. R. A., Möbius, M., van Opdorp, A., Becker, E. S., & Rinck, M. (2016). Can't look away: An Eye-Tracking based attentional disengagement training for depression. *Cognit. Ther. Res.*, 40(5), 672–686.

First, M. B., Williams, J. B. W., Karg, R. S., & Spitzer, R. L. (2015). *Structured Clinical Interview for DSM-5—Research Version (SCID-5 for DSM-5, Research Version; SCID-5-RV)*. American Psychiatric Association, Arlington, VA.

Fonov, V., Evans, A., McKinstry, R., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47, Supplement 1, S102.

Golomb, J. D., McDavitt, J. R. B., Ruf, B. M., Chen, J. I., Saricicek, A., Maloney, K. H., Hu, J., Chun, M. M., & Bhagwagar, Z. (2009). Enhanced visual motion perception in major depressive disorder. *J. Neurosci.*, 29(28), 9072–9077.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, 5, 13.

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., Poline, J.-B., Rokem, A., Schaefer, G., Sochat, V., Triplett, W., Turner, J. A., Varoquaux, G., & Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data*, 3, 160044.

Gorgolewski, K. J., Esteban, O., Markiewicz, C. J., Ziegler, E., Ellis, D. G., Notter, M. P., Jarecka, D., Johnson, H., Burns, C., Manhães-Savio, A., Hamalainen, C., Yvernault, B., Salo, T., Jordan, K., Goncalves, M., Waskom, M., Clark, D., Wong, J., Loney, F., Modat, M., Dewey, B. E., Madison, C., Visconti di Oleggio Castello, M., Clark, M. G., Dayan, M., Clark, D., Keshavan, A., Pinsard, B., Gramfort, A., Berleant, S., Nielson, D. M., Bougacha, S., Varoquaux, G., Cipollini, B., Markello, R., Rokem, A., Moloney,

B., Halchenko, Y. O., Wassermann, D., Hanke, M., Horea, C., Kaczmarzyk, J., de Hollander, G., DuPre, E., Gillman, A., Mordom, D., Buchanan, C., Tungaraza, R., Pauli, W. M., Iqbal, S., Sikka, S., Mancini, M., Schwartz, Y., Malone, I. B., Dubois, M., Frohlich, C., Welch, D., Forbes, J., Kent, J., Watanabe, A., Cumba, C., Huntenburg, J. M., Kastman, E., Nichols, B. N., Eshaghi, A., Ginsburg, D., Schaefer, A., Acland, B., Giavasis, S., Kleesiek, J., Erickson, D., Küttner, R., Haselgrove, C., Correa, C., Ghayoor, A., Liem, F., Millman, J., Haehn, D., Lai, J., Zhou, D., Blair, R., Glatard, T., Renfro, M., Liu, S., Kahn, A. E., Pérez-García, F., Triplett, W., Lampe, L., Stadler, J., Kong, X.-Z., Hallquist, M., Chetverikov, A., Salvatore, J., Park, A., Poldrack, R., Craddock, R. C., Inati, S., Hinds, O., Cooper, G., Perkins, L. N., Marina, A., Mattfeld, A., Noel, M., Snoek, L., Matsubara, K., Cheung, B., Rothmei, S., Urchs, S., Durnez, J., Mertz, F., Geisler, D., Floren, A., Gerhard, S., Sharp, P., Molina-Romero, M., Weinstein, A., Broderick, W., Saase, V., Andberg, S. K., Harms, R., Schlamp, K., Arias, J., Papadopoulos Orfanos, D., Tarbert, C., Tambini, A., De La Vega, A., Nickson, T., Brett, M., Falkiewicz, M., Podranski, K., Linkersdörfer, J., Flandin, G., Ort, E., Shachnev, D., McNamee, D., Davison, A., Varada, J., Schwabacher, I., Pellman, J., Perez-Guevara, M., Khanuja, R., Pannetier, N., McDermottroe, C., & Ghosh, S. (2018). Nipype. *Software*.

Gotlib, I. H. & Joormann, J. (2010). Cognition and depression: current status and future directions. *Annu. Rev. Clin. Psychol.*, 6, 285–312.

Grabner, G., Janke, A. L., Budge, M. M., Smith, D., Pruessner, J., & Collins, D. L. (2006). Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. *Med. Image Comput. Comput. Assist. Interv.*, 9(Pt 2), 58–66.

Greicius, M. D., Flores, B. H., Menon, V., Glover, G. H., Solvason, H. B., Kenna, H., Reiss, A. L., & Schatzberg, A. F. (2007). Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biol. Psychiatry*, 62(5), 429–437.

Greve, D. N. & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1), 63–72.

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nat. Neurosci.*, 7(5), 555–562.

Hallion, L. S. & Ruscio, A. M. (2011). A meta-analysis of the effect of cognitive bias modification on anxiety and depression. *Psychol. Bull.*, 137(6), 940–958.

Hamilton, J. P., Furman, D. J., Chang, C., Thomason, M. E., Dennis, E., & Gotlib, I. H. (2011). Default-mode and task-positive network activity in major depressive disorder: implications for adaptive and maladaptive rumination. *Biol. Psychiatry*, 70(4), 327–333.

Hamilton, J. P., Glover, G. H., Bagarinao, E., Chang, C., Mackey, S., Sacchet, M. D., & Gotlib, I. H. (2016). Effects of salience-network-node neurofeedback training on affective biases in major depressive disorder. *Psychiatry Res.*, 249, 91–96.

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends Cogn. Sci.*, 16(2), 114–121.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664), 1634–1640.

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), 2425–2430.

Hellrung, L., Dietrich, A., Hollmann, M., Pleger, B., Kalberlah, C., Roggenhofer, E., Villringer, A., & Horstmann, A. (2018). Intermittent compared to continuous real-time fMRI neurofeedback boosts control over amygdala activation. *Neuroimage*, 166, 198–208.

Hilland, E., Landrø, ', N. I., Harmer, C. J., Browning, M., Maglanoc, L. A., & Jonassen, R. (2019). Attentional bias modification is associated with fMRI response toward negative stimuli in individuals with residual depression: a randomized controlled trial. *J. Psychiatry Neurosci.*, 44(5), 1–11.

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. *J. Neurophysiol.*, 109(5), 1250–1258.

Holmes, A. J. & Pizzagalli, D. A. (2007). Task feedback effects on conflict monitoring and executive control: relationship to subclinical measures of depression. *Emotion*, 7(1), 68–76.

Huntenburg, J. M. (2014). *Evaluating nonlinear coregistration of BOLD EPI and T1w images*. Master's thesis, Freie Universität, Berlin.

Iacoviello, B. M., Wu, G., Alvarez, E., Huryk, K., Collins, K. A., Murrough, J. W., Iosifescu, D. V., & Charney, D. S. (2014). Cognitive-emotional training as an intervention for major depressive disorder. *Depress. Anxiety*, 31(8), 699–706.

James, W. (1918). *The Principles of Psychology*. New York: H. Holt and Company.

Jenkinson, M. (2003). Fast, automated, n-dimensional phase-unwrapping algorithm. *Magn. Reson. Med.*, 49(1), 193–197.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2), 825–841.

Jones, E. B. & Sharpe, L. (2017). Cognitive bias modification: A review of meta-analyses. *J. Affect. Disord.*, 223, 175–183.

Joormann, J. (2004). Attentional bias in dysphoria: The role of inhibitory processes. *Cogn. Emot.*, 18(1), 125–147.

Joormann, J. (2019). Is the glass half empty or half full and does it even matter? cognition, emotion, and psychopathology. *Cogn. Emot.*, 33(1), 133–138.

Joormann, J., Waugh, C. E., & Gotlib, I. H. (2015). Cognitive bias modification for interpretation in major depression: Effects on memory and stress reactivity. *Clin. Psychol. Sci.*, 3(1), 126–139.

Kadosh, K. C. & Staunton, G. (2019). A systematic review of the psychological factors that influence neurofeedback learning outcomes. *Neuroimage*, 185, 545–555.

Kaiser, R. H., Andrews-Hanna, J. R., Wager, T. D., & Pizzagalli, D. A. (2015). Large-Scale network dysfunction in major depressive disorder: A meta-analysis of Resting-State functional connectivity. *JAMA Psychiatry*, 72(6), 603–611.

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. *Behav. Res. Ther.*, 46, 1238–1243.

Keynan, J. N., Cohen, A., Jackont, G., Green, N., Goldway, N., Davidov, A., Meir-Hasson, Y., Raz, G., Intrator, N., Fruchter, E., Ginat, K., Laska, E., Cavazza, M., & Hendler, T. (2019). Electrical fingerprint of the amygdala guides neurofeedback training for stress resilience. *Nature Human Behaviour*, 3(1), 63–73.

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Neto, E. C., & Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLOS Computational Biology*, 13(2), e1005350.

Koster, E. H. W., Hoorelbeke, K., Onraedt, T., Owens, M., & Derakshan, N. (2017). Cognitive control interventions for depression: A systematic review of findings from training studies. *Clin. Psychol. Rev.*, 53, 79–92.

Kruijt, A.-W., Putman, P., & Van der Does, W. (2013). A multiple case series analysis of six variants of attentional bias modification for depression. *ISRN Psychiatry*, 2013, 414170.

LaConte, S. M., Peltier, S. J., & Hu, X. P. (2007). Real-time fMRI using brain-state classification. *Hum. Brain Mapp.*, 28(10), 1033–1044.

Lanczos, C. (1964). Evaluation of noisy data. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, 1(1), 76–85.

LANG & PJ (2005). International affective picture system (IAPS) : affective ratings of pictures and instruction manual. *Tech. Rep. NAVTRADEVCEN*.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cognition and Emotion*, 24(8), 1377–1388.

LeMoult, J., Colich, N., Joormann, J., Singh, M. K., Eggleston, C., & Gotlib, I. H. (2018). Interpretation bias training in depressed adolescents: Near- and Far-Transfer effects. *J. Abnorm. Child Psychol.*, 46(1), 159–167.

Li, H., Wei, D., Browning, M., Du, X., Zhang, Q., & Qiu, J. (2016a). Attentional bias modification (ABM) training induces spontaneous brain activity changes in young women with subthreshold depression: a randomized controlled trial. *Psychol. Med.*, 46(5), 909–920.

Li, Z., Tong, L., Wang, L., Li, Y., He, W., Guan, M., & Yan, B. (2016b). Self-regulating positive emotion networks by feedback of multiple emotional brain states using real-time fMRI. *Exp. Brain Res.*, 234(12), 3575–3586.

Linden, D. E. J. (2014). Neurofeedback and networks of depression. *Dialogues Clin. Neurosci.*, 16(1), 103–112.

Linden, D. E. J., Habes, I., Johnston, S. J., Linden, S., Tatineni, R., Subramanian, L., Sorger, B., Healy, D., & Goebel, R. (2012). Real-time self-regulation of emotion networks in patients with depression. *PLoS One*, 7(6).

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, 453(7197), 869–878.

Lui, S., Wu, Q., Qiu, L., Yang, X., Kuang, W., Chan, R. C. K., Huang, X., Kemp, G. J., Mechelli, A., & Gong, Q. (2011). Resting-state functional connectivity in treatment-resistant depression. *Am. J. Psychiatry*, 168(6), 642–648.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The karolinska directed emotional faces (KDEF). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91, 630.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The chicago face database: A free stimulus set of faces and norming data. *Behav. Res. Methods*, 47(4), 1122–1135.

MacDuffie, K. E., MacInnes, J., Dickerson, K. C., Eddington, K. M., Strauman, T. J., & Adcock, R. A. (2018). Single session real-time fMRI neurofeedback has a lasting impact on cognitive behavioral therapy strategies. *Neuroimage Clin*, 19, 868–875.

Marchewka, A., Zurawski, Ł., Jednoróg, K., & Grabowska, A. (2014). The nencki affective picture system (NAPS): introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behav. Res. Methods*, 46(2), 596–610.

Mathews, A. & Mackintosh, B. (2000). Induced emotional interpretation bias and anxiety. *J. Abnorm. Psychol.*, 109(4), 602–615.

Mehler, D. M. A., Sokunbi, M. O., Habes, I., Barawi, K., Subramanian, L., Range, M., Evans, J., Hood, K., Lührs, M., Keedwell, P., Goebel, R., & Linden, D. E. J. (2018). Targeting the affective brain—a randomized controlled trial of real-time fMRI neurofeedback in patients with depression. *Neuropsychopharmacology*.

Meir-Hasson, Y., Kinreich, S., Podlipsky, I., Hendler, T., & Intrator, N. (2014). An EEG Finger-Print of fMRI deep regional activation. *Neuroimage*, 102 Pt 1, 128–141.

Mennen, A. C., Norman, K. A., & Turk-Browne, N. B. (2019). Attentional bias in depression: understanding mechanisms to improve training and treatment. *Current Opinion in Psychology*, 29, 266–273.

Möbius, M., Ferrari, G. R. A., van den Bergh, R., Becker, E. S., & Rinck, M. (2018). Eye-Tracking based attention bias modification (ET-ABM) facilitates disengagement from negative stimuli in dysphoric individuals. *Cognit. Ther. Res.*, 42(4), 408–420.

Mogg, K. & Bradley, B. P. (2005). Attentional bias in generalized anxiety disorder versus depressive disorder. *Cognit. Ther. Res.*, 29(1), 29–45.

Montgomery, S. A. & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *Br. J. Psychiatry*, 134, 382–389.

Nguyen, M., Vanderwal, T., & Hasson, U. (2019). Shared understanding of narratives is correlated with shared neural responses. *Neuroimage*, 184, 161–170.

Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking rumination. *Perspect. Psychol. Sci.*, 3(5), 400–424.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.*, 10(9), 424–430.

Oblak, E. F., Lewis-Peacock, J. A., & Sulzer, J. S. (2017). Self-regulation strategy, feedback timing and hemodynamic properties modulate learning in a simulated fMRI neurofeedback environment. *PLoS Comput. Biol.*, 13(7), e1005681.

Oblak, E. F., Sulzer, J. S., & Lewis-Peacock, J. A. (2019). A simulation-based approach to improve decoded neurofeedback performance. *Neuroimage*, 195, 300–310.

Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H., & Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proc. Natl. Acad. Sci. U. S. A.*, 89(13), 5951–5955.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behav. Res. Methods*, 51(1), 195–203.

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), 1963–1966.

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., & Petersen, S. E. (2011). Functional network organization of the human brain. *Neuron*, 72(4), 665–678.

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fmri. *NeuroImage*, 84(Supplement C), 320–341.

Price, R. B., Siegle, G. J., Silk, J. S., Ladouceur, C., McFarland, A., Dahl, R. E., & Ryan, N. D. (2013). Sustained neural alterations in anxious youth performing an attentional bias task: a pupilometry study. *Depress. Anxiety*, 30(1), 22–30.

Ramot, M., Kimmich, S., Gonzalez-Castillo, J., Roopchansingh, V., Popal, H., White, E., Gotts, S. J., & Martin, A. (2017). Direct modulation of aberrant brain network connectivity through real-time NeuroFeedback. *Elife*, 6, e28974.

Sanchez, A., Everaert, J., & Koster, E. H. W. (2016). Attention training through gaze-contingent feedback: Effects on reappraisal and negative emotions. *Emotion*, 16(7), 1074–1085.

Sanchez, A., Vazquez, C., Marker, C., LeMoult, J., & Joormann, J. (2013). Attentional disengagement predicts stress recovery in depression: an eye-tracking study. *J. Abnorm. Psychol.*, 122(2), 303–313.

Sanchez-Lopez, A., De Raedt, R., van Put, J., & Koster, E. H. W. (2019a). A novel process-based approach to improve resilience: Effects of computerized mouse-based (gaze)contingent attention training (MCAT) on reappraisal and rumination. *Behav. Res. Ther.*, 118, 110–120.

Sanchez-Lopez, A., Everaert, J., Van Put, J., De Raedt, R., & Koster, E. H. W. (2019b). Eye-gaze contingent attention training (ECAT): Examining the causal role of attention regulation in reappraisal and rumination. *Biol. Psychol.*, 142, 116–125.

Satterthwaite, T. D., Cook, P. A., Bruce, S. E., Conway, C., Mikkelsen, E., Satchell, E., Vandekar, S. N., Durbin, T., Shinohara, R. T., & Sheline, Y. I. (2016). Dimensional depression severity in women with major depression and post-traumatic stress disorder correlates with fronto-amygdalar hypoconnectivty. *Mol. Psychiatry*, 21(7), 894–902.

Schapiro, A. C., McDevitt, E. A., Rogers, T. T., Mednick, S. C., & Norman, K. A. (2018). Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nat. Commun.*, 9(1), 3920.

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. *Biol. Mood Anxiety Disord.*, 5(1), 424.

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, 91(6), 1402–1412.

Segrave, R. A., Arnold, S., Hoy, K., & Fitzgerald, P. B. (2014). Concurrent cognitive control training augments the antidepressant efficacy of tDCS: a pilot study. *Brain Stimul.*, 7(2), 325–331.

Sheline, Y. I., Barch, D. M., Donnelly, J. M., Ollinger, J. M., Snyder, A. Z., & Mintun, M. A. (2001). Increased amygdala response to masked emotional faces in depressed subjects resolves with antidepressant treatment: an fMRI study. *Biol. Psychiatry*, 50(9), 651–658.

Sheline, Y. I., Barch, D. M., Price, J. L., Rundle, M. M., Vaishnavi, S. N., Snyder, A. Z., Mintun, M. A., Wang, S., Coalson, R. S., & Raichle, M. E. (2009). The default mode network and self-referential processes in depression. *Proc. Natl. Acad. Sci. U. S. A.*, 106(6), 1942–1947.

Sheline, Y. I., Price, J. L., Yan, Z., & Mintun, M. A. (2010). Resting-state functional MRI in depression unmasks increased connectivity between networks via the dorsal nexus. *Proc. Natl. Acad. Sci. U. S. A.*, 107(24), 11020–11025.

Shibata, K., Lisi, G., Cortese, A., Watanabe, T., Sasaki, Y., & Kawato, M. (2019). Toward a comprehensive understanding of the neural mechanisms of decoded neurofeedback. *Neuroimage*, 188, 539–556.

Shou, H., Yang, Z., Satterthwaite, T. D., Cook, P. A., Bruce, S. E., Shinohara, R. T., Rosenberg, B., & Sheline, Y. I. (2017). Cognitive behavioral therapy increases amygdala connectivity with the cognitive control network in both MDD and PTSD. *Neuroimage Clin*, 14, 464–470.

Siegle, G. J., Ghinassi, F., & Thase, M. E. (2007a). Neurobehavioral therapies in the 21st century: Summary of an emerging field and an extended example of cognitive control training for depression. *Cognit. Ther. Res.*, 31(2), 235–262.

Siegle, G. J., Price, R. B., Jones, N. P., Ghinassi, F., Painter, T., & Thase, M. E. (2014). You gotta work at it: Pupillary indices of task focus are prognostic for response to a neurocognitive intervention for rumination in depression. *Clin. Psychol. Sci.*, 2(4), 455–471.

Siegle, G. J., Thompson, W., Carter, C. S., Steinhauer, S. R., & Thase, M. E. (2007b). Increased amygdala and decreased dorsolateral prefrontal BOLD responses in unipolar depression: related and independent features. *Biol. Psychiatry*, 61(2), 198–209.

Sitaram, R., Ros, T., Stoeckel, L., Haller, S., Scharnowski, F., Lewis-Peacock, J., Weiskopf, N., Blefari, M. L., Rana, M., Oblak, E., Birbaumer, N., & Sulzer, J. (2017). Closed-loop brain training: the science of neurofeedback. *Nat. Rev. Neurosci.*, 18(2), 86–100.

Snyder, H. R. (2013). Major depressive disorder is associated with broad impairments on neuropsychological measures of executive function: a meta-analysis and review. *Psychol. Bull.*, 139(1), 81–132.

Sorger, B., Scharnowski, F., Linden, D. E. J., Hampson, M., & Young, K. D. (2019). Control freaks: Towards optimal selection of control conditions for fMRI neurofeedback studies. *Neuroimage*, 186, 256–265.

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, 245–255.

Sulzer, J., Sitaram, R., Blefari, M. L., Kollias, S., Birbaumer, N., Stephan, K. E., Luft, A., & Gassert, R. (2013). Neurofeedback-mediated self-regulation of the dopaminergic midbrain. *Neuroimage*, 83, 817–825.

Suslow, T., Konrad, C., Kugel, H., Rumstadt, D., Zwitserlood, P., Schöning, S., Ohrmann, P., Bauer, J., Pyka, M., Kersting, A., Arolt, V., Heindel, W., & Dannlowski, U. (2010). Automatic mood-congruent amygdala responses to masked facial expressions in major depression. *Biol. Psychiatry*, 67(2), 155–160.

Taschereau-Dumouchel, V., Cortese, A., Chiba, T., Knotts, J. D., Kawato, M., & Lau, H. (2018). Towards an unconscious neural reinforcement intervention for common fears. *Proc. Natl. Acad. Sci. U. S. A.*, 115(13), 3470–3475.

Taschereau-Dumouchel, V., Kawato, M., & Lau, H. (2019). Multivoxel pattern analysis reveals dissociations between subjective fear and its physiological correlates. *Mol. Psychiatry*.

Thibault, R. T., MacPherson, A., Lifshitz, M., Roth, R. R., & Raz, A. (2018). Neurofeedback with fMRI: A critical systematic review. *Neuroimage*, 172, 786–807.

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, B. J., & Nelson, C. (2009). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry Res.*, 168(3), 242–249.

Treiber, J. M., White, N. S., Steed, T. C., Bartsch, H., Holland, D., Farid, N., McDonald, C. R., Carter, B. S., Dale, A. M., & Chen, C. C. (2016). Characterization and correction of geometric distortions in 814 diffusion weighted images. *PLOS ONE*, 11(3), e0152472.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4itk: Improved n3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6), 1310–1320.

Uys, J. D. K., Marais, L., Faure, J., Prevoo, D., Swart, P., Mohammed, A. H., Stein, D. J., & Daniels, W. M. U. (2006). Developmental trauma is associated with behavioral hyperarousal, altered HPA axis activity, and decreased hippocampal neurotrophin expression in the adult rat. *Ann. N. Y. Acad. Sci.*, 1071, 542–546.

Vazquez, C., Blanco, I., Sanchez, A., & McNally, R. J. (2016). Attentional bias modification in depression through gaze contingencies and regulatory control using a new eye-tracking intervention paradigm: study protocol for a placebo-controlled trial. *BMC Psychiatry*, 16(1), 439.

Victor, T. A., Furey, M. L., Fromm, S. J., Bellgowan, P. S. F., Öhman, A., & Drevets, W. C. (2012). The extended functional neuroanatomy of emotional processing biases for masked faces in major depressive disorder. *PLoS One*, 7(10), e46439.

Victor, T. A., Furey, M. L., Fromm, S. J., Ohman, A., & Drevets, W. C. (2010). Relationship between amygdala responses to masked faces and mood state and treatment in major depressive disorder. *Arch. Gen. Psychiatry*, 67(11), 1128–1138.

Victor, T. A., Furey, M. L., Fromm, S. J., Öhman, A., & Drevets, W. C. (2013). Changes in the neural correlates of implicit emotional face processing during antidepressant treatment in major depressive disorder. *Int. J. Neuropsychopharmacol.*, 16(10), 2195–2208.

Vodrahalli, K., Chen, P.-H., Liang, Y., Baldassano, C., Chen, J., Yong, E., Honey, C., Hasson, U., Ramadge, P., Norman, K., & Arora, S. (2016). Mapping between fMRI responses to movies and their natural language annotations.

Wagner, G., Sinsel, E., Sobanski, T., Köhler, S., Marinou, V., Mentzel, H.-J., Sauer, H., & Schlösser, R. G. M. (2006). Cortical inefficiency in patients with unipolar depression: an event-related FMRI study with the stroop task. *Biol. Psychiatry*, 59(10), 958–965.

Wang, S., Peterson, D. J., Gatenby, J. C., Li, W., Grabowski, T. J., & Madhyastha, T. M. (2017). Evaluation of field map and nonlinear registration methods for correction of susceptibility artifacts in diffusion MRI. *Frontiers in Neuroinformatics*, 11.

Wang, Y., Keller, B., Capota, M., Anderson, M. J., Sundaram, N., Cohen, J. D., Li, K., Turk-Browne, N. B., & Willke, T. L. (2016). Real-time full correlation matrix analysis of fMRI data. In *2016 IEEE International Conference on Big Data (Big Data)* (pp. 1242–1251).

Watanabe, T., Sasaki, Y., Shibata, K., & Kawato, M. (2017). Advances in fMRI Real-Time neurofeedback. *Trends Cogn. Sci.*, 21(12), 997–1010.

Wells, T. T. & Beevers, C. G. (2010). Cognition and emotion biased attention and dysphoria: Manipulating selective attention reduces subsequent depressive symptoms. *Cognition & Emotion*, 24(4), 719–728.

Wiers, C. E. & Wiers, R. W. (2017). Imaging the neural effects of cognitive bias modification training. *Neuroimage*, 151, 81–91.

Williams, J. B. W. & Kobak, K. A. (2008). Development and reliability of a structured interview guide for the montgomery asberg depression rating scale (SIGMA). *Br. J. Psychiatry*, 192(1), 52–58.

Xia, C. H., Ma, Z., Ciric, R., Gu, S., Betzel, R. F., Kaczkurkin, A. N., Calkins, M. E., Cook, P. A., García de la Garza, A., Vandekar, S. N., Cui, Z., Moore, T. M., Roalf, D. R., Ruparel, K., Wolf, D. H., Davatzikos, C., Gur, R. C., Gur, R. E., Shinohara, R. T., Bassett, D. S., & Satterthwaite, T. D. (2018). Linked dimensions of psychopathology and connectivity in functional brain networks. *Nat. Commun.*, 9(1), 3003.

Yamada, T., Hashimoto, R.-I., Yahata, N., Ichikawa, N., Yoshihara, Y., Okamoto, Y., Kato, N., Takahashi, H., & Kawato, M. (2017). Resting-State functional Connectivity-Based biomarkers and functional MRI-Based neurofeedback for psychiatric disorders: A challenge for developing theranostic biomarkers. *Int. J. Neuropsychopharmacol.*, 20(10), 769–781.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods*, 8(8), 665–670.

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., & Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.*, 106(3), 1125–1165.

Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C. J., & Hasson, U. (2017). Same story, different story. *Psychol. Sci.*, 28(3), 307–319.

Young, K. D., Siegle, G. J., Zotev, V., Phillips, R., Misaki, M., Yuan, H., Drevets, W. C., & Bodurka, J. (2017). Randomized clinical trial of Real-Time fMRI amygdala neurofeedback for major depressive disorder: Effects on symptoms and autobiographical memory recall. *Am. J. Psychiatry*, 174(8), 748–755.

Young, K. D., Zotev, V., Phillips, R., Misaki, M., Drevets, W. C., & Bodurka, J. (2018). Amygdala real-time functional magnetic resonance imaging neurofeedback for major depressive disorder: A review. *Psychiatry Clin. Neurosci.*

Young, K. D., Zotev, V., Phillips, R., Misaki, M., Yuan, H., Drevets, W. C., & Bodurka, J. (2014). Real-time FMRI neurofeedback training of amygdala activity in patients with major depressive disorder. *PLoS One*, 9(2), e88785.

Yu, M., Linn, K. A., Shinohara, R. T., Oathes, D. J., Cook, P. A., Duprat, R., Moore, T. M., Oquendo, M. A., Phillips, M. L., McInnis, M., Fava, M., Trivedi, M. H., McGrath, P., Parsey, R., Weissman, M. M., & Sheline, Y. I. (2019). Childhood trauma history is linked to abnormal brain connectivity in major depression. *Proc. Natl. Acad. Sci. U. S. A.*

Yuan, H., Young, K. D., Phillips, R., Zotev, V., Misaki, M., & Bodurka, J. (2014). Resting-state functional connectivity modulation and sustained changes after real-time functional magnetic resonance imaging neurofeedback training in depression. *Brain Connect.*, 4(9), 690–701.

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*, 31(3), 1116–1128.

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1), 45–57.

Zotev, V., Yuan, H., Misaki, M., Phillips, R., Young, K. D., Feldner, M. T., & Bodurka, J. (2016). Correlation between amygdala BOLD activity and frontal EEG asymmetry during real-time fMRI neurofeedback training in patients with depression. *NeuroImage: Clinical*, 11, 224–238.