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

2014

FOCUS VERSUS BREADTH: THE EFFECTS OF NEURAL GAIN ON INFORMATION PROCESSING

Eldar, Eran

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp019k41zd65m

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

FOCUS VERSUS BREADTH:
THE EFFECTS OF NEURAL GAIN ON INFORMATION PROCESSING

Eran Eldar

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: Yael Niv

June 2014

© Copyright by Eran Eldar, 2014. All rights reserved.

Abstract

We propose that brain-wide variations in neural gain control the degree to which information is processed in a broad, integrative manner, or conversely, in a narrowly focused manner. Neural gain, which is thought to be modulated throughout the brain by the locus coeruleus-norepinephrine system, can be thought of as a contrast control mechanism. When gain is high, the contrast between weakly and strongly activated neurons is increased, and thus, we expect processing to be more narrowly focused on the most strongly represented sources of information. In contrast, low gain may allow the integration of a broader range of information. We first investigate the whole-brain effects of neural gain using functional connectivity and graph-theoretic analyses of neuroimaging data, in conjunction with pupil diameter indices of norepinephrine function, as well as in response to a norepinephrine-enhancing drug. The results reveal signs of brain-wide fluctuations in gain that are tracked by pupillary indices, and suggest that high gain has a focusing, clustering effect on neural interactions throughout the brain. We then present three sets if experiments designed to investigate the behavioral effects of variations in gain. In the first experiment, we show that pupillary and neuroimaging indices of high gain are associated with learning that is more focused on particular types of stimulus features, in accordance with individual predisposition. In the second set of experiments, we show that high gain has a similar effect on perception and memory, making them more focused and less integrative, and that the effects of gain do not have to be tied to individual predisposition, but rather, they can be flexibly manipulated by means of subliminal priming or experimental task. In the third set of experiments, we show that the reduced integration that is associated with high gain comes with a benefit – weaker susceptibility to classical decision

making biases. Finally, we end by integrating the results presented throughout the thesis into a coherent Bayesian account, and discuss how this account may serve as a basis for a cognitive theory of autism.

Contents

Abstract ..... **III**

Acknowledgments ..... **IX**

1 Introduction **1**

1.1 Gain modulation as contrast control 2

1.2 From molecules to systems to behavior 3

1.3 Organization of the thesis 4

2 Background **6**

2.1 Neural gain 6

2.2 Selective gain modulation 7

2.3 Brain-wide gain modulation 8

2.3.1 The locus coeruleus and norepinephrine 9

2.3.2 Norepinephrine and neural gain 9

2.3.3 Evidence at the system level 13

2.4 Conclusion 17

3 The whole-brain effects of neural gain **18**

3.1 Changes in activity 19

3.2 Changes in functional connectivity 20

3.2.1 Global fluctuations in local functional connectivity 21

3.2.2 Functional connectivity and pupil diameter 24

3.2.3 The distribution of functional connections 25

3.3 Pharmacological manipulation 30

3.3.1 Functional connectivity strength and clustering ... 31
3.3.2 Signs of decreased gain? ... 34
3.4 Discussion ... 36
3.5 Conclusion ... 38
3.6 Appendix: Methods ... 38
3.6.1 Pupil diameter study ... 38
3.6.2 Pharmacological manipulation study ... 41
3.6.3 fMRI data processing ... 43
3.6.4 Neural network model ... 47
4 Neural gain and the focus of learning ... 49
4.1 A neural network model ... 52
4.2 Pupil diameter and adherence to predispositions ... 53
4.3 Functional connectivity clustering and adherence to predispositions ... 59
4.4 Discussion ... 60
4.5 Conclusion ... 63
4.6 Appendix: Methods ... 63
4.6.1 Neural network model ... 63
4.6.2 Experimental methodology ... 66
5 Manipulating the effect of gain on perception and memory ... 70
5.1 Introduction ... 71
5.2 A neural network model of perceptual integration of letter shape and context ... 72
5.3 Pupil diameter and perceptual processing ... 74
5.4 An attentional manipulation ... 76
5.5 Within-participant variations ... 76

5.6 Voluntary direction of attention 79

5.7 Conclusion 81

5.8 Appendix: Methods 82

5.8.1 Neural network model 82

5.8.2 Ambiguous letters experiment 85

5.8.3 Recognition memory experiment 89

6 Neural gain and decision making biases 92

6.1 Introduction 92

6.2 Anchoring 94

6.3 Persistence of belief 95

6.4 Framing 96

6.4.1 Attribute framing 96

6.4.2 Risky choice framing 97

6.4.3 Task framing 98

6.5 Sample-size neglect 99

6.6 Overall susceptibility to biases 100

6.7 Computational model 101

6.8 The cost of weaker biases 102

6.9 Discussion 104

6.10 Appendix: Methods 105

6.10.1 Experimental methodology 105

6.10.2 Computational model 113

7 Theoretical and practical implications 115

7.1 A Bayesian perspective 115

7.2 A neural gain account of autism 118

7.3 Open questions 122

7.4 Future directions 123

7.4.1 Information processing styles 123

7.4.2 Dynamic interactions between gain and information processing 124

7.5 Conclusion 124

Bibliography 126

Acknowledgments

I am grateful to Yael Niv, my adviser, for close guidance and support during the past 5 years. Yael's methodical, perfectionistic approach, which was evident in all aspects of our work, has taught me by example what it means to be a professional researcher. Yael's professional attitude, however, did not prevent her from always keeping sight of what matters the most. I am particularly thankful for Yael's unqualified support and understanding during the prolonged illness of my mother, which allowed me to fully devote myself to my family when that was needed.

I also thank Jon Cohen, to whom I am indebted in multiple ways. First, for his personal guidance and teaching. I have learned greatly from Jon's broad and integrative scientific perspective, as well as from his passionate, inspiring approach. Moreover, Jon is responsible for laying the scientific foundations on which my own ideas were based. Finally, Jon co-founded the Princeton Neuroscience Institute, which provided me with an ideal environment to develop as a scientist.

I thank the members of my thesis committee – Matt Botvinick, Josh Gold and Ken Norman – for valuable feedback and advice. Matt and Ken, together with Yael and Jon, formed the core of the stimulating departmental PDP meeting series, which I will surely miss.

I also thank members of the Niv lab in the past and present, especially our spirited lab manager Angela Radulescu, but also Mike Todd, Bob Wilson, Carlos Diuk, Reka Daniel, Nico Schuck, Angela Langdon, Sam Gershman, Stephanie Chan, Andra Geana, Yuan Chang Long, and others, for sharing their experience and thoughts, and helping whenever I needed help.

I am grateful to my brother, Ofer Eldar, and to my partner, Jane Keung, for unconditional personal support. While I was serving in the army, pursuing a PhD seemed an improbable possibility, which was eventually realized largely thanks to my brother's foresight, help and encouragement. Then, at a time in which my home in Israel fell apart due to my mother's illness, Jane made Princeton a home to me. A place where there is always someone waiting for you.

Finally, I would like to dedicate this thesis to my mother, whose love, support and dedication I am forever grateful for. If she was with me today, she would have surely been the happiest person seeing me graduate.

Chapter 1

Introduction

Suppose the brain had a contrast knob that could be used to control the contrast between strongly and weakly active neurons. How would increasing the contrast affect neural processing? Figure 1.1 suggests a possible answer. Increasing the picture's contrast eliminates subtle differences and emphasizes the picture's darkest areas. As a result, many features such as the tree rings are discarded, whereas the darkest features of the picture (in this case, the wood cracks) are emphasized. Similarly, we may hypothesize that increasing the contrast between strongly and weakly active neurons would focus neural processing on features that are represented by the most strongly active neurons, while discarding more weakly represented features. The purpose of this thesis is to test this hypothesis – that the brain has a global contrast-control mechanism whose effect is to focus neural information processing on the most strongly represented features of the information – and to investigate the implications of brain-wide contrast modulation for human decision making in different domains.

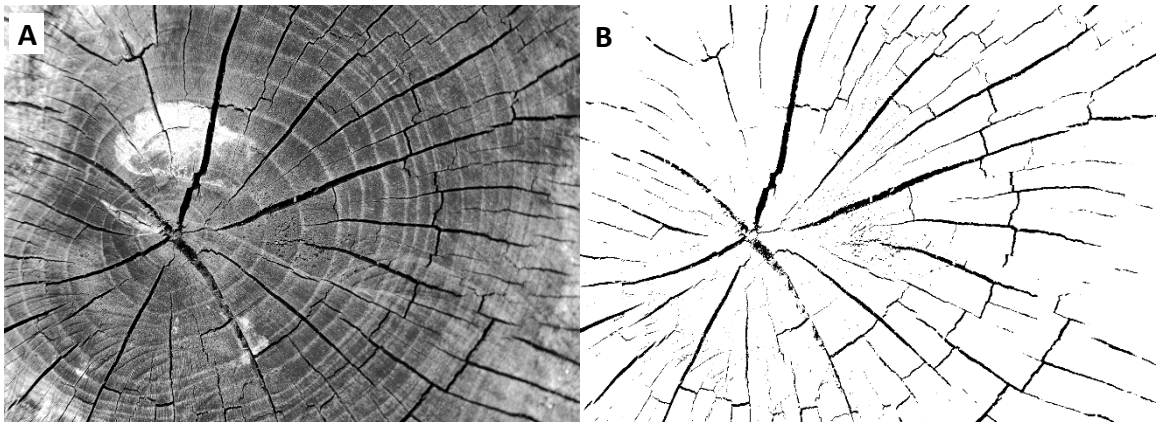

Figure 1.1. The effect of increasing contrast. (A) A grayscale picture of a tree stump. (B) The picture from panel A with increased contrast. Lighter features are not visible whereas the darkest features are emphasized.

1.1 Gain modulation as contrast control

This thesis was inspired by the adaptive gain theory (Aston-Jones & Cohen, 2005), according to which the locus coeruleus-norepinephrine system serves to modulate neural gain throughout the brain (Servan-Schreiber et al., 1990). Brain-wide gain modulation can be naturally understood as modulation of the contrast between excited and inhibited neurons (Figure 1.2). We will later see that this should result, more generally, in modulation of the contrast between strongly and weakly activated neural units, provided that units compete through mutual inhibition. My own contribution to the theory lies in proposing that gain enhancement has an effect that is akin to that of contrast enhancement (Figure 1.1), that is, the focusing of neural processing on the most strongly represented features. Throughout this thesis, I present converging evidence in support of this proposal, including neural network simulations, neuroimaging data and behavioral data.

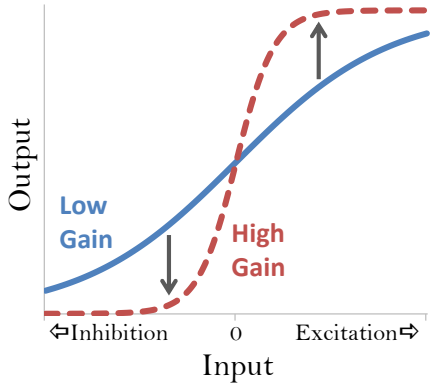

Figure 1.2. Input-output function of a model neuron (or population of neurons) with low and high neural gain. Adapted from Aston-Jones & Cohen (2005).

1.2 From molecules to systems to behavior

The idea of neural gain modulation originated from studies of single-cell electrophysiology in response to the neurotransmitter norepinephrine. Based on this earlier work, I simulate the effect of gain modulation on neural activity and connectivity in neural network models (Rumelhart & McClelland, 1986), which I use to predict whole-brain neuroimaging measures, as well as the behavior of human participants in different experimental tasks. In so doing, I attempt to link a low-level principle of neural function via computational modeling to system-level neural and behavioral phenomena.

While I believe that this constitutes a powerful and promising approach, this approach suffers from the inherent disadvantage of relying on a relatively broad set of assumptions. In particular, the link between the neural network models presented in this thesis and neural processes that take place on the single-cell level is far from obvious. It is not even clear whether the models' units best correspond to single neurons or to population of neurons. This problem is compounded by the indirect and imprecise nature of the indices of neural function that are available in human participants. Thus, the models that I present are best viewed as high-level

abstractions of neural processes, whose validity should be primarily assessed by the accuracy of the system-level predictions that they generate. My hope is that the wide range of model predictions that were borne out by the experimental results will convince the reader of the strength of the hypotheses that the models embody.

1.3 Organization of the thesis

The thesis begins with a background chapter (chapter 2) in which I explain the concept of neural gain and survey evidence in support of a brain-wide gain modulation mechanism that is subserved by the locus coeruleus-norepinephrine system. Existing evidence includes a wide range of findings from cellular electrophysiology studies and behavioral studies, but no system-level neural observations. Thus, in Chapter 3, I investigate the whole-brain effects of neural gain using whole-brain functional connectivity and graph-theoretic analyses of neuroimaging data, in conjunction with pupillary indices of norepinephrine function, as well as in response to a norepinephrine-enhancing drug. The analyses are based on predictions generated from neural network simulations, and are designed to uncover signs of brain-wide fluctuations in gain, and to examine the relationship of these fluctuations with pupillary indices, and with the focusing of neural connectivity.

Next follow three chapters about the behavioral effects of variations in gain. In each of these chapters, I generate a different prediction concerning the effects of gain, illustrated using a neural network model, and then test the prediction in a behavioral experiment in conjunction with pupillometry (and in chapter 4 only, also neuroimaging). Chapter 4 is concerned with the focusing effect of high gain on learning, which is manifested in accordance with individual predisposition. In Chapter 5, I show that high gain has a similar effect on perception and

memory, making them more focused and less integrative, and that the effects of gain do not have to be tied to individual predisposition, but rather, they can be flexibly manipulated by means of subliminal priming or experimental task. Finally, in Chapter 6, I show that the reduced integration that is associated with high gain comes with a benefit – weaker susceptibility to classical decision making biases.

I end the thesis with Chapter 7 in which I integrate the results presented throughout the thesis into a coherent Bayesian account, and discuss how this account may serve as a basis for a cognitive theory of autism.

Chapter 2

Background

A principal aim of cognitive neuroscience is to understand how humans process information. Significant progress has been made in this endeavor by using general principles, abstracted from biological properties of the central nervous system (CNS), to explain system-level neural or behavioral phenomena (e.g., Rumelhart & McClelland, 1986). In keeping with this tradition, this thesis builds on the concept of brain-wide gain modulation, derived from the biological study of the neuromodulator norepinephrine (Servan-Schreiber et al., 1990; Aston-Jones & Cohen, 2005), to explain a range of whole-brain neuroimaging and behavioral data. In this chapter, I will introduce the concept of neural gain and the way it is typically modeled. I will then survey previous evidence in support of the existence of a brain-wide gain modulation mechanism that is realized by the locus-coeruleus-norepinephrine (LC-NE) system. In addition, I will discuss the use of pupillometry as a noninvasive method for tracking levels of LC-NE function and neural gain.

2.1 Neural gain

Neural gain modulation refers to modulation of the impact of incoming neural signals on neural activity. Thus, with higher gain, a neural unit responds with a larger increase in activity to an

excitatory signal and with a larger decrease in activity to an inhibitory signal, as compared to with lower gain (Figure 1.2). Gain is typically modeled as a multiplicative factor that enhances the effect of input on output (Servan-Schreiber et al., 1990):

$$\text{output} = f(gain \cdot \text{input}) \qquad (2.1)$$

where *f* is the activation function, which is used to compute a neural unit's level of activity (i.e., its output) given the input that it receives.

2.2 Selective gain modulation

The concept of neural gain modulation originated from observations of the response profiles of single neurons. For instance, a key study showed that the response of posterior parietal neurons to visual stimuli that lie in their receptive field is multiplicatively modulated by eye position (Andersen et al., 1985). The concept of gain was later generalized to explain the modulation of neural responses throughout sensory and motor cortex (Salinas & Abbott, 1995; McAdams & Maunsell, 1999; Treue & Martınez-Trujillo, 1999) as a function of attention. The simple, yet powerful, idea is that selective attention to specific features of sensory input is brought about by a selective increase in neural gain in neurons that represent these features (Figure 2.1). Selective gain modulation has been suggested to result from changes in overall levels of background synaptic input, both excitatory and inhibitory (Chance et al., 2002).

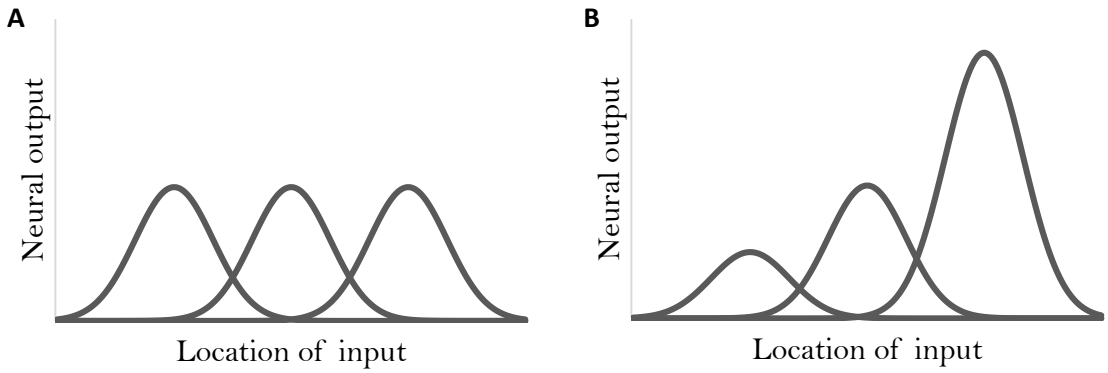

Figure 2.1. Schematic illustration of the way selective gain modulation is thought to mediate selective attention. (A) Response profile of three model neurons as a function of the location of the input (e.g., in visual field). (B) Response profile of the same model neurons when attention is directed to the right. The gain of the neuron that responds to right-sided input is enhanced, whereas the gain of the neuron that responds to left-sided input is diminished. As a result, the network as a whole is more sensitive to right-sided input. Adapted from Salinas & Thier (2000).

2.3 Brain-wide gain modulation

Neural gain may also be modulated non-selectively, increasing or decreasing simultaneously all over the brain. From a modeling perspective, brain-wide gain modulation should follow the same mathematical form as selective gain modulation (Eq. 2.1), except that all network units share the same setting of the gain parameter. Brain-wide gain modulation was first used to model behavioral effects of the neuromodulators norepinephrine and dopamine, which had been suggested to enhance neural gain (Servan-Schreiber et al., 1990). Following up on this early work, and integrating a wide range of more recent evidence, Aston-Jones and Cohen (2005) proposed that the LC-NE system primarily acts by modulating neural gain throughout the brain. This proposal depends on two assertions: first, the LC-NE system issues a uniform brain-

wide NE signal, and second, NE modulates neural gain. In what follows I review evidence in support of each of these assertions in turn.

2.3.1 The locus coeruleus and norepinephrine

The LC nucleus, located in the dorsorostral pons, is the sole source of NE to the cerebral, cerebellar, and hippocampal cortices (Aston-Jones, 2004; Aston-Jones et al., 1984; Moore & Bloom, 1979). Although small in number, LC neurons project widely, to the entire neocortex, as well as to the rest of the forebrain with the exception of the striatum (Jones et al., 1977; Jones & Moore, 1977; Jones & Yang, 1985; Figure 2.2). LC projections are unmyelinated and therefore slowly conducting (Aston-Jones et al., 1985), and some have extra-synaptic release sites that may necessitate further diffusion to exert an effect (Beaudet & Descarries ,1978; Seguela et al., 1990). Taken together, these findings indicate that signals sent by the LC have low spatial and temporal specificity, and thus, are well suited to providing a uniform brain-wide neuromodulatory signal.

2.3.2 Norepinephrine and neural gain

NE, also called noradrenaline, is a monoamine neurotransmitter that is thought to exert its effect in the CNS by means of α and β adrenergic receptors (Devilbiss & Waterhouse, 2004). Early studies of NE function focused on its direct impact on neural activity. Both local application of NE and stimulation of LC input pathways showed that NE suppresses spontaneous spiking of neurons in the cerebellum (Hoffer et al., 1971; Hoffer et al., 1973), cerebral cortex (Olpe et al., 1980; Stone, 1973), hippocampus (Segal & Bloom, 1974a; Segal &

Bloom, 1974b), thalamus (Nakai & Takaori, 1974; Phillis et al., 1967) and hypothalamus (Miyahara & Oomura, 1982). However, the direct effect of a neurotransmitter on neural activity may not fully capture the neurotransmitter's contribution to the operation of a neural circuit. Thus, later studies tested for potential second-order (i.e., neuromodulatory) effects of NE, by examining its impact on neural responses to independent input signals that were mediated by other neurotransmitters.

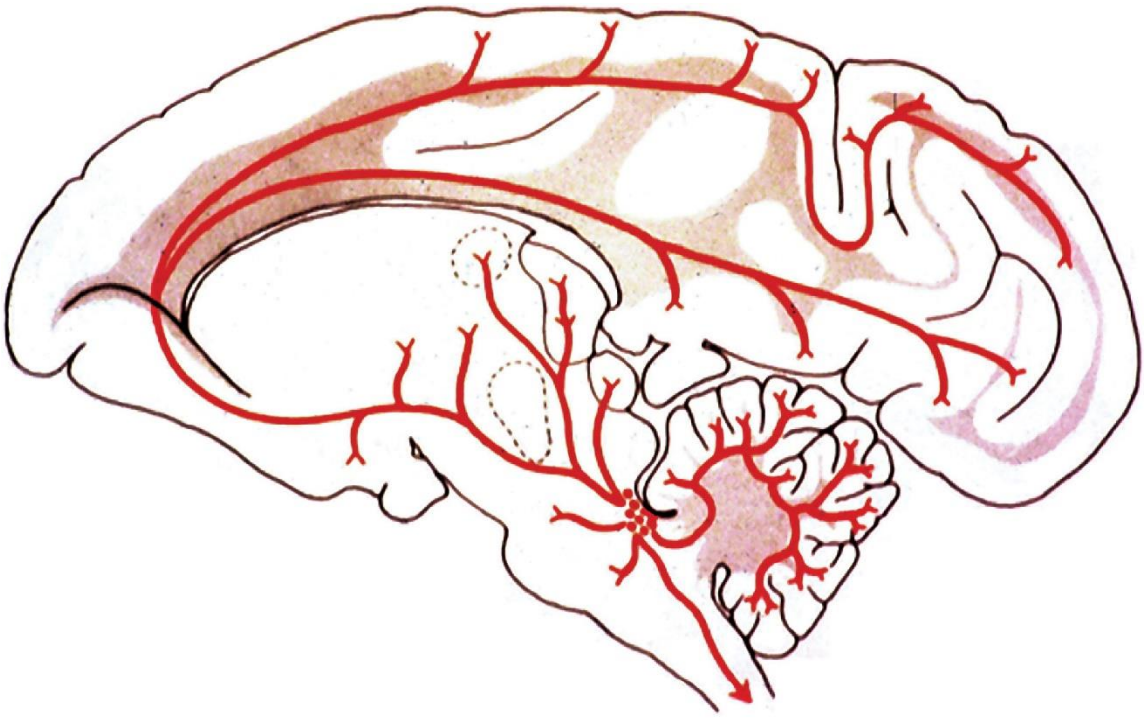

Figure 2.2. Illustration of projections of the LC system. Sagittal view of a monkey brain showing LC neurons located in the pons with efferent projections throughout the central nervous system. Note that only few areas do not receive LC innervation (e.g., hypothalamus and caudate-putamen). Reproduced from Aston-Jones & Cohen (2005).

2.3.2.1 Norepinephrine as a neuromodulator

First attempts to examine the second-order effects of NE revealed that it suppresses spontaneous neural activity more strongly than responses to a co-occurring input signal (Foote et al., 1975; Freedman et al., 1977). A different line of studies, which focused on low levels of NE and weak co-occurring input signals, found that NE actually potentiates, and in some cases even gates, neural response to co-occurring input (Moises et al., 1983; Moises et al., 1981; Moises & Woodward, 1980; Moises et al., 1979; Waterhouse et al., 1980; Waterhouse & Woodward, 1980; Woodward et al., 1979; Waterhouse et al., 1988). NE's neuromodulatory effects were further observed in a wide range of brain regions including visual cortex (Videen et al., 1984), lateral geniculate nucleus (Rogawski & Aghajanian, 1980), lateral hypothalamus (Sessler et al., 1988), hippocampus (Segal & Bloom, 1976), and superior colliculus (Sato & Kayama, 1983). These results suggested that NE sharpens the contrast between neural response to synaptic input and spontaneous neural activity, thus facilitating neural communication.

2.3.2.2 Conflicting evidence

The finding that NE facilitates neural signaling naturally lead to the modeling of NE's effect as increasing neural gain (Servan-Schreiber et al., 1990). Further investigations, however, revealed a more complicated pattern. Brain slice recordings showed that while NE facilitates response to excitation in most cells by means of α_1 receptor activation, in some cells it in fact suppresses response to excitation by activating β receptors (Devilbiss & Waterhouse, 2000). A similar result was observed when the effect of NE was studied in rats by means of tonic LC

stimulation in conjunction with sensory whisker stimulation (Devilbiss & Waterhouse, 2004). Moreover, while LC stimulation mostly facilitated response to whisker stimulation in the ventroposterior medial thalamus, it was as likely to facilitate as to suppress response further downstream, in the barrel field somatosensory cortex. Even in those cells whose response to excitation was facilitated by NE, higher levels of NE facilitated response less strongly, resulting in an inverted U-shaped relationship between NE levels and response to excitation. Finally, other studies reported that NE has more specific effects on neural activity, such as an increase in the selectivity of neural response (Hurley et al., 2004), or a specific facilitation of bottom-up inputs at the expense of top-down inputs (Hasselmo et al., 1997).

These observations challenge the conception that NE strictly increases neural gain. However, many of the observations that seem to conflict with the gain theory may be explained in terms of increased neural gain acting in particular network architectures. For instance, both the suppression of neural responsivity and increased selectivity could result from the potentiation of inhibitory connections. Indeed, some of the neural network simulations that will be presented in later chapters produce just such effects. Furthermore, the higher NE levels that lead to a U-shaped pattern of facilitation may not be reached under physiological conditions. In addition, the gain theory of the LC-NE system is primarily concerned with system-level effects. According to this view, changes in neural gain characterize processes at the system level, which may emerge from a more diverse set of processes acting at the level of single neurons.

2.3.3 Evidence at the system level

We have seen that evidence at the cellular level in support of the LC's role in brain-wide gain modulation, while suggestive, cannot be regarded as conclusive. Further support for the gain theory, however, is provided by studies of the system-level effects of LC activity.

2.3.3.1 The effect of tonic LC-NE activity

Baseline (i.e., tonic) levels of LC activity seem to control the state of arousal and wakefulness of an animal (Aston-Jones & Bloom, 1981a; Hobson et al., 1975; Rajkowski et al., 1997; Rasmussen et al., 1986). Behaviorally, arousal and wakefulness are characterized by higher sensitivity to sensory stimulation and more vigorous motor activity, both of which could naturally be explained by the facilitation of neural signaling that is brought about by increased neural gain. Indeed, increased tonic levels of gain have been used to model the observed boosting of signal-detection performance by CNS stimulants that increase NE and dopamine levels (Servan-Schreiber et al., 1990; Servan-Schreiber et al., 1998a; Servan-Schreiber et al., 1998b).

2.3.3.2 The effect of phasic LC-NE activity

In addition to its tonic levels of activity, the LC responds transiently (i.e., in a phasic manner) to salient sensory stimuli that elicit a behavioral response (Aston-Jones & Bloom, 1981b; Foote et al., 1980; Grant et al., 1988). More specifically, within the context of an experimental task, the LC responds strongly to target stimuli but minimally or not at all to distractors (Aston-Jones et al., 1994). This finding was confirmed in reversal experiments, in which the LC

responded preferentially to a stimulus when it constituted a target, but not when the same stimulus appeared as a distractor (Aston-Jones et al., 1997). Critically, LC response was time locked to the behavioral response, not to stimulus onset, though it only preceded a response induced by a task stimulus (Figure 2.3; Clayton et al., 2004; Bouret & Sara, 2004). These findings suggest that phasic LC responses might serve to facilitate the execution of task decisions, an effect that has been theoretically explained by an increase in neural gain (Usher et al., 1999).

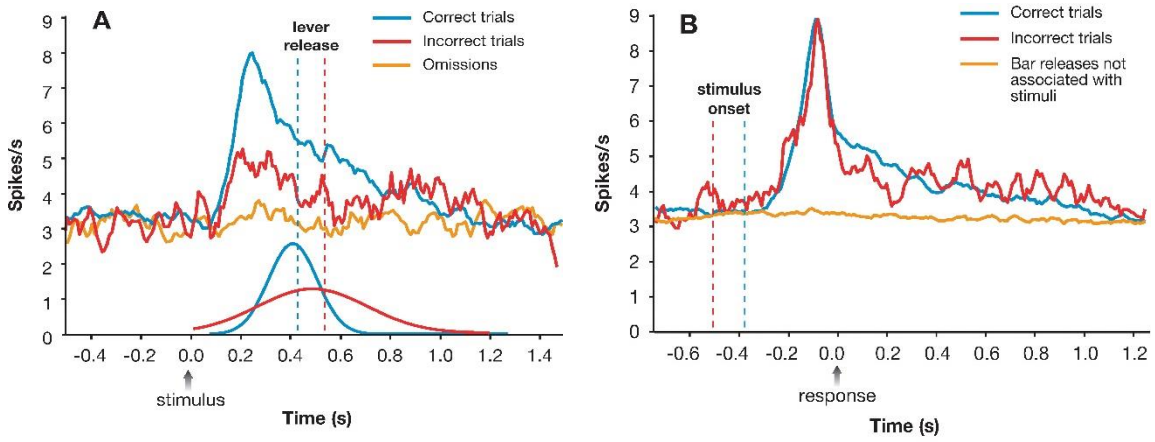

Figure 2.3. Phasic activation of monkey LC neurons in a two-alternative forced choice task. Stimulus- and response-locked population peri-event time histograms (PETH) showing LC responses for trials yielding correct and incorrect behavioral responses. (A) Stimulus-locked population PETHs showing LC response to cues (presented at time 0) for trials yielding correct or incorrect behavioral responses. Note that the LC response rises sooner and is less prolonged on correct compared with incorrect trials in this analysis. No LC activation was detected on omission trials (orange line). Vertical dashed lines indicate the mean behavioral RTs. Curves below represent the normalized RT distributions for correct and incorrect trials. (B) The difference in the phasic LC response between correct and incorrect trials was not evident in response-locked population PETHs. In addition, no LC activation occurred prior to or following lever releases not associated with stimulus presentation (orange line). Dashed vertical lines indicate the mean stimulus onset times. Reproduced from Aston-Jones & Cohen (2005).

2.3.3.3 A tradeoff between tonic and phasic activity

In times of high tonic LC activity, phasic responses tend to be weaker. A state of high tonic and low phasic LC activity was shown to follow changes in reward contingencies, when the monkey does not yet recognize the rewarding stimulus and thus makes more errors (Aston-Jones et al., 1997). According to the gain theory of the LC-NE system, high tonic LC activity should facilitate behavioral responses in a *nonspecific* manner, which is what the monkey needs to do (and in fact does) to re-explore the task space when reward contingencies change. Thus, Aston-Jones & Cohen (2005) proposed that the balance between phasic and tonic modes of LC-NE activity is controlled so as to optimize the balance between task engagement and exploitation (low tonic, high phasic) on the one hand, and task disengagement and exploration (high tonic, low phasic) on the other.

2.3.3.4 Pupil diameter as an index of LC-NE activity

It is not possible to measure neural gain non-invasively in humans, nor is it possible to directly measure the LC activity thought to regulate it. However, converging evidence suggests that pupil diameter is closely correlated with LC-NE activity. LC responses are coupled with activity of the peripheral sympathetic nervous system, which is known to modulate pupil diameter (Svensson, 1987; Elam et al., 1981; Elam et al., 1984). Furthermore, a series of carefully controlled studies showed that adrenergic stimulation in the CNS leads to pupil dilation in rats and cats (Koss, 1986). Finally, pupil diameter was shown to strongly correlate with LC activity in the monkey (Figure 2.4; Rajkowski et al., 1993). Building on these findings,

recent studies have begun using pupil diameter to test predictions of the gain theory of the LC-NE system in humans (Aston-Jones & Cohen, 2005).

2.3.3.5 Pupillometry-based evidence in humans

Paralleling LC findings in monkeys, Gilzenrat et al. (2010) found that increases in baseline pupil diameter are associated with decreases in task utility and disengagement from the task, whereas reduced baseline diameter (but increases in phasic dilations) are associated with task engagement. Similarly, Jepma and Nieuwenhuis (2011) showed that pupil diameter predicts changes in the balance between exploration and exploitation. Closely related findings concerning the relationship between pupil diameter and behavior were also observed in humans performing an auditory oddball task (Murphy et al., 2011) or perceiving ambiguous stimuli (Einhäuser et al., 2008).

These results support the gain theory of the LC-NE system, and moreover, they suggest that gain can be assessed non-invasively in human participants using pupillometry. Specifically, baseline pupil diameter may track tonic LC activity, while task-evoked pupil responses may track phasic LC activity. Furthermore, while differences in baseline pupil diameter across individuals may reflect factors other than tonic LC-NE activity (e.g., physical pupil size), phasic pupil dilation responses can be normalized to the baseline diameter to obtain a between-participants measurement of neural gain: because phasic responses are inversely related to baseline pupil diameter, and thus, presumably, to tonic LC-NE activity, the average phasic response may provide an inverse measure of sustained levels of neural gain.

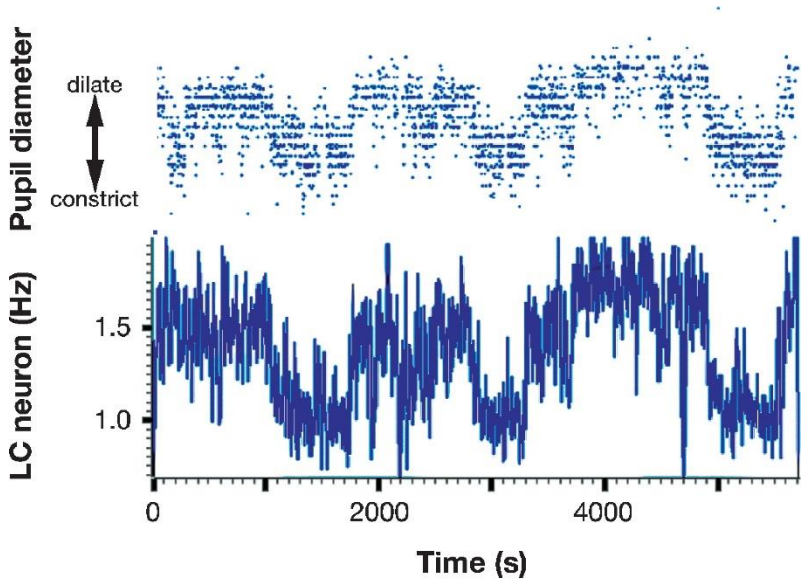

Figure 2.4. Relationship between tonic pupil diameter and baseline firing rate of an LC neuron in monkey. Pupil diameter measurements were taken by remote eye-tracking camera at each instant in time when the monkey achieved fixation of a visual spot during a signal-detection task. Note the close relationship between the pupil diameter and the rate of LC activity. Reproduced from Aston-Jones & Cohen (2005).

2.4 Conclusion

A wide range of pharmacological, electrophysiological, pupillometric and behavioral evidence suggests that the LC-NE system serves to modulate neural gain non-selectively throughout the brain. Brain-wide variations in gain may be tracked non-invasively in humans by measuring pupil diameter. Some of the observed effects of LC-NE activity on single neurons seemingly conflict with the neural gain theory, but many of these findings could in fact result from increased gain acting in a particular network architecture. More importantly, the gain theory is primarily concerned with high-level abstractions of brain-wide neural processes, rather than with the activity of single neurons, and thus, the theory should be primarily assessed by its ability to predict system-level neural and behavioral phenomena.

Chapter 3

The whole-brain effects of neural gain*

A wide range of anatomical, electrophysiological and behavioral evidence suggest that the locus coeruleus modulates neural gain throughout the brain. However, the brain-wide effects of gain modulation have yet to be studied. If brain-wide gain modulation indeed exists, it should affect neural function all over the brain, and thus, give rise to system-level neural effects that are evident when brain function is considered as a whole. Such effects may manifest throughout the brain in changes in neural activity. However, since gain modulation specifically relates to the enhancement of neural communication, its effects may be more clearly revealed by examining changes in functional connectivity. An effect on whole-brain neural function may manifest not only in the strength of functional connectivity, but also in the way that functional connections are distributed throughout the brain. In this chapter, I examine these three types of effects theoretically and experimentally, using whole-brain functional Magnetic Resonance Imaging

* Parts of this chapter appeared in Eldar, E., Cohen, J. D., & Niv, Y. The effects of neural gain on attention and learning. *Nature neuroscience* 16, 1146-1153 (2013), and were presented at SfN 2012, the Third Symposium on Biology of Decision Making 2013, and MathPsych 2013.

(fMRI) in conjunction with pupillary indices of neural gain, as well as in response to pharmacological manipulation.

3.1 Changes in activity

Increased gain entails that neural activation is driven toward maximal or minimal levels (Figure 1.2). Thus, large baseline pupil diameter should be associated with more extreme fMRI activations. Indeed, we found that the fMRI blood-oxygen-level-dependent (BOLD) signal, taken while participants were performing a learning task, was farther from its mean when baseline pupil diameter was large (mean absolute deviation from the mean 8.36) compared to when it was small (mean absolute deviation 8.02; $t_{29} = 3.79$, $p < 10^{-4}$, paired *t*-test comparing 10% of trials with highest pupil diameter to 10% of trials with lowest pupil diameter, out of a total of 216 trials per participant).

An additional prediction that stems from the hypothesized relationship between pupil diameter and gain is that the magnitude of pupil dilation in response to task-relevant stimuli should correlate with the magnitude of the BOLD response to task-relevant stimuli, but not to task-irrelevant stimuli (Aston-Jones & Cohen, 2005). This is so because the low tonic/high phasic state of LC-NE function is thought to specifically enhance the processing of task stimuli. To test this prediction, our experiment included random, task-irrelevant, auditory stimuli that participants were instructed to ignore. As predicted, both low baseline pupil diameter and high pupil dilation response were associated with stronger BOLD responses to task-relevant stimuli, but not to task-irrelevant stimuli (baseline diameter: $t_{27} = -5.04$, $p < 10^{-5}$; dilation response: $t_{27} = 2.56$, $p < 0.05$; Figure 3.1).

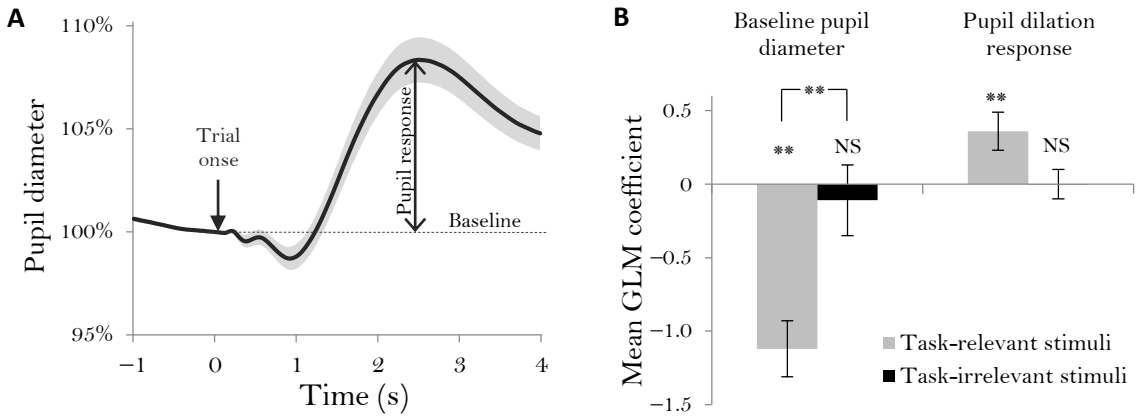

Figure 3.1. Relationship between pupil diameter and BOLD response to task-relevant and task-irrelevant stimuli. (A) Pupil diameter normalized by its value at trial onset (time 0), averaged within participants across trials, and then across participants (lighter shade: s.e.m. across participants; $n = 28$). Pupil dilation response was computed as the difference between the peak pupil diameter during the 4 s that followed trial onset and the pre-trial baseline diameter, normalized by the pre-experiment resting diameter. As expected, baseline pupil diameter and pupil response were anticorrelated in all participants (mean $r = -0.77$, range -0.89 to -0.54, $t_{27} = -28.9$, $p < 10^{-21}$). While baseline diameter is thought to correlate positively with tonic LC-NE function, the normalized pupil dilation response can serve as an inverse index that is comparable between individuals. (B) High baseline pupil dilation was associated with a weaker response to task-relevant stimuli compared to task-irrelevant stimuli, whereas high pupil dilation response was associated with a stronger response to task-relevant stimuli compared to task-irrelevant stimuli. $n = 28$, *: $p < 0.05$, **: $p < 10^{-4}$, errors bars: between subject s.e.m.

3.2 Changes in functional connectivity

Increased gain means that neural signals are enhanced, which predicts that interactions between connected parts of the network should increase. Indeed, gain modulation has been proposed as a mechanism for flexible control of network functional connectivity (Salinas, 2004; Salinas & Bentley, 2009; Haider & McCormick, 2009). We first tested this prediction theoretically by simulating the effect of gain on functional connectivity using randomly constructed neural network models, in which we varied the gain parameter and measured the resulting unit-to-unit correlations. The simulations suggested that global fluctuations in neural gain should be associated with global fluctuations in the mean strength of functional

connectivity (Figure 3.2A-C; $r = 0.99$, $p < 10^{-13}$). Different simulations, in which each unit was only connected to a minority of other units (10%), or in which correlations were measured between the mean activity of groups of 10 units, yielded qualitatively similar results, indicating that our simulation results are robust to network density and measurement granularity.

3.2.1 Global fluctuations in local functional connectivity

We thus examined the fMRI data for evidence of global fluctuations in the strength of functional connections. Towards this end, we measured functional connectivity while participants performed a learning task, which consisted of a series of games, and assessed the extent to which game-to-game fluctuations in functional connectivity in different brain areas were correlated with each other. To do this, we arbitrarily divided each participant's brain into 32 boxes that contained roughly similar volumes of gray matter (27.1 ±2.6 cm^3; Figure 3.3A). We then measured the mean strength of functional connectivity within each box during each game (quantified as the mean absolute correlation of the time series of the fMRI signal among pairs of voxels within the box). Finally, we correlated the time series of mean functional connectivity values over games for each pair of boxes. Mean functional connectivity strength across games was positively correlated for 96% of all box pairs, and the mean correlation coefficient was 0.72 (range 0.27 to 0.95 across participants, $t_{29} = 10.85$, $p < 10^{-10}$; Figure 3.3B). Notably, this correlation did not simply reflect a common global signal component, since the mean gray-matter signal was regressed out of the data prior to the functional connectivity analysis (see Section 3.6 for detailed methods). Thus, even though our measurements of functional connectivity in different boxes involved strictly disjoint brain areas, we found very strong correlations in fluctuations of these measurements throughout the brain.

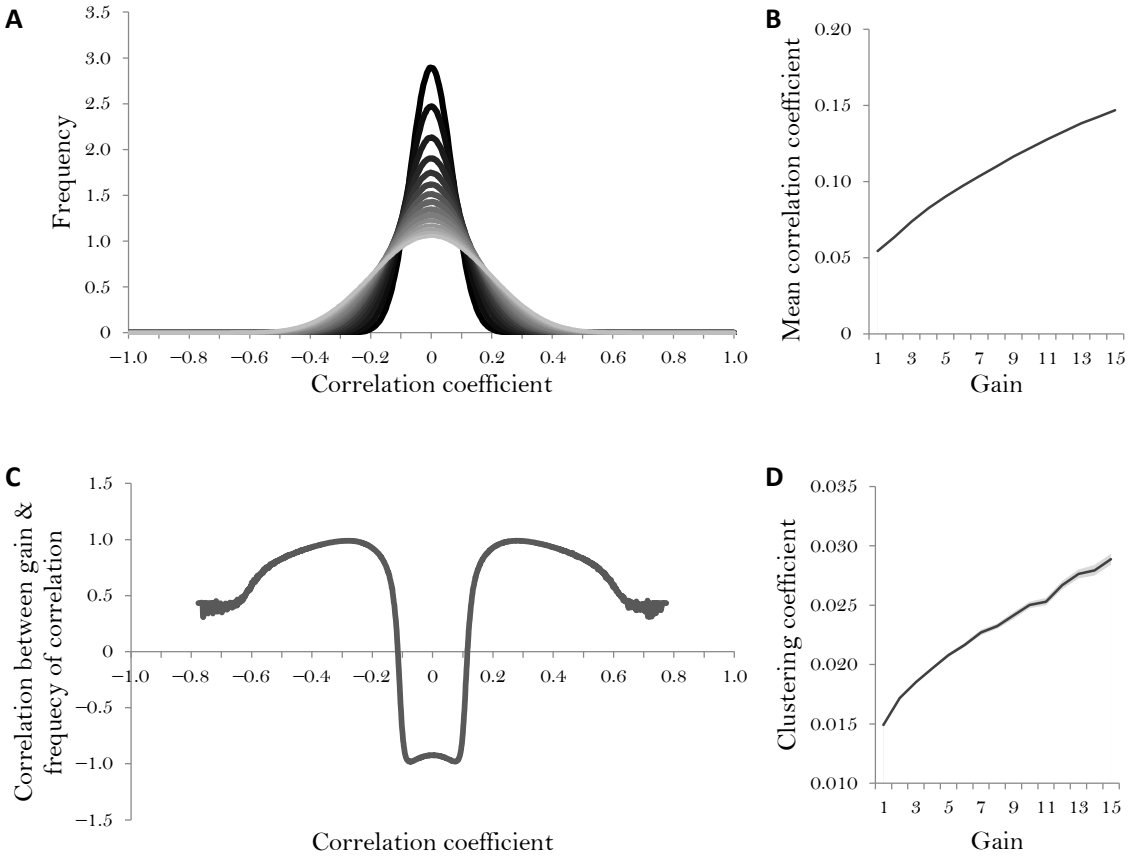

Figure 3.2. Simulation of the effect of global changes in gain on functional connectivity strength and clustering. Recurrent neural networks were composed of 1000 fully connected units with random connection weights. Unit-to-unit correlations were computed across 500 trials for each level of gain, for each of 100 networks. (A) Distribution of correlation coefficients for each of 15 different levels of gain. Higher gain results in stronger functional connections (correlations or anti-correlations). (B) Mean correlation coefficient increases as a function of gain. S.e.m. is too small to observe. (C) Correlation between the global gain parameter and the frequency of correlation coefficients as a function of correlation coefficient. Stronger correlations are more prevalent (and weaker correlations are less prevalent) when gain is higher. (D) Clustering coefficient of the networks' functional connectivity graphs as a function of gain. Clustering coefficient tended to increase with gain. Ligher shade: s.e.m.

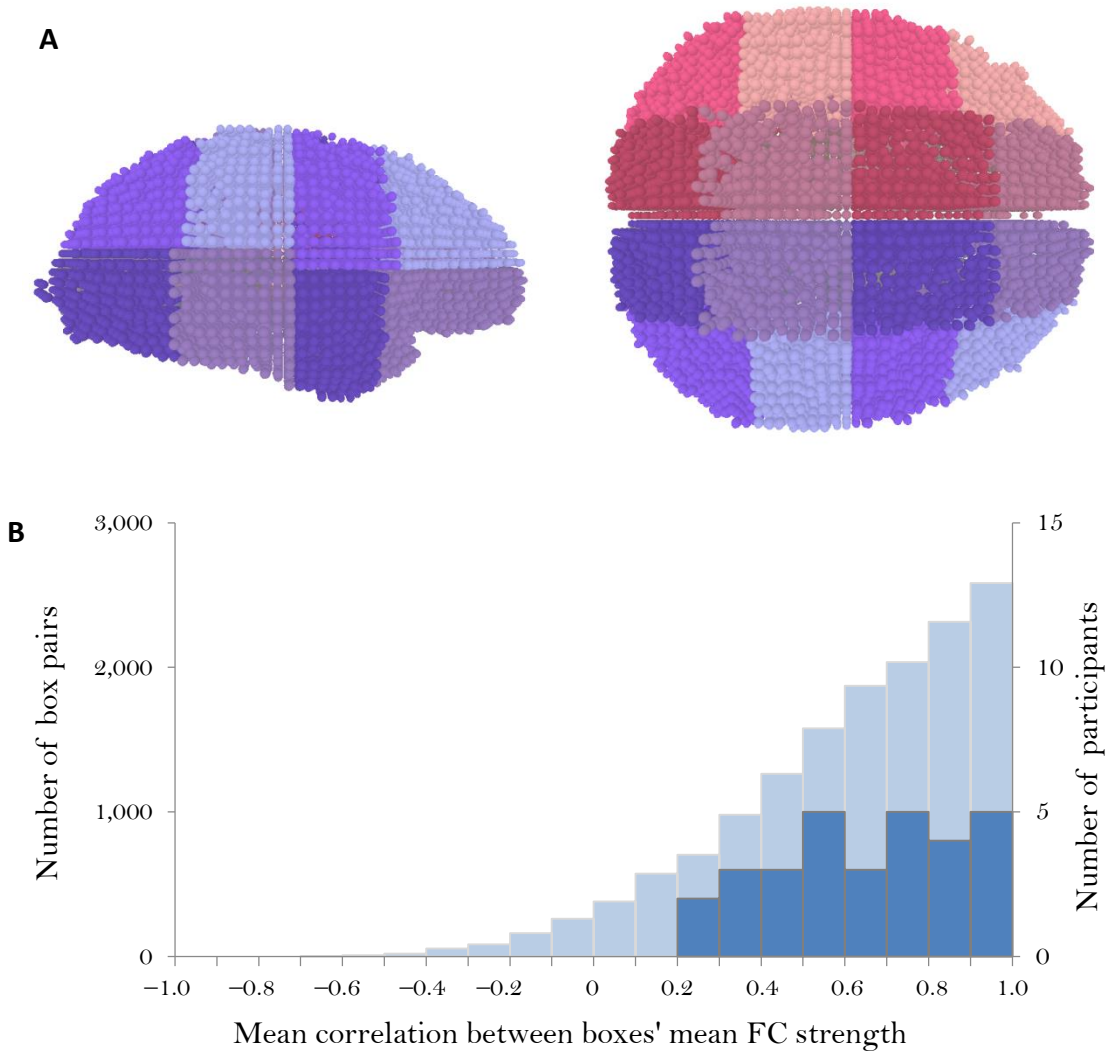

Figure 3.3. Global fluctuations in local functional connectivity. (A) 3D rendering of one participant's gray-matter voxels divided into 32 boxes, viewed from the right and from above. Each sphere represents a voxel. Adjacent boxes are denoted in different colors. Voxel division is visualized using custom-made software created in the Processing programming environment (Reas & Fry, 2007). (B) Histogram of between-box correlations of mean within-box functional connectivity strength (light blue, left Y axis), and of participants' mean correlation values (dark blue, right Y axis).

3.2.2 Functional connectivity and pupil diameter

While this result is strongly suggestive of global modulation of neural signaling, it is nevertheless possible that global fluctuations in local functional connectivity reflect correlated local instabilities of the MRI scanner. Such a confound could be dismissed if the measured fluctuations in functional connectivity were also to covary with the separately attained measures of baseline pupil diameter. Indeed, we found that when baseline pupil diameter was highest (indicative of high gain), high functional connectivity measurements were more prevalent (Figure 3.4A, gray), whereas when baseline pupil diameter was lowest (low gain) weaker functional connectivity was more prevalent (Figure 3.4A, black). Accordingly, baseline pupil diameter was positively correlated with mean functional connectivity strength (mean $r = 0.27$ across participants, $t_{27} = 2.98$, $p < 0.01$), and, similarly, pupil responses were anticorrelated with mean functional connectivity strength (mean $r = -0.24$ across participants, $t_{27} = -3.63$, $p < 0.01$). In particular, baseline diameter was positively correlated with the number of functional connectivity measurements stronger than ± 0.17, and anticorrelated with the number of weaker functional connectivity measurements (Figure 3.4B). Remarkably, this non-monotonic relationship between functional connectivity strength and its correlations with baseline diameter was predicted by our simulation of the effects of gain on functional connectivity in randomly connected neural networks (*cf.* Figure 3.2B).

To verify that the relationship between functional connectivity and pupil diameter was not specific to particular brain regions, but rather was manifest throughout the brain, we examined this relationship separately in each of the 32 boxes. Functional connectivity strength was positively correlated with baseline pupil diameter in 70% of the boxes (22 out of 32 boxes per

participant on average; Figure 3.5A), and the mean correlation coefficient was 0.19 ($t_{27} = 3.19$, $p < 0.01$; Figure 3.5B). Notably, baseline pupil diameter was not correlated with the mean box fMRI signal (mean $r = -0.004$, $t_{27} = -0.35$, $p = 0.73$), indicating that the relationship with functional connectivity strength did not reflect pupil-related variations in signal strength. Further, the relationship between baseline diameter and functional connectivity strength was fairly consistent throughout the brain: for every box, functional connectivity was positively correlated with baseline pupil diameter in at least half of the participants. As expected, functional connectivity strength was also anticorrelated with pupil dilation response in 75% of the boxes (Figure 3.5C), and the mean correlation coefficient was –0.20 ($t_{27} = -3.54$, $p < 0.01$; Figure 3.5D). Our results thus suggest that the strength of functional connectivity fluctuated in a similar manner throughout the brain, and these fluctuations were tracked closely by both pupil diameter indices.

3.2.3 The distribution of functional connections

Thus far, we provided evidence of brain-wide fluctuations in gain, as manifested in fMRI activity and connectivity, which reflects the enhancement of neural signals. We now turn to a more specific effect of gain modulation that is central to the theme of this thesis. Recall that we proposed that increasing gain has an effect that is akin to that of increasing contrast, namely the effect of focusing information processing on a more limited set of features (Figure 1.1). Accordingly, we hypothesized that high gain would be associated with a more tightly clustered pattern of neural interactions, reflecting selective processing of particular input streams. In contrast, we expected low gain to be associated with widely distributed neural interactions, which mediate concurrent processing of multiple input features.

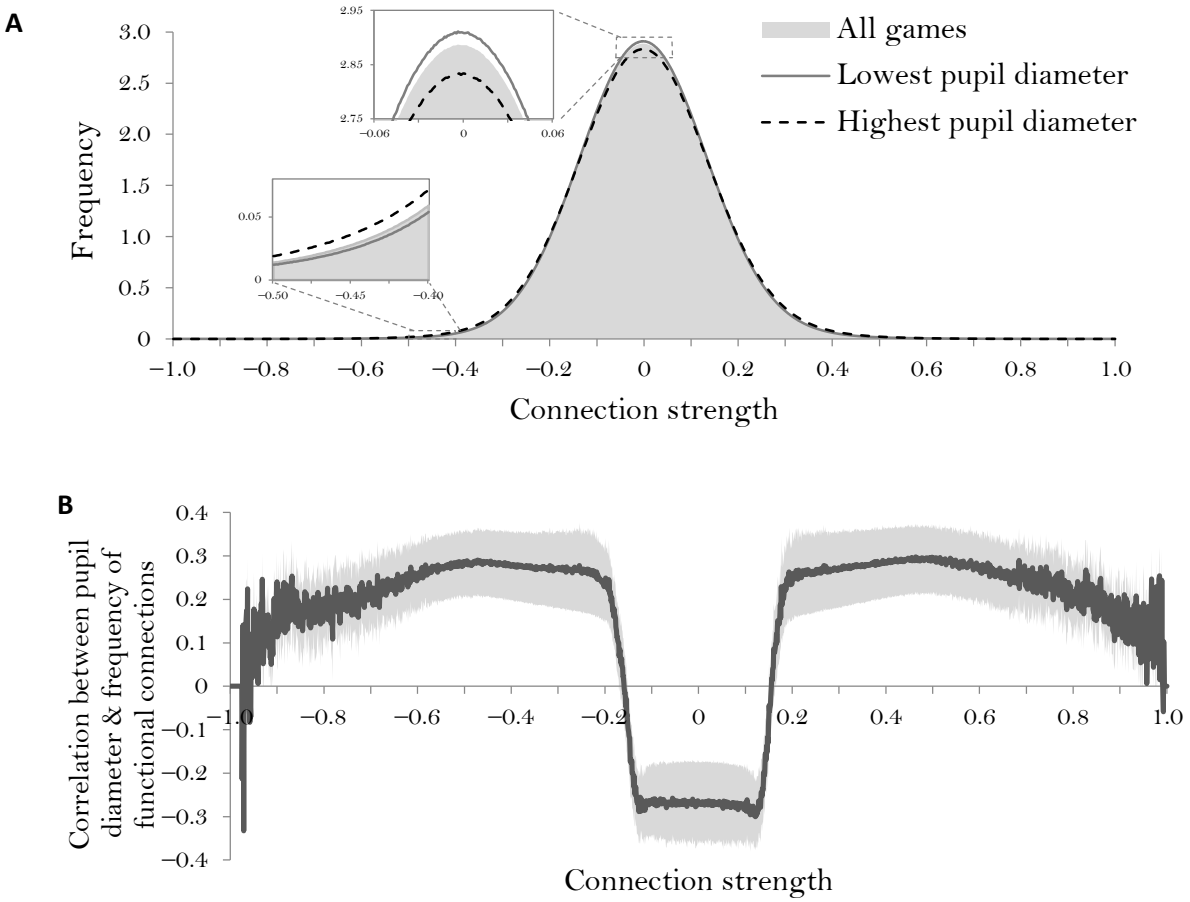

Figure 3.4. Pupil diameter and whole-brain functional connectivity. (A) Distribution of functional connections by connection strength (n = 28). The distribution is shown separately for all games (gray shading), for the third of each participant's games in which the participant's baseline pupil diameter was lowest (solid line), and for the third of games in which pupil diameter was highest (dashed line). Insets: magnification of boxed areas to show differences between lowest and highest pupil diameter games. (B) Game-by-game correlation between baseline pupil diameter and frequency of functional connectivity measurements as a function of functional connectivity value (n = 28). The Y axis indicates whether large pupil diameter was associated with more (positive values) or fewer (negative values) voxel pairs for which functional connectivity strength is indicated on the X axis. For each participant, we computed the distribution of functional connections during each game, and then computed the correlation across games between baseline pupil diameter and the number of voxel pairs in each bin of the distribution. The curve shows the correlations averaged over participants with s.e.m. indicated by the lighter shading. Larger pupil diameter was associated with more strong functional connectivity measurements (absolute strength > 0.17) and fewer weak functional connectivity measurements (between –0.17 and +0.17).

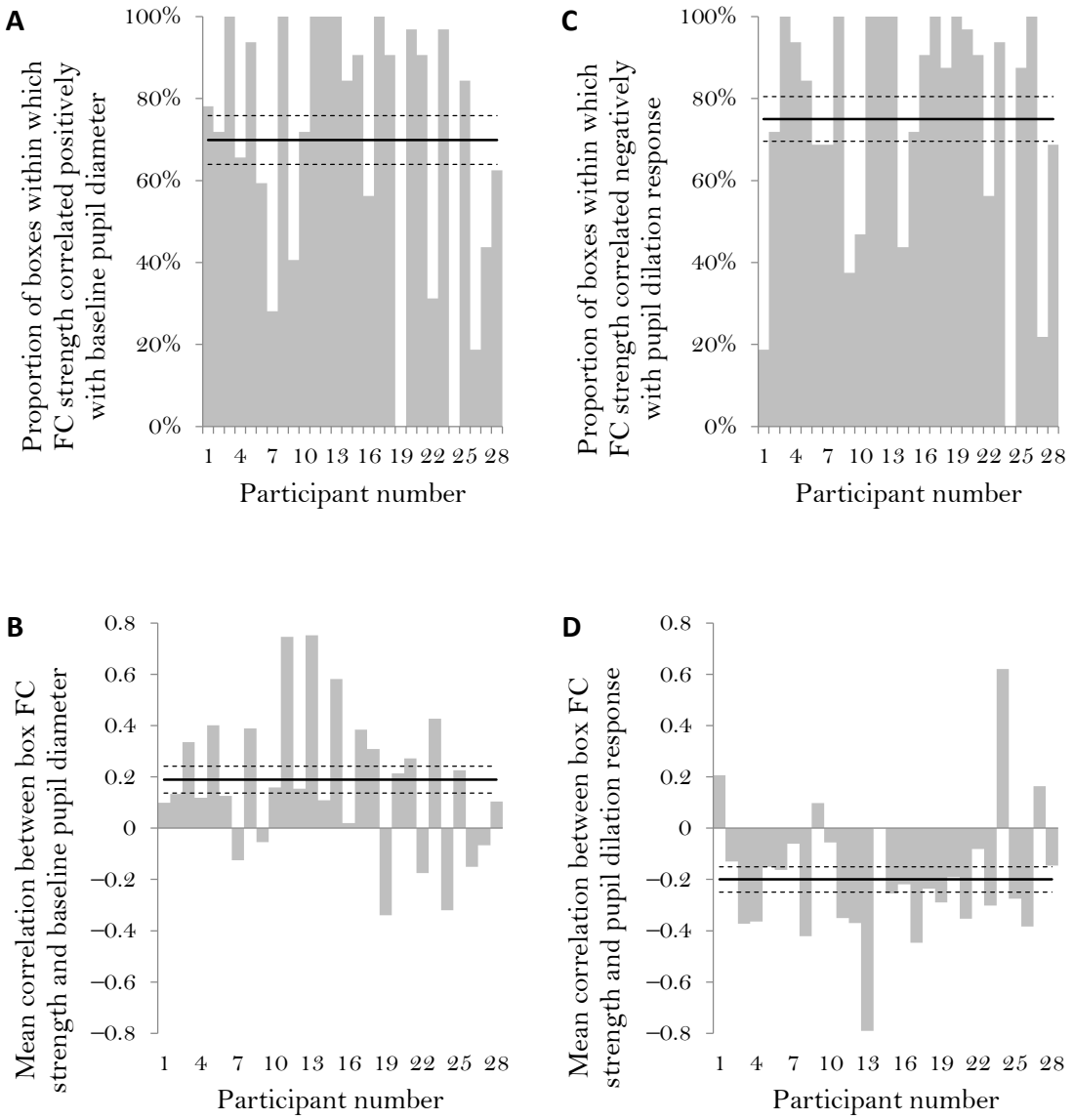

Figure 3.5. Pupil diameter and local functional connectivity. (A, C) Proportion of boxes within which mean functional connectivity strength was positively correlated with baseline pupil diameter (A) or negatively correlated with pupil dilation response (C) for each participant. (B, D) Mean correlation between within-box functional connectivity strength and baseline pupil diameter (B) or pupil dilation response (D) for each participant. Solid horizontal line: group means, dashed horizontal lines: s.e.m.

To examine this idea theoretically, we constructed a functional connectivity graph (Eguíluz et al., 2005) for each of the random 1000-unit networks described above. Each of the graphs' nodes represented a unit and two units were connected by an edge if the correlation of activity between them was in the top 1% of all such correlations. The clustering coefficient (Luce & Perry, 1949) of such a graph indicates the degree to which functional connectivity is tightly clustered in the network. As predicted, higher gain was associated with higher clustering coefficients (Figure 3.2D; $r = 0.99$, $p < 10^{-12}$).

To test the degree to which functional connectivity was tightly clustered in the brain, we similarly constructed a functional connectivity graph for each participant and each game (18 graphs per participant). The graphs were constructed as done for the simulated networks, except that in this case each of the graphs' nodes represented a voxel, and connectivity was determined by the correlation between voxels' time series (see Figure 3.6 A and B for example graphs). As predicted, we found a significant game-by-game correlation between the clustering coefficient of these graphs and baseline pupil diameter (mean $r = 0.14$ across participants, $t_{27} = 1.82$, $p < 0.05$ one tailed; Figure 3.6C). That is, when participants' pupil diameter indicated high gain, their neural functional connectivity tended to be more tightly clustered. Moreover, we found a similar correlation when the analysis was restricted to prefrontal cortex, an area that is not involved in primary visual processing (mean $r = 0.14$ across participants, $t_{27} = 2.05$, $p < 0.05$), suggesting that the relationship between pupil diameter and clustering was indeed due to global fluctuations in gain and not due to differences in activation to the visual stimuli.

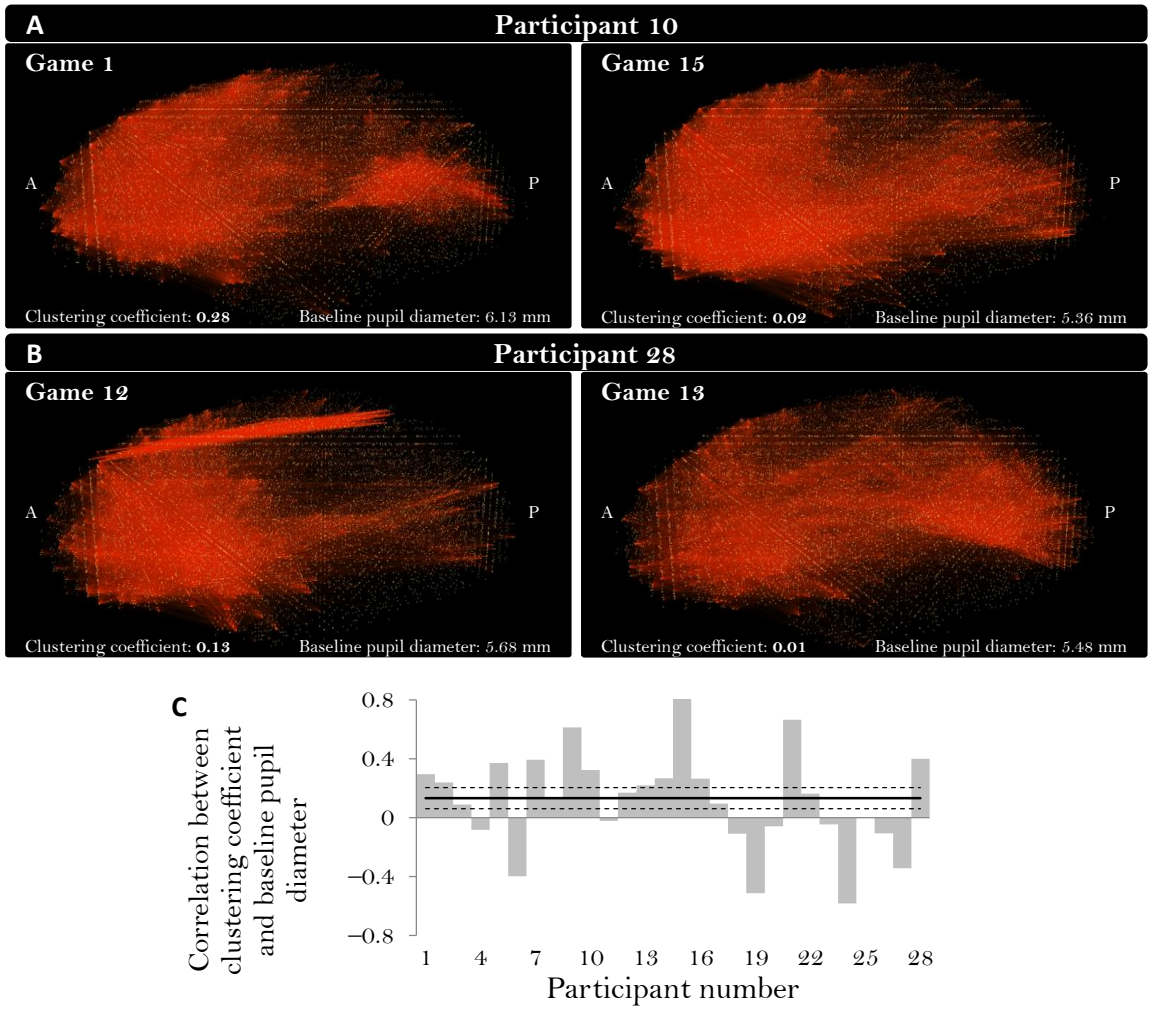

Figure 3.6. The distribution of functional connection. (A) Connectivity graphs from participant 10 in two different games. This participant's baseline pupil diameter was highest in game 1 and lowest in game 15. In line with our hypothesis, the clustering coefficient was higher in game 1 (0.28) than in game 15 (0.02). As can be seen, connectivity formed two disparate clusters in game 1, whereas in game 15 it was more globally distributed. (B) Connectivity graphs from participant 28 in two different games. The clustering coefficient was highest for this participant in game 12 and lowest in game 13. As shown, functional connections were mostly clustered in fontal cortex in game 12, whereas in game 13 they were distributed over most of the brain. Indeed, the participant's baseline pupil diameter was higher in game 12 (5.68) than in game 13 (5.48). A – anterior, P – posterior. For the purpose of visual rendering only, connectivity graphs were compressed to a size of 10,000 vertices using a k-means clustering algorithm, merging together vertices whose voxels' MNI coordinates are closest. The correlation values of merged vertices were averaged, and the strongest 0.05% of the resulting correlations were displayed as edges. (C) Game-by-game correlation between clustering coefficient and baseline pupil diameter by participant. $n = 28$.

3.3 Pharmacological manipulation

In studying the effects of LC-NE function and neural gain, pupil diameter measurements only provide correlational evidence. Causal evidence, however, can be obtained using pharmacological manipulation. Thus, we next examined variations in functional connectivity strength and clustering in response to the norepinephrine-specific reuptake inhibitor reboxetine.

Reboxetine has been in use for the treatment of depression, anxiety and attention deficit hyperactivity disorder (Hajós et al., 2004). Its binding affinity is highly selective to the norepinephrine transporter, and both acute and chronic administration of the drug have been shown to raise extracellular levels of norepinephrine in frontal cortex and hippocampus (Hajós et al., 2004; Millan et al., 2001; Sacchetti et al., 1999). However, reboxetine does suppress LC-NE activity, and thus reduces physiologic norepinephrine function (Szabo and Blier, 2001). Still, since reboxetine increases cortical extracellular norepinephrine levels, we predicted that its administration would be associated with signs of increased gain in fMRI – that is, with stronger and more tightly clustered functional connectivity networks.

To test this prediction, we analyzed a pharmacological fMRI data set that was shared with us by Andrea Reinecke and Catherine Harmer from the University of Oxford (Papadatou-Pastou et al., 2012). Half of the participants received reboxetine, and half received placebo, 2 hours before performing an autobiographical memory task in the MRI scanner for a period of 9 minutes. In analyzing the fMRI data, we used the same methods as in the pupillometry study. That is, absolute functional connectivity was measured throughout the brain, and graph-

theoretic analysis was used to compute for each participant the degree to which functional connectivity was clustered.

3.3.1 Functional connectivity strength and clustering

Salivary cortisol levels, which are indicative of the efficacy of reboxetine, were similar in the two study groups at baseline (reboxetine group: 15.71 ±1.80 mmol/L; placebo group: 15.08 ±1.75 mmol/L), but higher in the reboxetine group before entering the scanner (reboxetine group: 17.48 ±1.93 mmol/L; placebo group: 11.57 mmol/L ±1.36 mmol/L) and at the end of the study (reboxetine group: 14.85 ±1.53 mmol/L; placebo group: 9.29 mmol/L ±1.09 mmol/L; $F_{1,20} = 6.62$, $p < 0.05$, ANOVA).

In contrast to our prediction, mean whole-brain functional connectivity strength was *lower* in the reboxetine group compared to the placebo group ($t_{20} = -2.2$, $p < 0.05$; Figure 3.7A, left). Furthermore, graph-theoretic analysis showed that functional connections were less clustered in the reboxetine groups ($t_{20} = -3.1$, $p < 0.01$; Figure 3.7A, right) as compared to controls. Both of these results are consistent with lower rather than higher neural gain in response to reboxetine.

Functional connectivity and clustering coefficient were strongly correlated across participants ($r = 0.65$, $t_{20} = 3.8$, $p < 0.005$; Figure 3.7B), suggesting that low clustering in the reboxetine group might have simply reflected weaker functional connectivity. However, clustering was still lower in the reboxetine group after the effect of mean functional connectivity was regressed out ($t_{20} = -1.82$, $p < 0.05$ one tailed). The same result was found when clustering was compared between the reboxetine group and those participants in the placebo group whose

mean functional connectivity was in the same range (functional connectivity < 0.07; $t_{15} = -2.06$, $p < 0.05$ one tailed). Thus, the effect of reboxetine on the clustering of functional connections was only partially predicted by its effect on mean functional connectivity.

Notably, reboxetine administration seemed to restrict mean functional connectivity and clustering coefficients to a specific range (0.06 to 0.07 for the former, 0.004 to 0.012 for the latter), as compared to the more varied measurements in the placebo group (Figure 3.7B). To gain further insight into these results, we examined the number of strong connections made by each voxel (i.e. its cardinality) in participants' functional connectivity networks. In participants with strong mean functional connectivity (FC > 0.07), all of which belonged to the placebo group, connectivity mostly involved relatively few densely connected voxels (cardinality > 500), while most voxels were sparsely connected (cardinality < 100; variance of cardinality: 17026 ±2024; Figure 3.7C). Densely connected voxels were not limited to particular brain regions, but rather, were scattered throughout the brain (Figure 3.7D). In contrast, in participants from both study groups whose mean functional connectivity was weaker (FC < 0.07), connectivity involved a large number of voxels with an intermediate number of connections (100 to 500; variance of cardinality: 7588 ±754; difference from high FC participants: $t_{20} = 4.9$, $p < 10^{-5}$). Seemingly, without pharmacological intervention (i.e., in the placebo group) participants functioned in one of two modes: either few voxels were significantly involved in neural communication, or almost all voxels were. In contrast, only the latter mode was evident in participants who received reboxetine suggesting that reboxetine restricted whole-brain neural communication to this specific mode.

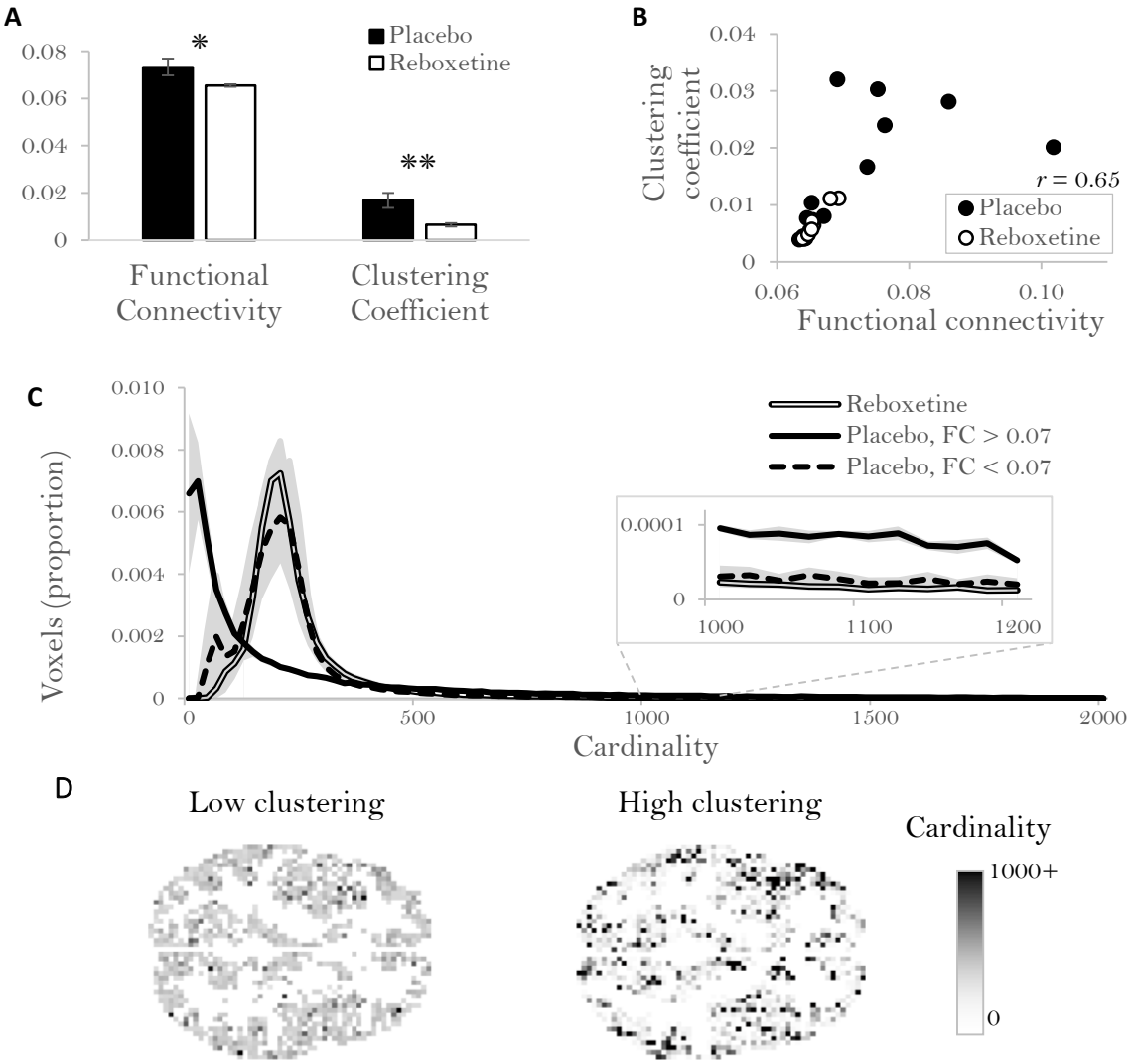

Figure 3.7. Functional connectivity strength and clustering in response to reboxetine. (A) Mean whole-brain unsigned functional connectivity and clustering coefficient for the placebo (n = 11) and reboxetine (n = 11) study groups. *: $p < 0.05$, *: $p < 0.01$, error bars: s.e.m. (B) Clustering coefficient as a function of mean functional connectivity for the placebo and reboxetine study groups. (C) Distribution of voxels by the number of graph connections that each voxel makes (i.e., cardinality). The data are shown separately for the reboxetine study group (n = 11), for participants of the placebo group for whom mean functional connectivity was in the same range as in the reboxetine group (< 0.07, n = 6), and for participants of the placebo group for whom functional connectivity was stronger (> 0.07, n = 5). Inset: magnification of marked area to show differences between groups in the high cardinality range. Shaded area: s.e.m. (D) Exemplary horizontal brain images (MNI Z = 0), showing the spatial arrangement of high and low cardinality voxels in participants with low (0.007, reboxetine group) and high (0.03, placebo group) clustering coefficients. Note that cardinality is less uniform in the high clustering image, and that high cardinality voxels are distributed throughout the brain.

3.3.2 Signs of decreased gain?

While weaker and more distributed functional connectivity is suggestive of low gain, it can also result from an *increase* in norepinephrine and gain that exceeds physiological levels. This can be demonstrated in our neural network models: when gain was increased in the model to such a degree that, on average, half of the model's units were at saturation levels of activity, unit-to-unit correlations started decreasing with increasing gain (Figure 3.8A). However, this decrease in correlations was not coupled with a decrease in the clustering coefficient as was the case in our fMRI data. Nevertheless, clustering coefficients are sensitive to the underlying structure of the network, prompting us to examine these effects in a more structured network. Our simulations showed that in a network composed of multiple groups of units with strong within-group connections and weak between-group connections the clustering coefficient did drop with the weakening of correlations at high levels of gain (Figure 3.8B).

To determine whether weaker functional connectivity in the reboxetine group resulted from low gain or from exceedingly high levels of gain, we thus examined the absolute mean-corrected blood-oxygen-level dependent (BOLD) signal. By definition, increased gain should always drive activations towards maximal or minimal levels (Figure 1.2), regardless of network structure. Thus, provided that activations are mean corrected, absolute activation levels should increase monotonically with gain (Figure 3.8, mean activation level). Indeed, we have seen previously that pupillary indices of high norepinephrine are associated with a higher mean BOLD signal (see section 3.1). Here, however, mean BOLD signal was similar in the reboxetine group and in participants from the placebo group with similarly low functional connectivity levels (FC < 0.07; $t_{15} = 0.3$, $p = 0.73$), but higher in participants from the placebo group that

had stronger functional connectivity (FC > 0.07; $t_{14} = 2.3$, $p < 0.05$; Figure 3.9). This result indicates that weaker functional connectivity in the reboxetine group was most likely a result of low, not high gain.

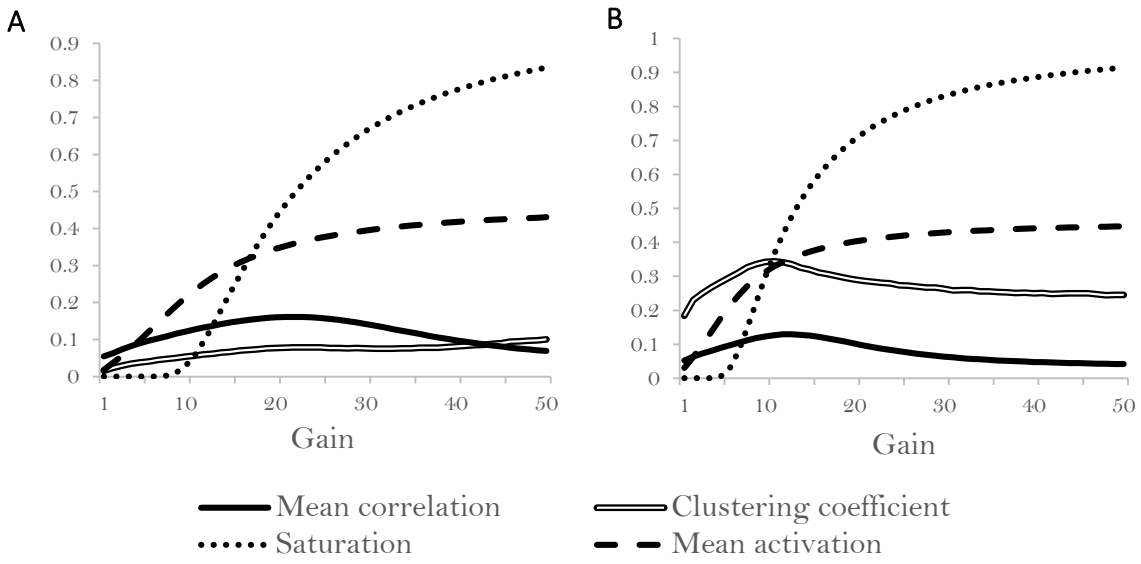

Figure 3.8. Neural network simulations of the effect of gain on activations and correlations. Mean unsigned unit-to-unit correlation, clustering coefficient, saturation level, and mean unsigned activation are shown as a function of gain. Saturation level reflects the proportion of units whose activation is close to maximal (>0.95) or minimal (<0.05) levels. 500 trials were conducted for each of 100 randomly constructed networks. Error bars are too small to observe. (A) Results of a random 1000-unit model. All weights were drawn from the same uniform distribution. (B) Results of a "structured" model, consisting of ten 100-unit groups, with strong weights between units of the same group, and weak weights between units of different groups.

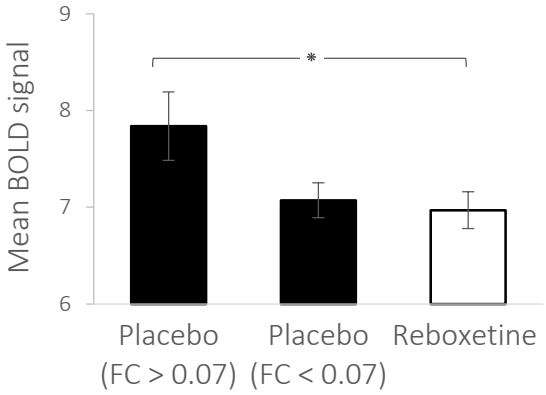

Figure 3.9. Mean absolute mean-corrected BOLD signal. Data are shown separately for the reboxetine study group ($n = 11$), for participants of the placebo group for whom mean functional connectivity was in the same range as in the reboxetine group (< 0.07, $n = 6$), and for participants of the placebo group for whom functional connectivity was stronger (> 0.07, $n = 5$). Weaker functional connectivity was associated with lower mean activation levels, which is suggestive of lower gain. *: $p < 0.05$.

3.4 Discussion

We investigated the existence of brain-wide fluctuations in neural gain using pupillometry, fMRI and a pharmacological manipulation. Pupil diameter indices of LC-NE function were associated with changes in the variance of the BOLD signal, and in BOLD response to task stimuli, that are predicted by changes in gain. In addition, fMRI data were characterized by global fluctuations in the strength of functional connectivity, as would be expected from global modulation of gain, and these fluctuations were tracked by pupillary indices of LC-NE function. Moreover, these pupillary indices were correlated with the degree to which functional connectivity was clustered, as predicted by our neural network modeling. These findings support existing theory that implicates the LC-NE system in global modulation of neural gain (Servan-Schreiber et al., 1990; Aston-Jones & Cohen, 2005). Additionally, our clustering analysis extends this theory by suggesting that high gain is associated with a shift from a widely distributed pattern of neural processing to a more tightly clustered pattern, which potentially mediates more selectively focused processing. The behavioral consequences of this shift in the mode of processing will be the topic of the remaining chapters.

In contrast to our predictions, functional connectivity strength and clustering decreased in response to reboxetine, a NE reuptake inhibitor. These results are surprising since NE is thought to increase, not decrease gain. Although these results could potentially be explained by the chaotic dynamics that may characterize unusually high levels of gain, weaker functional connectivity in our data was associated with a lower mean BOLD signal, making such an interpretation unlikely. Thus, our findings suggest that neural gain is decreased in response to reboxetine.

How can norepinephrine reuptake inhibition, presumably accompanied by an increase in levels of norepinephrine, result in a decrease in gain? First, reuptake inhibition does not only increase a neurotransmitter's extracellular level, but also changes its spatial profile. For example, inhibition of dopamine reuptake increases extrasynaptic dopamine levels, but not its intrasynaptic levels (Cragg and Rice, 2004). A similar effect has been suggested to follow norepinephrine reuptake inhibition (Bönisch and Brüss, 2006). A shift in the balance between intrasynaptic and extrasynaptic norepinephrine transmission may be further facilitated by the suppressive effect of reboxetine on LC-NE activity (Szabo and Blier, 2001). In addition, differences in dose, spatial and temporal profile of norepinephrine transmission may also change the type of adrenergic receptor that is primarily activated by norepinephrine.

Indeed, further investigation of the neuromodulatory effect of NE indicated that adrenergic stimulation may in fact suppress response to glutamate (i.e., reduce neural gain) when applied in high doses or if β receptors are selectively stimulated (Devilbiss & Waterhouse, 2000). It has thus been suggested that NE increases neural gain at low doses, but that further increases in NE can decrease gain. Since our pupil diameter study progressed at a slow pace (inter-trial intervals were 6 to 10 s long), it is possible that participants were at a state of low arousal, and

thus had low LC-NE activity. Therefore, in sum, our findings seem to suggest an inverted U-shaped relationship between NE and gain, echoing recent single-cell electrophysiology findings. In general, though, any one of the aforementioned differences between physiological NE function and the effect of reuptake inhibition may explain why the effect of reboxetine in our data was consistent with lower, not higher gain.

3.5 Conclusion

The present findings suggest that brain function is characterized by global fluctuations in neural gain, which are modulated by the LC-NE system, and that increased gain is associated with a shift from a widely-distributed to a tightly clustered pattern of neural interactions. Response to pharmacological manipulation indicates that gain may increase in response to NE at low NE levels, but decrease in response to NE at high NE levels. It is unclear, however, whether a decrease in gain would be observed in response to physiological NE activity as well.

3.6 Appendix: Methods

3.6.1 Pupil diameter study

3.6.1.1 Participants

36 participants (mean age 25.1, age range 18-61, 22 females) performed the behavioral experiment and 35 participants (mean age 20.5, age range 18-30, 25 females) performed the fMRI experiment. Participants were from the Princeton University area, and gave written informed consent before taking part in the study, which was approved by the university's

institutional review board. Participants received monetary compensation for their time, as well as a bonus according to their performance (4 cents per reward point, $8.04–$10.72, mean $9.47).

3.6.1.2 Stimuli

To minimize luminance-related changes in pupil diameter, all stimuli (which were either words or images) were made isoluminant with the background, to best approximation. Word colors were adjusted to be isoluminant using the flicker-fusion procedure (Lambert et al., 2003) on the display systems that were used in the experiments. More complex images, which consisted of many colors, were adjusted by scaling all colors so as to equate the mean estimated luminance with the background. For this purpose, luminance of each color was estimated based on its RGB values as 0.2126·R + 0.7152·G + 0.0722·B (http://www.w3.org/Graphics/Color/sRGB). The mean deviation of luminance within images was 29% (range 0% to 76%). Since within-image variance and deviation of the display system from the sRGB standard might cause slight differences in luminance perception, all of the analyses based on pupil dilation response were repeated using pupil responses to word stimuli only, which did not suffer from these sources of variance. The results of these analyses were similar to those reported here, and are thus not reported for the sake of brevity.

Stimuli were presented using MATLAB software (MathWorks) and the Psychophysics Toolbox (Brainard, 1997) using a projector outside the MRI scanner that displayed the stimuli onto a translucent screen located at the end of the scanner bore (fMRI experiment), which participants viewed through a mirror attached to the head coil. To compare BOLD responses to task-relevant and task-irrelevant stimuli, 72 task-irrelevant auditory stimuli (phonemes), which participants were instructed to ignore, were played at random times during the inter-trial

intervals in the fMRI experiment (4 stimuli per game). The phonemes were obtained from http://www.wikipedia.org and were at most 1 s long.

3.6.1.3 Behavioral task

The task consisted of 18 games, in which participants chose between pairs of stimuli and received monetary reward according to their choices. Each game included 12 trials, each of which took up to 5 seconds. Inter-trial interval was varied randomly (uniformly) between 6 s and 10 s. The task is described in detail in Chapter 4.

3.6.1.4 Eye tracking

An ASL Long Range Optics unit (Applied Science Laboratories, MA) was used to measure pupil diameter at a rate of 60 samples per second during the fMRI experiment. Pupil diameter data were processed in MATLAB to detect and remove blinks and other artifacts. Specifically, samples below 66.67% or above 150% of each block's median sample, or differing by more than 10% from the previous sample, were labeled as artifactual. Then, based on an examination of the artifact-triggered average, we labeled as artifactual, in addition, 6 samples before and 2 samples after every sample already labeled as artifactual. All artifactual samples were replaced using linear interpolation. At the beginning of the experiment, a measurement of pupil diameter at rest was taken for a period of 45 s. All subsequent pupil dilation responses were normalized by the pre-experiment resting pupil diameter. For each trial, baseline pupil diameter was computed as the average diameter over a period of 1 s prior to the beginning of the trial (at the end of the inter-trial interval, at which point pupil activity from the trial itself should have subsided). Pupil dilation response was computed as the difference between the peak diameter

recorded during the 4 s following trial onset and the preceding baseline diameter (Figure 3.1A). Baseline pupil diameter and dilation response measurements in which more than half of the samples contained artifacts were considered invalid and excluded from the analysis. Only participants with at least 30 valid trials were included in the across-participant analysis of mean pupil dilation (n = 30). Only participants for whom at least 6 games included 6 valid trials each were included in the game-by-game analysis of baseline pupil diameter (n = 28).

3.6.1.5 fMRI Data Acquisition

Functional (EPI sequence; 34 slices covering whole cerebrum; resolution = 3 × 3 × 3 mm with 1-mm gap; repetition time (TR) = 2.0 s; echo time (TE) = 30 ms; flip angle = 90°) and anatomical (MPRAGE sequence; 256 matrix; TR, 2.5 s; TE, 4.38 ms; flip angle, 8°; 1 × 1 × 1 mm resolution) images were acquired using a 3T Allegra MRI scanner (Siemens, Erlangen, Germany).

3.6.2 Pharmacological manipulation study

3.6.2.1 Participants

Twenty-two healthy volunteers participated in the study (age range: 23–38 years). All participants were free of medication, apart from contraceptive pills. Exclusion criteria included a current or previous history of psychiatric disorder (assessed with the *Structured Clinical Interview for DSM: Clinical Version* [SCID-CV]; Frances et al., 1995), substance abuse, and serious physical and neurological problems. Participants who reported any current use of illicit drugs were excluded. Participants gave written informed consent and were reimbursed for

their time and traveling expenses. The study was undertaken with ethics approval granted by the Oxfordshire Psychiatric Research Ethics Committee.

3.6.2.2 Procedure

The study followed a between-groups, double-blind, randomized design with two groups: a reboxetine group, and a placebo group. The reboxetine group received a single 4mg oral dose of the drug. The placebo group received a matched placebo capsule. The two groups were matched with respect to age (reboxetine group: 28.1 ±3.0 years, placebo group: 26.5 ±4.5 years). Participants attended the hospital having fasted for 3 hours prior to and during study participation to ensure similar rates and levels of reboxetine absorption. They were briefed on scanner safety and gave written consent before the study commenced. In order to confirm the absorption of reboxetine at the time of testing, salivary cortisol, which is indicative of central norepinephrine levels, was measured at baseline, before entering the scanner, and at the end of the study, using an in-house double antibody radioimmunoassay (intra- and inter-assay coefficients of variation were 3% and 10%, respectively; lower limit of detection was 0.5 mmol/L). Previous work has shown that levels of salivary cortisol peak approximately 2 hr after the administration of reboxetine and remain elevated for at least 2 hr (Hill et al., 2003). Testing therefore began 2 hrs after administration of the drug.

3.6.2.3 Behavioral Task

Participants performed an autobiographic memory retrieval task, in which they recalled specific memories in response to different word cues, for a total time of 9 min. All participants were debriefed after exiting the scanner and were asked to report their memories in order to ensure

that they had performed the task as instructed. Further details about the task, as well as results of task-based analyses of the fMRI data are reported elsewhere (Papadatou-Pastou et al., 2012).

3.6.2.4 fMRI Data Acquisition

Imaging was performed at the University of Oxford Centre for Clinical Magnetic Resonance Research Unit, at the John Radcliffe Hospital in Oxford, using a whole body 1.5-T scanner (Siemens Sonata Medical Systems) with a standard quadrature birdcage head coil. Structural images were acquired with a 3-dimensional T1-weighted FLASH sequence (repetition time [TR] = 12 ms, echo time [TE] = 5.6 ms, flip angle = 19°, 1-mm isotropic voxels, matrix = 256 × 160 × 208; elliptical sampling, orientation = coronal, acquisition time = 5 m 14 s). Functional images were acquired with a T2*-weighted echoplanar imaging sequence (TR = 3 s, TE = 50 ms, 32 slices, matrix = 64 × 64, 3 mm^3 isotopic voxels, 180 volumes per participant).

3.6.3 fMRI data processing

3.6.3.1 fMRI Data Preprocessing

Data were processed using MATLAB and SPM8 (Wellcome Trust Centre for Neuroimaging, UCL). Functional data were motion corrected, and low-frequency drifts were removed with a temporal high-pass filter (cutoff of 0.0078 Hz). Images were normalized to Montreal Neurological Institute (MNI) coordinates. No spatial smoothing was applied. Brains were segmented into gray matter, white matter and cerebrospinal fluid (CSF). Mean Gray-matter, white-matter and CSF fMRI signals, and movement parameters were regressed out of

functional data. Cerebral and frontal lobe MNI coordinates provided with xjView (www.alivelearn.net/xjview8) were used to restrict analysis to the cerebri.

Data from 4 participants in the pupil diameter study whose head moved by more than 2 mm or 2° were excluded from further analysis, leaving 30 participants. To further validate the results of the regional and whole-brain functional connectivity analyses in the pupil diameter study, we repeated these analyses with alternative preprocessing in which: 1. stimulus and outcome presentation events (convolved with SPM's canonical hemodynamic response function) were regressed out of the data. 2. Mean gray-matter signal was not regressed out. 3. The analysis was restricted to voxels that were activated in response to task stimuli or outcomes ($p < 0.001$ uncorrected), as determined by a general linear model that included regressors for stimulus and outcome presentation and for movement parameters. Results were qualitatively similar to the original analysis and are thus not presented.

3.6.3.2 General Linear Model analysis

Two general linear models were used to compare the way the fMRI BOLD signal response to task-relevant and task-irrelevant stimuli varied with baseline pupil diameter and pupil dilation response. Each model included regressors for task-relevant stimuli onset, task-irrelevant stimuli onset, and for each of these – a parametric regressor that reflected the trial-to-trial variability of either the baseline pupil diameter or the pupil dilation response. In addition, regressors that reflect head movement parameters were included in both models.

3.6.3.3 Regional functional connectivity analysis

To divide gray-matter voxels into uniformly-sized regions, we partitioned each brain recursively into 32 boxes as follows: first, the median X coordinate was used to split all voxels into two boxes. Then, each of the resulting subsets of voxels was divided by its median Y coordinate into two boxes. The same procedure was then repeated recursively with the median Z coordinates, with the median X coordinates and finally, with the median Y coordinates, resulting in 32 boxes of voxels. Mean functional connectivity strength was measured for each box in each game as the mean absolute correlation between all voxel pairs within the box. We then computed the across-games correlations of boxes' mean functional connectivity strength, both between the boxes, and with the baseline pupil diameter and the pupil dilation response.

3.6.3.4 Whole-brain functional connectivity analysis

To examine functional connectivity throughout the brain, we first computed a full voxel-to-voxel correlation matrix for each participants using the time series of all cerebral gray matter voxels (20786-29254 voxels per participant). In the pupil diameter study, this was done for each game separately. We then constructed a 2000-bin histogram of correlation (connectivity) strengths using the values of each correlation matrix. Game-by-game correlation between the pupil measurements and the number of functional connections was then computed for each bin separately, in order to assess whether there were fewer or more connections of this strength when gain increased (as assessed by pupil measurements). We also calculated the correlation between the pupil measurements and the mean functional connection strength, computed as the average absolute correlation coefficient.

3.6.3.5 Functional-connectivity clustering analysis

For each functional-connectivity correlation matrix, we constructed a 'functional connectivity graph' (Eguíluz et al., 2005) in which each voxel was represented by a vertex and two vertices were connected if the absolute value of the correlation between their respective voxels was in the top 0.05% of all voxel-voxel correlations. The 0.05% threshold was chosen so as to limit computing time to an acceptable level, resulting in 233929 to 431794 connections per graph. In the pharmacological manipulation study, which involved a much lower number of correlation matrices (22 vs. 540), we used a threshold of 1%, resulting in 2160185 to 3342793 connections per graph. To quantify the degree to which functional connectivity was tightly clustered, rather than broadly distributed, we computed each graph's clustering coefficient (Luce & Perry, 1949), defined as the number of closed triplets of vertices divided by the number of all connected triplets of vertices. Images of connectivity graphs were produced using custom-made software in the Processing programming environment (Reas & Fry, 2007).

3.6.3.6 Statistical analysis

Statistical analysis was carried out using MATLAB. All correlations values reported are Pearson correlation coefficients. Averaging of correlation coefficients was preceded by Fisher *r*-to-*z* transformation and followed by Fisher's *z*-to-*r* transformation, so as to mitigate the problem of the non-additivity of correlation coefficients (Fisher, 1921). Group-level significance of within-participant correlations was tested statistically by converting the correlation coefficients to *z* values, and then using a *t* test to determine whether the mean of this set of values is significantly different from 0. Significance of across-participant Pearson correlation

coefficients was computed using the Student's t-distribution. All tests were two tailed except where indicated otherwise.

3.6.4 Neural network model

To examine theoretically the relationship that should hold between gain, activations and functional connectivity, we constructed a recurrent neural network of 1000 fully connected units. Weights were randomly sampled from a uniform distribution between –0.01 and 0.01. On every trial, activations (a) were randomly sampled from a uniform distribution between 0 and 1, and then updated in a random order until each unit was updated 5 times. Unit i activation was computed as:

$$a_i = f\left(\sum_j w_{ij} a_j\right) \qquad (3.1)$$

where w_{ij} is the connection weight from unit j to unit i, and $f(x)$ is the sigmoid activation function:

$$f(x) = \frac{1}{1 + e^{-gain \cdot x}}. \qquad (3.2)$$

until each unit was updated 5 times. The same gain was used for all units. The end state was considered as the activation pattern of that trial. For each level of gain, we conducted 500 trials and computed the degree to which each pair of units was correlated across trials. The full unit-to-unit correlation matrix was used to compute the clustering coefficient as described below for the fMRI data. We repeated the simulation 100 times, each time with a different randomly determined weight matrix. Finally, we also tested the relationship between gain and unit-to-

unit correlations in an alternative "structured" network, which consisted of 10 groups of 100 units, with stronger connections within the groups (range: -0.05 to 0.05), and weaker connections between the groups (range: -0.0056 to 0.0056). The total sum of weights was equal to that of the non-structured network.

Chapter 4

Neural gain and the focus of learning*

Having examined in the previous chapter the *neural* effects of gain, as manifested in whole-brain fMRI, we now turn to the *behavioral* effects of gain. In this chapter we will examine the effects of gain on learning from multidimensional information. When presented with a set of stimuli, some people may attend to, and therefore learn most about concrete visual details, while others may attend to abstract semantic concepts associated with those stimuli. Evidence suggests that such variations in attention and learning may reflect stable individual predispositions (Felder & Silverman, 1988; Coffield et al., 2004; Felder & Spurlin, 2005). Here, we hypothesize that the expression of these predispositions is modulated by global variations in neural gain. Specifically, we propose that high gain focuses attention and learning on dimensions of the environment to which one is predisposed to attend, whereas low gain broadens attention, thereby weakening the constraint of prior dispositions on attention and learning.

---

* The material in this chapter appeared in Eldar, E., Cohen, J. D., & Niv, Y. The effects of neural gain on attention and learning. *Nature neuroscience* 16, 1146-1153 (2013), and was presented at SfN 2012, the Third Symposium on Biology of Decision Making 2013, and MathPsych 2013.

We first use a simple neural network model in which different neural representations compete through mutual inhibition, to demonstrate that applying a high global level of gain to all network units can make strong neural representations even more dominant, while further weakening weaker competing representations. Accordingly, we hypothesized that high gain results in processing that is more narrowly focused on the most strongly represented features of perceived information.

To test our hypotheses, we then use a novel task that quantifies the degree of learning about perceptual versus semantic features of stimuli (Figure 4.1), together with a standard trait questionnaire that assesses predispositions to attend to and learn about perceptual versus semantic dimensions of stimuli (Felder & Spurlin, 2005). In general, we expected participants to exhibit better learning for the type of features (perceptual or semantic) to which they are predisposed. More importantly, we hypothesized that the degree to which participants would selectively learn about their preferred type of features would be modulated by neural gain.

While it is impossible to directly measure gain in human participants, we have seen in previous chapters that pupil diameter may provide an easily measurable indirect index of gain (Section 2.3.3 & Chapter 3). Specifically, while baseline pupil diameter can be used to monitor changes in gain for an individual, phasic pupil dilations, which are inversely related to the baseline diameter, may provide an inverse index of gain that is better suited for between-participant comparisons, since they can be normalized to the baseline diameter, and thus, dissociated from factors such as physical pupil size that can confound baseline measurements.

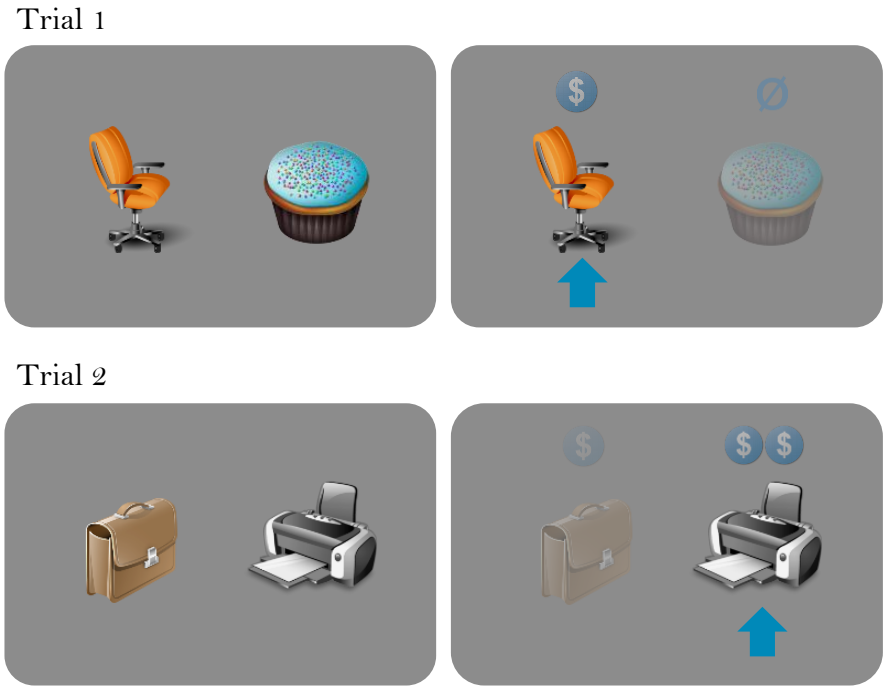

Figure 4.1. Experimental design of the visual/semantic learning task. In each trial, participants were presented with a choice between two images (objects or words). Participants were rewarded according to their choices, with counterfactual rewards also displayed. In this particular game, to maximize reward participants had to learn by trial and error that office-related images provide a higher reward than food-related images (semantic features), and that grayscale images provide a higher reward than color images (visual features). Each trial involved two new stimuli.

In addition, our neural network modeling of the effect of gain on learning suggested that the link between gain and focused learning should be mediated by a more tightly clustered pattern of neural interactions through which processing is selectively focused on particular input streams. In contrast, when gain in the model was low, widely distributed interactions mediated the concurrent processing of multiple stimulus features. Accordingly, we predicted that the degree to which functional connectivity is clustered, as measured using graph-theoretic analysis (Eguíluz et al., 2005), would correlate with a bias in learning performance toward the type of features that individual participants are predisposed to process.

4.1 A neural network model

First, to formalize our hypothesis about the effect of gain on attention and learning, we constructed a simple neural network model of the task, that learned a stimulus-reward relationship from examples (Figure 4.2A,B). The input to the network consisted of two separate streams of information, each representing one dimension (e.g., visual or semantic). One feature in each dimension was associated with a monetary reward whereas the other was not. We simulated a predisposition to attend to one dimension more than to the other by making connection weights in one stream stronger than those in the other stream (while maintaining a fixed sum of weights). We then examined the degree to which the network learned to associate the reward-predicting feature in each stream with a reward output, as a function of both the predisposition of the network and the level of gain.

With low gain, inputs from both the strong and weak streams propagated to the subsequent layers (Figure 4.2A), and the relationship with reward was learned for both types of features (that is, predisposition did not significantly bias learning; Figure 4.2C, black). In contrast, when gain was high, inputs in the strong stream dominated representations in the middle layer (Figure 4.2B) and learning of the input-reward relationship tended to proceed only on strongly-represented features (i.e., learning was biased towards features that the network was predisposed to represent; Figure 4.2C, gray). Thus, the simulations demonstrate that increased gain can focus learning on those features that the network is predisposed to represent. The simulations also demonstrate that gain affects communication patterns in the network: with lower gain multiple input streams interact (Figure 4.2A); but with higher gain weak input

streams have less of an effect on other parts of the network, with the result that network connectivity is more tightly clustered and separate subnetworks are formed (Figure 4.2B).

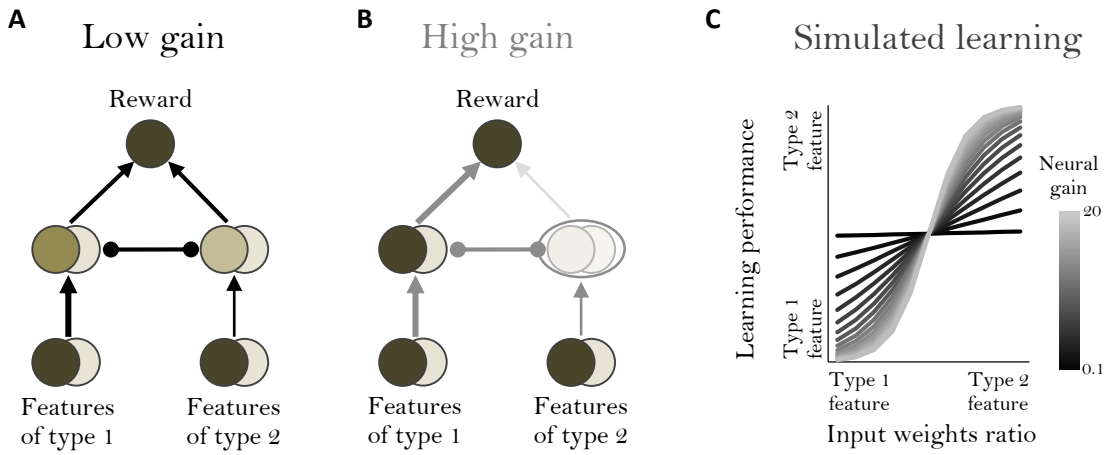

Figure 4.2. Simple reward-learning neural network. (A, B) Arrows denote excitatory connections, round edges denote inhibitory connections. Darker fill color indicates more activity and thicker lines indicate stronger weights in examples with low (A) and high (B) gain. With high gain, activity of weakly represented features (type 2) is blocked at the middle layer (circled), so the mapping between type 2 features and reward cannot be learned. This condition effectively separates the second input stream from the rest of the network. (C) Simulated learning of mapping between the reward-predicting features and reward. The relative strength of learning for the two features is shown as a function of the ratio between the input weights (varied between 1/2 and 2/1), for different levels of gain. The higher the gain, the more learning performance depends on the relative weight of each input stream.

4.2 Pupil diameter and adherence to predispositions

To test for the predicted relationship between pupil responses (as an index of neural gain) and the influence of attentional predispositions on learning, we asked participants to choose between pairs of multidimensional images (comprised of visual and semantic features) and rewarded them according to their choices. Unbeknownst to the participants, within each

stimulus set one visual feature and one semantic feature predicted monetary reward (Figure 4.1). For example, in one stimulus set, office-related images but not food-related images (semantic features) yielded reward and, similarly, grayscale images but not color images (visual features) yielded reward (rewards were additive so that a grayscale office-related image yielded twice the reward). Throughout 18 games, each with different semantic and visual dimensions and unique stimuli, we measured participants' visual and semantic performance separately using trials in which stimuli differed on either the visual or the semantic features, but not both.

In addition, we assessed each participant's predisposition to process either the visual or semantic features using the Index of Learning Styles (ILS) questionnaire (Felder & Spurlin, 2005). The ILS questionnaire contrasts a 'sensing' learning style that indicates a predisposition to process and learn about sense-related data such as visual features, with an 'intuitive' learning style that indicates a predisposition to learn about abstract concepts such as semantic categories.

The results showed that a more intuitive (and less sensing) learning style was correlated with better performance on the semantic trials compared to the visual trials ($r = 0.28$, $p = 0.05$ one tailed; Figure 4.3A; see Figure 4.4 for overall performance levels), consistent with a predisposition to attend to and learn about semantic vs. visual features of the stimuli. Critically, the degree to which task performance matched individual predisposition was strongly anticorrelated with mean pupil dilation response across individuals ($r = -0.96$, $p < 0.01$; Figure 4.3B). Given the inverse relationship between pupil response and gain discussed above, our finding suggests that the association between task performance and individual predisposition was itself associated with high gain. These behavioral results were fully replicated in a second experiment in which a different group of participants performed the same

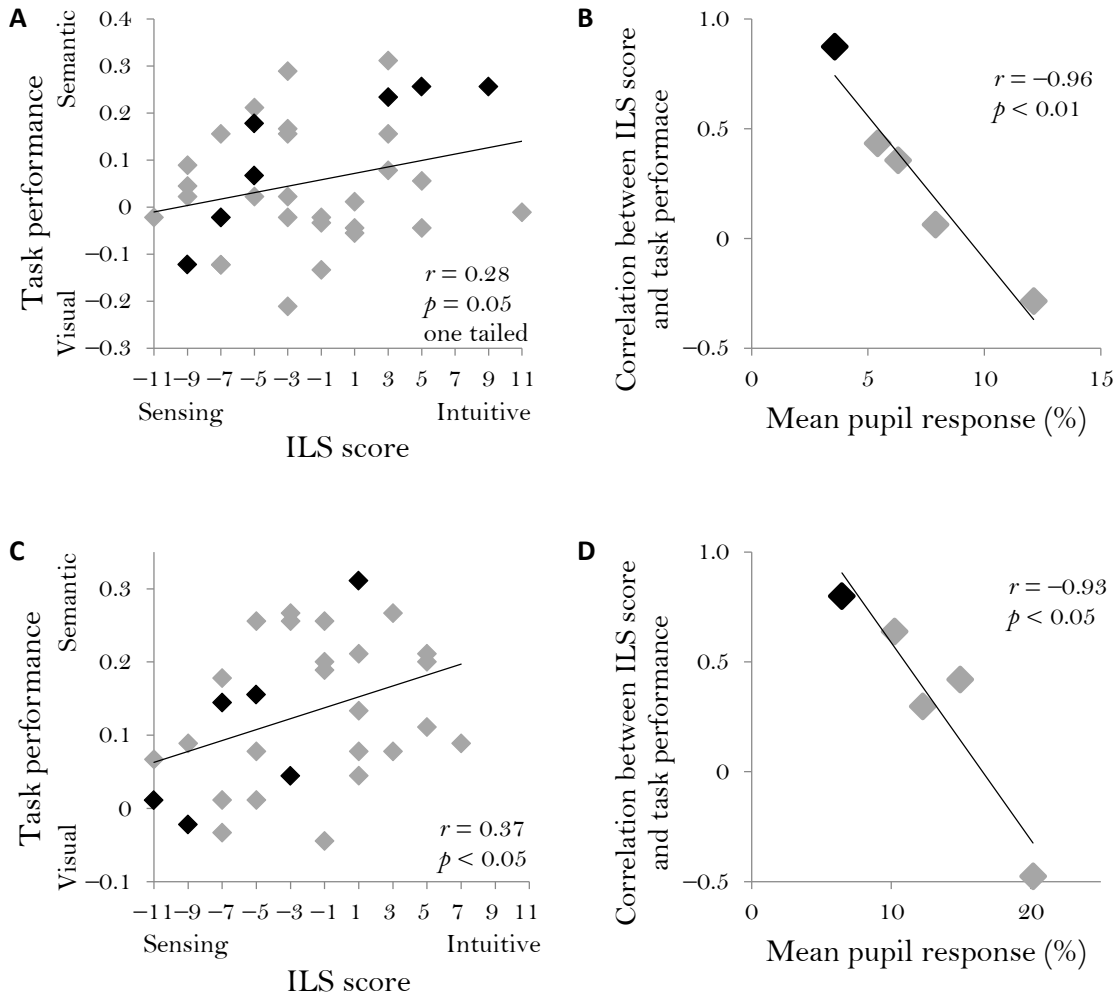

Figure 4.3. Relationship between learning performance and ILS scores. (A) Difference in learning about semantic and visual features in the behavioral experiment as a function of sensing-intuitive score on ILS questionnaire. Negative values indicate better visual performance (Y axis) and a sensing learning style (X axis), while positive values indicate better semantic performance and an intuitive learning style. $n = 35$. (B) Correlation between ILS sensing-intuitive score and visual-semantic performance difference on the task (as shown in (A)), as a function of mean pupil dilation response. To examine the degree to which task performance matched ILS score in participants with different levels of pupil response, participants were divided into 5 bins according to mean pupil dilation. Each data point represents a group of 7 participants. To illustrate, data points from the individual members of the group with lowest mean pupil response appear in black in (A). (C, D) Replication of behavioral results in the fMRI experiment with a different group of participants. $n = 30$. In (D) each data point represents a group of 6 participants.

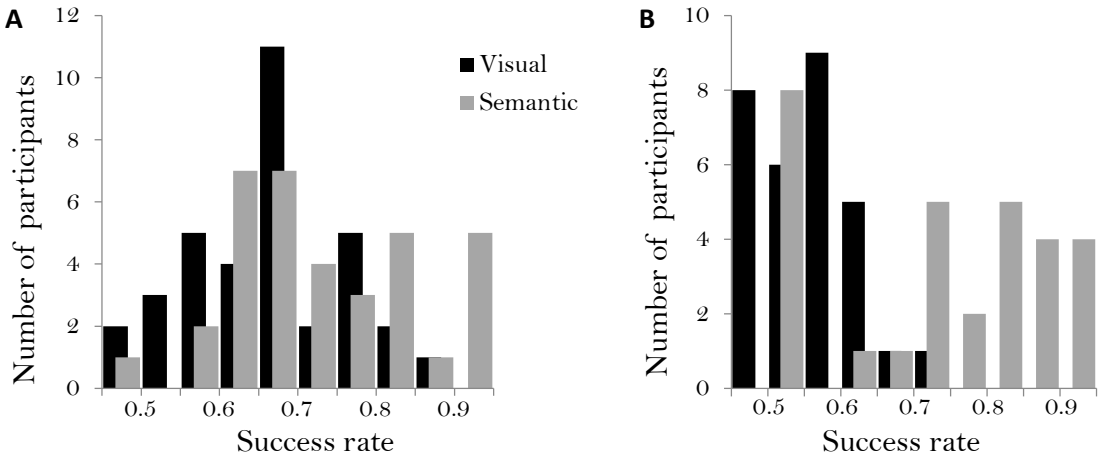

Figure 4.4. Performance on visual and semantic trials. Chance performance level is 0.5. (A) Behavioral experiment. $n = 35$. (B) Imaging experiment. $n = 30$. Performance on visual trials was lower in the imaging experiment compared to the behavioral experiment, most probably due to a lower quality visual display, but it was still significantly above chance (mean 0.54, $t_{29} = 5.05$, $p < 10^{-5}$).

task while being scanned using fMRI (Figure 4.3C,D). Moreover, in both experiments mean pupil dilation response did not correlate with overall task performance (behavioral experiment: $n = 35$, $r = -0.13$, $p = 0.44$; imaging experiment: $n = 30$, $r = 0.04$, $p = 0.82$), mean reaction times (following log transform; behavioral experiment: $n = 35$, $r = 0.17$, $p = 0.33$; imaging experiment: $n = 30$, $r = 0.06$, $p = 0.77$), or with answers to debriefing questions regarding interest, motivation and attention (Table 4.1). These results suggest that the relationship between pupil responses and adherence to one's learning predisposition cannot be explained in terms of fluctuations in overall level of arousal or attention to the task. Further analysis confirmed that the decrease in correlation between ILS score and task performance for participants with higher pupil response (lower gain) was not simply a result of a more limited range of ILS scores for these participants (Figure 4.5).

Correlation with mean pupil response	**Behavioral Experiment (*n* = 35)**	**Imaging Experiment (*n* = 29)**
Interest	$r = 0.26$, $p = 0.14$	$r = -0.39$, $p = 0.038$
Motivation	$r = 0.23$, $p = 0.19$	$r = -0.45$, $p = 0.013$
Difficulty to maintain attention		$r = 0.21$, $p = 0.27$
Correlation with match between task performance and ILS score (i.e., adherence to predispodition)	**Behavioral Experiment (*n* = 5 groups of 7)**	**Imaging Experiment (*n* = 29) (*n* = 5 groups of 6)**
Interest	$r = -0.46$, $p = 0.44$	$r = 0.43$, $p = 0.47$
Motivation	$r = -0.20$, $p = 0.75$	$r = 0.72$, $p = 0.17$
Difficulty to maintain attention		$r = -0.33$, $p = 0.58$

Table 4.1. Post-experiment ratings, pupil response and task performance. Relationship of post-experiment ratings of interest, motivation and difficulty to maintain attention, with pupil diameter (top) and adherence to predispositions in the task (bottom). Following the experiment, participants were asked to rate between 1 to 5 how interesting they found the experiment (Interest), how motivated they were to earn as much money as possible (Motivation), and, in the imaging experiment, how difficult it was for them to maintain attention (Difficulty to maintain attention). One participant in the imaging experiment did not fill out the debriefing questionnaire. Thus, in the group-based analysis (bottom panel), the group with the lowest ratings consists of 5 participants only.

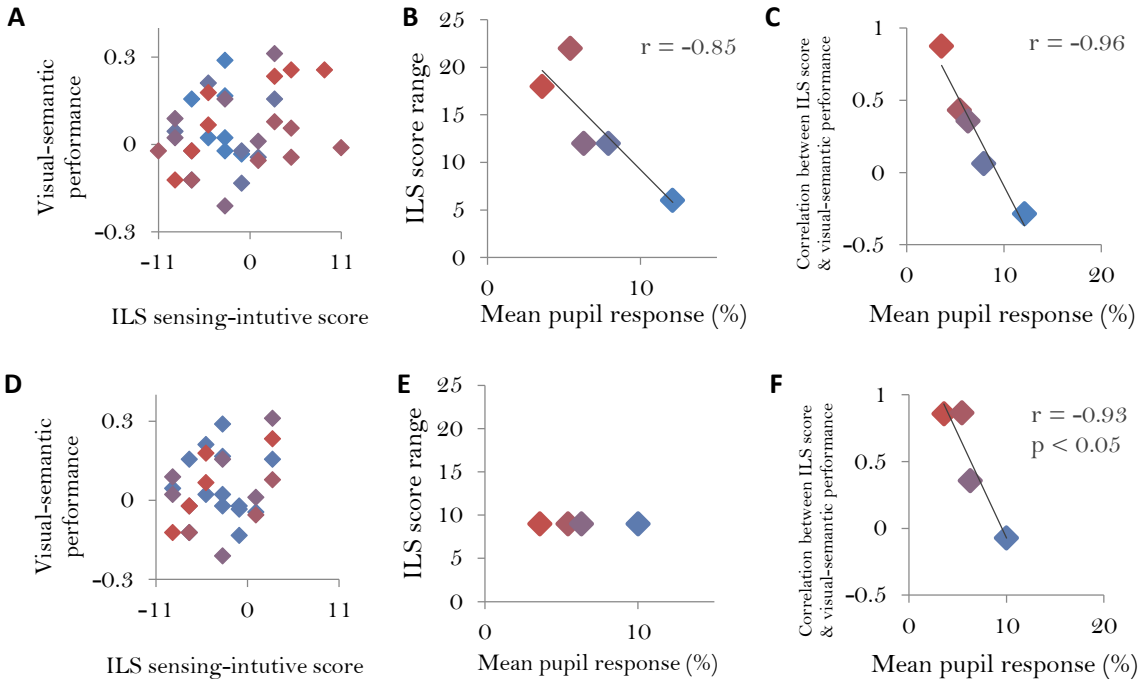

Figure 4.5. ILS score range and the relationship between ILS score and task performance. In the first experiment, high mean pupil response was associated not only with a decrease in correlation between ILS scores and task performance (A, C; r = -0.96, $p < 0.005$), but also with a decrease in the range of ILS scores (B; r = -0.85, p = 0.07). Decrease in range may thus serve as an alternative explanation for the decrease in correlation. However, in the second (imaging) experiment the range of ILS scores did not vary with mean pupil response (r = -0.03, p = 0.97), while the decrease in correlation between ILS scores and task performance recurred (r = -0.93, $p < 0.01$; Figure 4.3D). Furthermore, equating the range of ILS scores between the different groups of participants of the first experiment (D, E), did not eliminate the decrease in correlation between ILS scores and task performance that was observed in the first experiment (F). Thus, we can conclude that low mean pupil response was associated with a decrease in correlation between ILS scores and task performance irrespective of ILS score range.

(A) Visual-semantic performance difference on the behavioral task as a function of sensing-intuitive score on the ILS questionnaire. Negative values indicate better visual performance (Y axis) and a 'sensing' learning style (X axis), while positive values indicate better semantic performance and an 'intuitive' learning style. Color indicates binning according to mean pupil response, with a redder color indicating lower pupil response. n = 35. (B) Range of ILS sensing-intuitive scores for each group of participants. Participants were divided into 5 groups according to mean pupil response. Each data point represents a group of 7 participants. (C) Correlation between ILS sensing-intuitive score and visual-semantic performance difference in the task, as a function of mean pupil response. (D, E, F) ILS score range was equated for all groups by discarding the data of 6 participants whose score was lower than -9 or larger than 3, and merging the two groups whose mean pupil response was lowest.

4.3 Functional connectivity clustering and adherence to predispositions

To test whether the link between gain and focused learning was mediated by a more tightly clustered pattern of neural interactions, as predicted by our neural network simulations, we constructed a functional connectivity graph for each participant and each game (18 graphs per participant), and computed the clustering coefficient of each of these graphs (see Section 3.2.3 for details). In Chapter 3, we have already seen that, as predicted, the degree of clustering of functional connections was correlated with a pupillary index of gain (Figure 3.6). Here, we examine whether clustering also correlated with the degree to which learning was focused on stimulus features to which the individual was predisposed to attend. Consistent with our hypothesis, we found a significant game-by-game correlation between the clustering coefficient and a shift in learning performance toward the type of feature that the ILS scores indicated as preferred by each participant (mean $r = 0.08$ across participants, $t_{29} = 2.2$, $p < 0.05$). Concordantly, ILS score was correlated with the relationship between clustering coefficient and task performance ($r = 0.35$, $p < 0.05$; Figure 4.6). Thus, when participants' neural functional connections were more tightly clustered, task performance more strongly reflected individual predispositions.

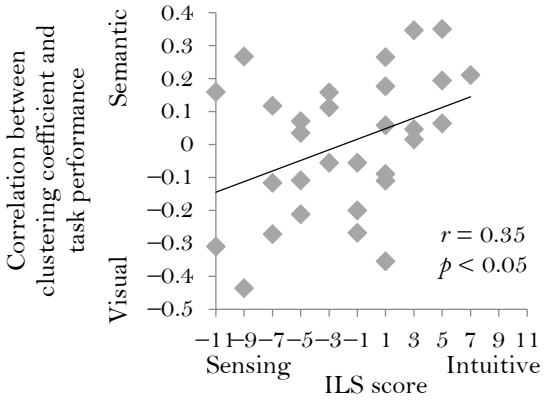

Figure 4.6. The clustering of functional connections and task performance: game-by-game correlation between clustering coefficient and visual-semantic performance difference in task as a function of sensing-intuitive score on the ILS questionnaire. $n = 30$.

4.4 Discussion

We investigated the relationship between global, brain-wide fluctuations in neural gain and the effect of individual priors or attentional predispositions (so-called 'learning styles') on trial-and-error learning behavior. More specifically, we used pupil-diameter measures as a proxy for global levels of neural gain, to test the hypothesis that predispositions constrain learning more strongly when gain is higher. In two experiments, the degree to which learning performance followed individual predisposition was strongly correlated with pupil response. In addition, we showed that within participants, the relationship between pupil diameter and focused learning was mediated by the clustering of functional connections. Taken together, these results support the hypothesis that high gain constrains the type of information that is learned from multidimensional sensory input in accordance with one's prior processing dispositions.

The results of our study provide a neural-computational framework within which past findings concerning the relationship of stress and norepinephrine levels to cognitive function can be understood. A large body of psychological research in humans suggests that stress (which is associated with high levels of norepinephrine) reduces the breadth of attention (Easterbrook,

1959; Staal, 2004). Another line of studies shows that stress and norepinephrine shift rat and human behavior from a flexible mode of behavior to a more rigid habitual mode in which previously established stimulus-response associations are followed (Dias-Ferreira et al., 2009; Schwabe & Wolf, 2011; Schwabe et al., 2010; Schwabe et al., 2011). Stress and norepinephrine have also been linked to diminished performance in tasks requiring cognitive flexibility (Alexander et al., 2007; Campbell et al., 2008). Our findings suggest an explanation of these previously observed phenomena in terms of the influence of the LC-NE system in globally modulating neural gain. Increased gain narrows attention by strengthening already strong neural representations at the expense of competing weaker representations. This, in turn, favors previously established patterns of behavior, which are subserved by well-established neural circuits and thus tend to form stronger representations.

We attempted to identify the effects of neural gain – a computational concept defined in terms of the input-output function of neural units – on behavior and on whole-brain fMRI metrics. This constitutes a novel, promising approach by which low-level principles of neural function may be linked via computational modeling to system-level neural and behavioral phenomena. However, the disadvantage of our approach is that it necessarily relies on a broad set of assumptions. Specifically, in making our predictions, in this and the previous chapter, we assumed that changes in pupil diameter would track changes in neural gain. Furthermore, our fMRI predictions were based on the assumption that the BOLD signal would reflect the neural effects of gain simulated by changes in firing rates in our computational models. This last assumption is particularly tenuous, since several studies have found dissociations between spiking activity and the BOLD signal specifically under conditions that are thought to involve changes in neuromodulation (Maier et al, 2008; Sirotin & Das, 2009; Logothetis, 2008).

Nevertheless, we presented here a diverse set of behavioral and imaging results that precisely match the predictions made by our neural network simulations of the effect of gain on neural activity, connectivity and behavior. This set of converging results, in addition to evidence from past studies, provides substantial support for the assumptions underlying this study.

The focusing effect of neural gain on processing may at first glance seem to be in conflict with previous accounts suggesting that tonically high gain reduces task-focused attention (Aston-Jones & Cohen, 2005). However, while our findings suggest that increased gain focuses attention on predisposed dimensions of sensory stimuli, these need not be related to the task at hand. Rather, if distracting stimuli are salient enough to evoke strong neural representations, our theory predicts that high gain would be associated with increased attention to distracters, and thus, with reduced task-focused attention. Our findings also fit well with a previous suggestion (Dayan & Yu, 2005) that phasic norepinephrine responses, which are stronger in low gain states (low tonic LC-NE activity), facilitate behavioral flexibility in response to unexpected target stimuli.

Several of our results drew upon graph-theoretic methods, which have been increasingly used to analyze both structural and functional brain imaging data (Bullmore & Sporns, 2009; Bullmore & Sporns, 2012). The strength of these methods lies in their ability to capture, by simple quantitative measures, characteristics of networks that are comprised of a very large number of elements. Most previous studies employing graph-theoretic analyses have investigated stationary aspects of neural processing networks, but a few recent studies have begun to examine how measures of functional brain network topology vary with behavior (Bassett et al., 2011; Kitzbichler et al., 2011; Nicol et al., 2012). The latter, however, analyzed relatively small networks (<120 nodes). In contrast, we used graph-theoretic measures to

examine how the topology of high-resolution whole-brain networks (>20000 nodes) varies with behavior. Our results indicate that such an analysis can provide meaningful insights into the way sensory information is processed and learned.

4.5 Conclusion

Our findings suggest that processing predispositions can influence learning, but that these priors are not always binding. Rather, brain-wide fluctuations in neural gain affect the distribution of neural interactions, thereby modulating the breadth of attention, and thus the extent to which processing and learning are constrained by prior dispositions.

4.6 Appendix: Methods

4.6.1 Neural network model

We modeled learning of stimulus-reward mapping from examples using a three-layer neural network. The network consisted of a 4-node "stimulus" input layer, in which the stimulus was represented using two types of features (e.g., in the case of our task, semantic and visual features of the stimulus), a "representation" middle layer and a "reward" output layer in which activity represented the expected reward (Figure 4.2A,B). As in our task, there were two possible features in each type, one of which was associated with a reward output. Our aim was to examine how associative learning changes as a function of gain and of the network's predisposition to represent either of the stimulus features more strongly. Thus, each stimulus consisted of a binary input vector where one input feature of each type was set to 1 (and the

rest were set to 0), and the weights associated with each input reflected the degree to which the network is predisposed to represent that type of feature. In addition, middle layer units inhibited each other (weight = −1), to simulate competition for attention between different representations. Unit *i* activation was computed as:

$$a_i = f\left(\sum_j w_{ij} a_j\right) \qquad (4.1)$$

where w_{ij} is the connection weight from unit *j* to unit *i*, and $f(x)$ is the sigmoid activation function:

$$f(x) = \frac{1}{1 + e^{-gain \cdot x}}. \qquad (4.2)$$

The parameter *gain* reflected the level of neural gain in the network, and had the same value for all units. To determine how much inhibition each middle layer unit should exert, the activation level of each unit was first computed based on the input from the input layer. Then, the resulting values were used to compute the magnitude of lateral inhibition in the middle layer, and activation levels were recomputed.

We ran the simulation with 15 different values of gain between 0.1 and 20, and with 16 different settings of predisposition to one of the input streams for each level of gain. Network predisposition to represent feature 1 relative to feature 2 was varied between 1/2 and 2/1, and input-to-middle layer weights were set accordingly, under the constraint that both weights sum to one (e.g. for a ratio of 1/2, weight 1 was set as 0.33 and weight 2 was set as 0.67).

On each run, each of the 4 possible stimuli and its associated reward output were presented to the network. The network's task was to learn to associate between the reward-associated

features and a reward output, and we examined the extent to which that was learned for each of the input streams. Learning proceeded as follows: weights from the middle layer units to the output unit (w_{oi}) were initialized to 0, and the output unit activation (o) was computed according to equation 1. Then, the difference between the target (t) and actual output activation was used to update the weights to the output unit using the delta rule (McClelland & Rumelhart, 1988):

$$w_{oi} = w_{oi} + (t - o)f'\left(\sum_j w_{oj}a_j\right)a_i \qquad (4.3)$$

where $f'(x)$ is the derivative of the activation function, which in this case is:

$$f'(x) = gain \cdot f(x)\big(1 - f(x)\big). \qquad (4.4)$$

The learned weights w_{oi} reflected the degree to which the network learned to associate each of the stimulus features with the reward output. We thus used the ratio between these weights to represent the bias in learning performance towards either of the reward-associated features. It is easy to see that the only term that differentiates between the update equations of the two weights is a_i, the activation of the respective middle layer unit. Indeed, the ratio between the learned weights followed the ratio between the activation of the two middle layer units. Each run was repeated 100 times, with a random ordering of the stimuli, and the resulting weight ratios were averaged.

4.6.2 Experimental methodology

4.6.2.1 Participants

36 naïve participants (mean age 25.1, age range 18-61, 22 females) performed the behavioral experiment and 35 naïve participants (mean age 20.5, age range 18-30, 25 females) performed the fMRI experiment. Participants were from the Princeton University area, and gave written informed consent before taking part in the study, which was approved by the university's institutional review board. Participants in the behavioral experiment received monetary compensation according to their performance on the task (6 cents per reward point, $13.5-$16.2 total, mean $14.88). fMRI Participants received monetary compensation for their time, as well as a bonus according to their performance (4 cents per reward point, $8.04-$10.72, mean $9.47).

4.6.2.2 Stimuli

The experiment involved 18 stimulus sets, half of which consisted of images of objects and the other half consisted of images of words. Words were generated using the Processing programming environment (Reas & Fry, 2007), and object images were collected from various sources on the internet using the Creative Commons search interface (http://search.creativecommons.org/), and edited using Adobe Photoshop CS5 (Adobe Systems Inc.). To minimize luminance-related changes in pupil diameter, all stimuli were made isoluminant with the background, to best approximation (see Section 3.6.1.2 for further details).

4.6.2.3 Behavioral task

Participants chose between pairs of stimuli and received monetary reward according to their choices. On each trial, participants had 3 s to choose between two stimuli, after which the reward was presented for 2 s. Inter-trial interval was varied randomly (uniformly) between 6 s and 10 s. We used a relatively long inter-trial interval to allow enough time following each trial for the pupil dilation response to evolve (Hoeks & Levelt, 1993). To minimize inter-subject variability, all participants encountered the same stimulus sets in the same order. For the same reason, as well as to speed up learning, participants were presented with both the reward for their choice above the chosen stimulus, and (slightly dimmed) the reward that they could have received if they had chosen the other stimulus (Figure 4.1). No stimulus appeared more than once.

Participants were instructed that stimuli had some properties that predict reward. They then underwent a short training session with a few example trials before starting the task. Unbeknownst to the participants, each stimulus set had one visual feature (bright background, blurry texture, etc.) and one semantic feature (food, sea-related, etc.) that was rewarded, and these differed from game to game. For example, in a particular game, choosing a grayscale image or an image of food led to reward while choosing a color image or of office equipment did not lead to reward. Rewards for the two features were additive such that choice of a stimulus that possessed both rewarding features resulted in two reward points.

Each of the first two trials included one stimulus that possessed the rewarding visual and semantic features, and thus yielded two reward points, and one stimulus that possessed neither of the rewarding features, and thus yielded no reward. In the following ten trials stimuli

differed on either the visual (5 trials) or the semantic (5 trials) dimension, but not both. These trials allowed us to measure performance on the visual and semantic dimensions separately. Performance was computed as the proportion of trials in which the more highly rewarding stimulus was chosen. One fMRI participant was excluded from the analysis due to lack of cooperation, as evidenced by performance that was lower than chance and frequent eye closing. Performance of all other participants was better than chance. Following completion of the task, participants completed the Index of Learning Styles questionnaire (Felder & Spurlin, 2005). Finally, participants filled out a standard debriefing questionnaire in which they were asked to rate on a scale of 1 to 5 how interesting they found the experiment, how motivated they were to earn as much as possible, and, in the imaging experiment, how difficult it was for them to maintain attention during the task.

4.6.2.4 Eye tracking

Eye tracking methodology is described in detail in section 3.6.1.4. A desk-mounted ASL model 504 eye-tracker (Applied Science Laboratories, MA) was used to measure participants' left pupil diameter at a rate of 60 samples per s while they were performing the behavioral task with their head fixed on a chinrest. Only participants with at least 30 valid (i.e., mostly artifact free) trials were included in the across-participant analysis of mean pupil dilation ($n = 35$ for the behavioral experiment, $n = 30$ for the imaging experiment).

4.6.2.5 fMRI methodology

fMRI data acquisition and processing is described in detail sections 3.6.1.5 and 3.6.3.

4.6.2.6 Statistical analysis

Statistical analysis was carried out as described in section 3.6.3.6.

Chapter 5

Manipulating the effect of gain on perception and memory*

In the previous chapter, we saw that increased gain focuses information processing on features to which one is predisposed to attend. This finding could be explained in terms of a direct relationship between neural gain and predispositions. Instead, I argued, high gain has the general effect of focusing attention on the most strongly represented features, and in the specific case that we studied, those features happened to be determined by individual predisposition. Thus, had participants been biased by means of an attentional manipulation to attend to a different set of features, these should have been the features on which high gain focused processing. In this chapter, I test whether the effects of neural gain are indeed sensitive to manipulations of attention.

---

* Parts of this chapter were presented at Cosyne 2012.

5.1 Introduction

Perception depends not only on features of a physical stimulus, but rather, it is also determined by the context in which a stimulus is perceived and the stimulus' relationship to prior knowledge and experience. For instance, the middle letter in the stimulus "CAT" may look more like an H if you focus on its shape, but more like an A if you focus on the surrounding letters and the word they could form (Figure 5.1). Extensive work (Reicher, 1969; Palmer, 1975; Massaro, 1979; McClelland & Rumelhart, 1981; Paap et al., 1982; Pellicano & Rhodes, 2003) has investigated the relative influence of stimulus information and context on perceptual processing, and the extent to which these are integrated in perception. Here, we propose that the extent of this integration is subject to modulation by neural gain. Specifically, we test the hypothesis that the extent to which perceptual processing is dominated by a particular source of information (e.g., stimulus or context) or integrates the different sources, is determined by brain-wide levels of neural gain (Servan-Schreiber et al., 1990; Aston-Jones & Cohen, 2005; Eldar et al., 2013).

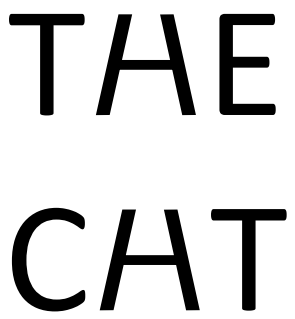

Figure 5.1. Two example stimuli. The resemblance of the trigram stimuli to known words favors perception of the ambiguous middle character as an H in the top stimulus and A in the bottom stimulus.

Our hypothesis follows from the idea that, as neural gain increases throughout the brain, excited neural units become even more active and inhibited units become even less active (Figure 1.2). Consequently, processing occurs faster, and is thus dominated by proximal sources of information that have the strongest, most immediate influence. In contrast, when gain is lower, processing is slower, permitting the integration of more weakly activated or distal sources of information, that take longer to exert their influence. We explored this hypothesis using a neural network model, and then tested it experimentally using an ambiguous-letter perception task in conjunction with pupillometry.

5.2 A neural network model of perceptual integration of letter shape and context

The sharpening of perceptual focus as a result of higher neural gain is illustrated by the neural network model shown in Figure 5.2A. Here we simulated the influence of different sources of information on the processing of ambiguous visual stimuli (based on a previous model developed by McClelland & Rumelhart, 1981). Consider, for example, the character "*A*" flanked by C and T (i.e., C*A*T). We simulated the ambiguity of *A* by providing partial bottom-up input to both the A and H letter units, while the C and T units received maximal input. We chose to provide stronger input to the H unit, simulating stronger visual similarity to that letter, so that we may contrast the contributions of visual input with semantic information: Since the letter units corresponding to the word CAT project to the CAT word unit, activation of the word unit generated semantic top-down feedback, further exciting the word-congruent letter units (i.e., C, A and T) and inhibiting word-incongruent letter units, such as 'H'. As it is not possible

to interpret the same stimulus simultaneously in two different ways (Necker, 1832), the A and H units competed through mutual inhibition, until only one prevailed in any given trial. The extent to which bottom-up vs. top-down influences affected this competition interacted with the level of gain. As seen in Figure 5.2B, under low neural gain top-down information could influence processing, so that despite the stronger bottom-up input to the H unit, A and H were equally likely to prevail as the interpretation of the middle letter. In contrast, high neural gain enhanced the initial advantage provided by the bottom-up input in support for interpreting the letter as 'H', thus allowing the H unit to outcompete the A unit and determine network output before the top-down support for the 'A' interpretation could have its influence (Figure 5.2B, blue line).

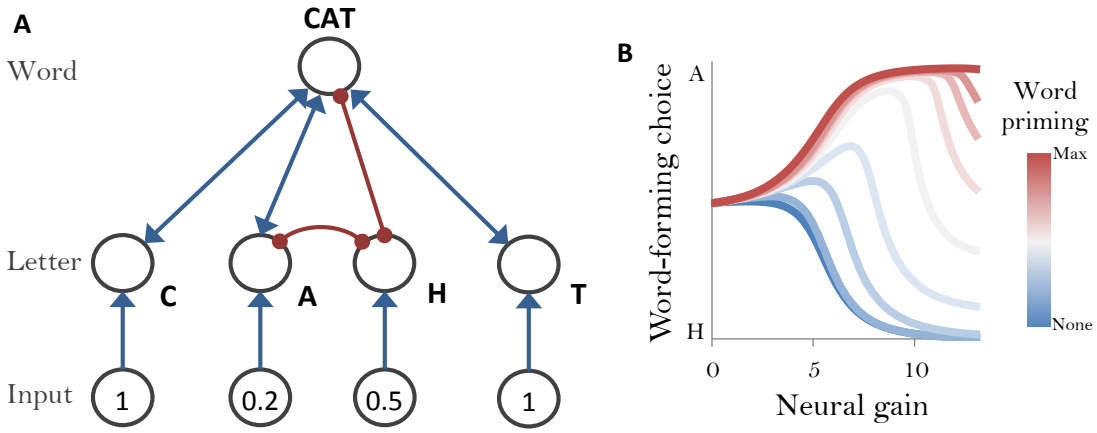

Figure 5.2. A neural network model of the effect of neural gain on ambiguous letter perception. (A) A model simulating perception of a stimulus similar to the bottom one in Figure 5.1. Blue lines: excitatory connections, red lines: inhibitory connections. (B) Simulated letter perception as a function of neural gain and the degree of priming of the CAT word-layer unit.

5.3 Pupil diameter and perceptual processing

To test for an effect of neural gain on perception of ambiguous letters in humans, we showed participants letter strings such as CAT, and asked them to indicate which letter the ambiguous character resembled the most, irrespective of whether it formed a word. Based on the simulation results, we predicted that high levels of neural gain would correlate with choices of the letter most visually resembling the character (e.g., H), whereas low gain would be associated with choices of the letter that formed the word (e.g., A).

As mentioned previously, it is not possible to directly measure neural gain, nor the noradrenergic activity thought to regulate gain (Aston-Jones & Cohen, 2005), using non-invasive methods in human participants. Thus, in line with the findings reported in chapters 3 and 4, we used pupil diameter as an indirect index of gain. Specifically, we measured the mean pupil dilation response to task stimuli throughout the experiment, and normalized the responses to participants' baseline pupil diameter to obtain a stable measure that is thought to be inversely related to tonic LC-NE activity and sustained levels of neural gain, and that can be compared across participants.

As predicted, lower mean pupil response (indicating higher sustained neural gain) was associated with shape-related (versus word-related) perception of ambiguous letters ($r = 0.37$, $t_{65} = 3.2$, $p < 0.005$; Figure 5.3A). Thus, high neural gain was associated with a greater influence of bottom-up information about the visual stimulus than the influence of top-down information about known words that the letter might complete. Our model suggests that this is because the letter's shape provided a more immediate and stronger source of activation than the top-down influence of the word. However, it is also possible that high neural gain generally

favors processing of bottom-up (e.g., visual) input over processing of top-down (e.g., semantic) influences, irrespective of the immediacy or relative strength of these signals.

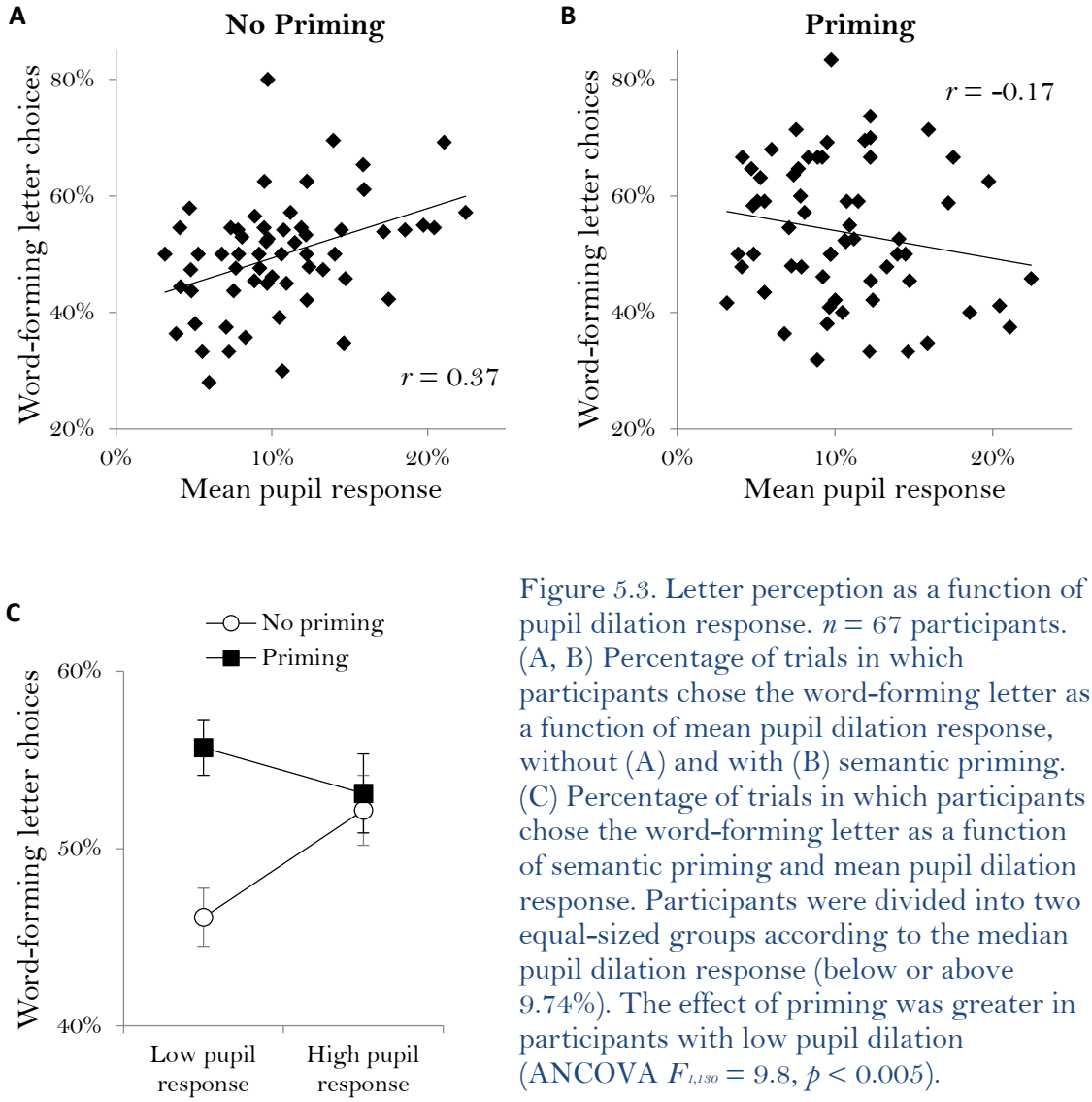

Figure 5.3. Letter perception as a function of pupil dilation response. $n = 67$ participants. (A, B) Percentage of trials in which participants chose the word-forming letter as a function of mean pupil dilation response, without (A) and with (B) semantic priming. (C) Percentage of trials in which participants chose the word-forming letter as a function of semantic priming and mean pupil dilation response. Participants were divided into two equal-sized groups according to the median pupil dilation response (below or above 9.74%). The effect of priming was greater in participants with low pupil dilation (ANCOVA $F_{1,130} = 9.8$, $p < 0.005$).

5.4 An attentional manipulation

To determine whether gain enhances sources of information that are strongest and/or most immediate, or rather it favors a particular source (bottom up vs. top down), we used semantic priming to pre-activate the top-down word information prior to presenting the stimulus. In our network model, pre-activating the word unit for CAT before presenting the stimulus caused the model to settle on the word-congruent interpretation of the ambiguous character more frequently and, critically, this interacted with gain: When the word was adequately primed, higher gain became associated with a higher frequency of word-congruent letter perception (Figure 5.2B, red line). Thus, the model predicted that gain would enhance processing of the most activated source of information. To test this empirically, on half the trials (priming condition) we preceded the letter strings by subliminal presentation of a semantically related prime word (e.g. DOG preceded CAT). On the other half of trials (no-priming condition), a non-word was presented subliminally. Based on the model, we predicted that semantic priming would shift the positive correlation between pupil response and letter perception in the opposite direction. Our findings were consistent with this prediction. When words were semantically primed, pupillary responses indicating high neural gain were no longer associated with shape-related perception (r = -0.17 vs. r = 0.37, z = 3.17, $p < 0.005$; Figure 5.3B). Moreover, while semantic priming generally increased word-related perception (main effect of priming: ANCOVA $F_1 = 16.2$, $p < 0.005$), it did so only in participants whose pupil responses indicated high gain (i.e., mean pupillary response below median; Figure 5.3C; priming × pupil response interaction: ANCOVA $F_{1,130} = 9.8$, $p < 0.005$). This group of participants was more likely than chance to perceive the shape-related letter in the absence of priming ($t_{32} = 2.38$, $p < 0.05$), and

the word-related letter in the priming condition ($t_{32} = 2.89$, $p < 0.01$). In contrast, participants whose pupil responses indicated low neural gain (i.e., whose pupil response was higher than the median) were relatively unaffected by the priming manipulation ($t_{32} = 0.35$, $p = 0.73$), exhibiting almost equal sensitivity to letter shape and word in both conditions (no priming condition: $t_{32} = 1.40$, $p = 0.17$; priming condition: $t_{32} = 1.40$, $p = 0.17$).

5.5 Within-participant variations

While we observed the predicted relationship between neural gain and letter perception across participants, we did not find a similar relationship within participants. That is, we did not find a significant relationship between letter perception and trial-by-trial variations in pupillary response. One reason for this may be that neural gain did not vary sufficiently within individual participants during the course of the experiment for such a relationship to be detectable. Consistent with this possibility, the difference in mean pupil response between the first and second halves of the experiment was significantly lower within participants (mean 2.3%) compared to between participants (mean 5.3%; $t_{66} = -7.92$, $p < 10^{-11}$).

In addition, it is possible that the high level of noise associated with pupillometric measurements does not allow trial-by-trial within-participant effects to be detected. This problem may be circumvented by using reaction time as a mediating variable between pupil response and letter choice. Our neural network model indicated that the relationship of neural gain to reaction time (with high gain associated with faster responses) may be more robust, and thus easier to detect, than the relationship with choice behavior (Figure 5.4A). The model also predicted that the effects of priming on letter choice should interact with reaction time in the

same way that it does with gain – an interaction that can be detected without pupillometry. In agreement with these predictions, we found that trial-by-trial pupil responses were positively correlated with reaction time, during both the no-priming (mean $r = 0.09$, $t_{66} = 2.72$, $p < 0.01$) and priming conditions (mean $r = 0.06$, $t_{66} = 2.01$, $p < 0.05$). Critically, reaction times were faster for shape-related letter perceptions in the no-priming condition ($t_{66} = -2.15$, $p < 0.05$) but not in the priming condition ($t_{66} = 0.40$, $p = 0.69$; ANOVA priming × letter-choice interaction analysis: $F_{1,1} = 3.36$, $p < 0.05$ one-tailed; Figure 5.4B). This priming-dependent relationship between reaction time and letter choice mirrors the priming-dependent relationship between pupil response and letter choice found in the between-participant analysis (Figure 5.3C).

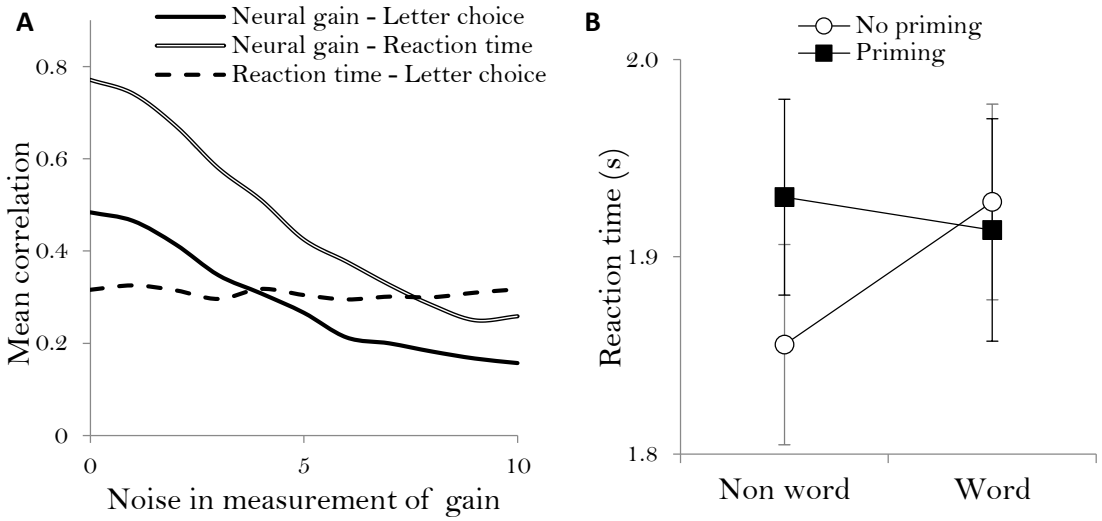

Figure 5.4. Pupil response, reaction time and letter perception. (A) Correlations, in model simulations, between pupillometric neural gain, reaction times and letter choices, as a function of noise in the "measurement" of gain. The latter was used to simulate noise assumed to be associated with pupillometric measurements as an index of actual neural gain (see Methods for details). For high levels of noise, the relationship between neural gain and letter choice is easier to detect through their relationship with reaction time. (B) Mean reaction time in human participants as a function of letter choice and semantic priming ($n = 67$). Faster reaction times were associated with choice of the non-word-forming letter, only when the word was not semantically primed.

5.6 Voluntary direction of attention

So far, strength of activation was manipulated by varying factors typically associated with involuntary processes (i.e., stimulus salience and subliminal priming). In an additional experiment, we tested whether the same effects obtain when activation is manipulated by the voluntary allocation of attention. To this end, we presented participants with words in one of two highly dissimilar fonts. To focus participants' attention on the shape of the words, we asked them to rate how readable each word was. Then, to test the degree to which participants had selectively processed word shape, we had them perform a recognition memory test, in which half of the target words assumed the same shape as in the readability rating phase, and the other half appeared in a different font. Participants with low pupillary response during the readability task (i.e., mean pupil response below median) were significantly affected by the change of font (mean d' difference: -0.36 ±0.1, $t_{20} = 3.5$, $p < 0.005$) whereas participants with high pupillary response were not (mean d' difference: +0.07 ±0.1, $t_{20} = -0.6$, $p = 0.53$), suggesting that low pupillary response (high gain) was associated with more selective processing of word shape (correlation between pupil response and d' difference: $n = 43$, $r = -0.41$, $p < 0.01$; Figure 5.5A,C). Moreover, this effect was not evident in words for which participants did not rate readability, but instead, performed a control, semantic task (correlation between pupil response and d' difference: $n = 29$, $r = 0.07$, $p = 0.72$; difference between readability and semantic tasks: $z = 2.01$, $p < 0.05$; Figure 5.5B,C). This indicates that the effect of high gain augmented the influence of attention in accordance with task demands. Finally, pupil response did not significantly correlate with general recognition performance levels ($n =$

43, r = -0.12, p = 0.44), suggesting that pupillary indices of gain primarily reflected an interaction with the distribution of attention, not overall task engagement.

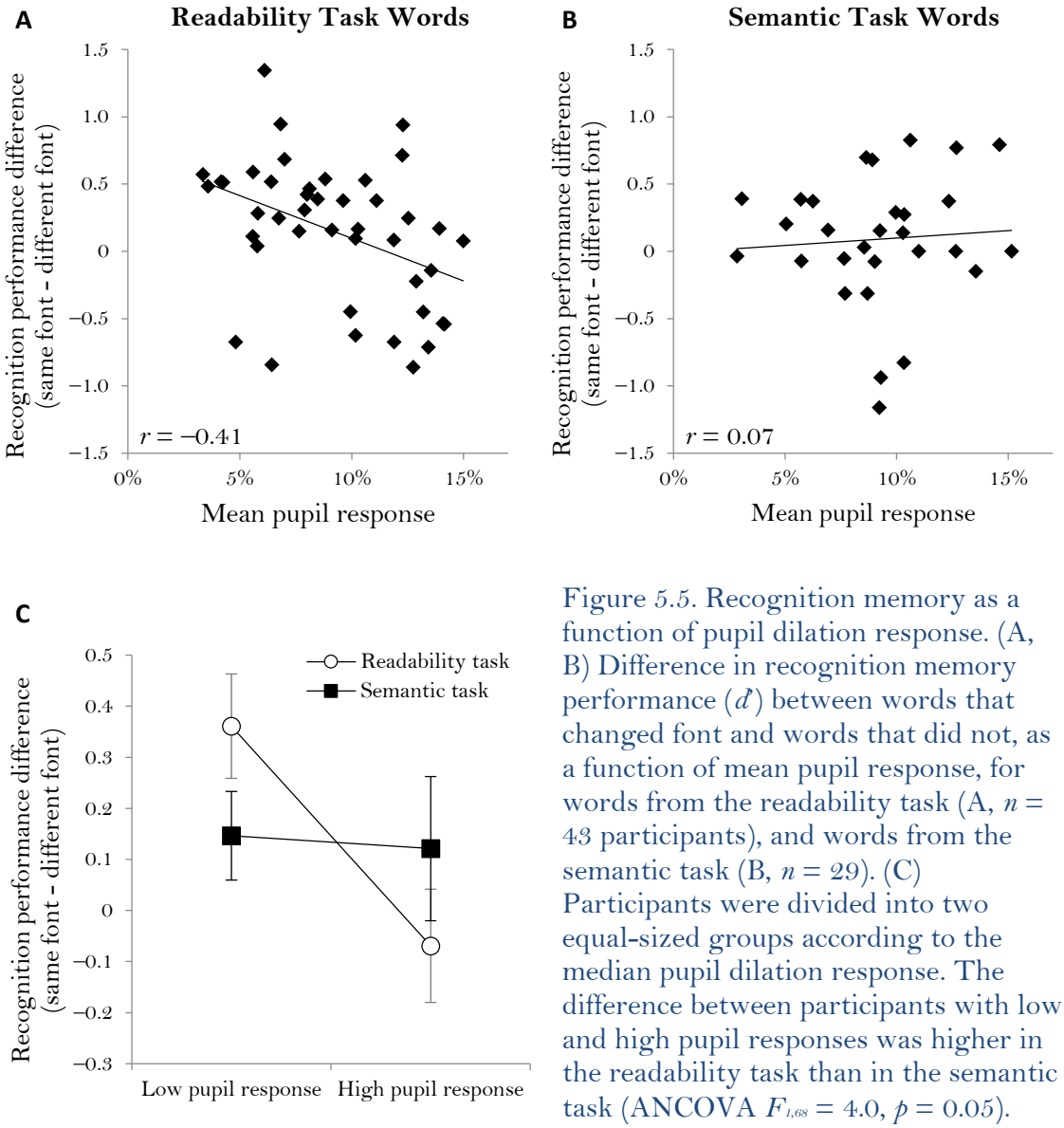

Figure 5.5. Recognition memory as a function of pupil dilation response. (A, B) Difference in recognition memory performance (d') between words that changed font and words that did not, as a function of mean pupil response, for words from the readability task (A, n = 43 participants), and words from the semantic task (B, n = 29). (C) Participants were divided into two equal-sized groups according to the median pupil dilation response. The difference between participants with low and high pupil responses was higher in the readability task than in the semantic task (ANCOVA $F_{1,68}$ = 4.0, p = 0.05).

5.7 Conclusion

We showed that individual differences in a pupillary index of neural gain are correlated with the degree to which perception and memory are influenced by the strongest and most immediate source of information. Our findings are consistent with the hypothesis that increases in neural gain affect network dynamics, favoring the influence of strong and immediate sources of information over weaker or more distal sources. It is important to note that, while we illustrated this point theoretically in a neural network that used top-down feedback connections to implement the effects of context, our theory would make similar predictions for models in which context effects are mediated through feedforward connections, so long as these are less direct (Paap et al., 1982) or weaker (Massaro & Cohen, 1991) than those mediating stimulus information. The critical interaction is between gain and the strength (and corresponding timing) of the influence of a source of information on the process of interest. Our priming results supported this conclusion, demonstrating that increases in gain favored the strongest source of information irrespective of its source (i.e., bottom-up vs. top-down). Furthermore, our findings from the final experiment suggest that gain interacts with the strength of processing irrespective of its cause (i.e., whether it is due to automatic processes or the influence of attention). Finally, these findings lend further support to a dimension of individual differences in information processing — breadth of integration — as well as a practical way of measuring it, which may help explain differences in behavior in real world domains, including disturbances of behavior in psychiatric disorders. We will expand on this last point in Chapters 6 and 7.

5.8 Appendix: Methods

5.8.1 Neural network model

We simulated the effect of neural gain on perception of the ambiguous letter in the stimulus CAT using a neural network model loosely based on McClelland and Rumelhart's (1981) interaction activation model. The network consisted of three layers: a 'visual' input layer, a letter layer and a word layer (Figure 5.2A). All weights were set to 1 (for excitatory connections) or -1 (for inhibitory connections). Since C and T are unambiguous, their respective letter-layer units received maximal input (input = 1). In contrast, since the middle letter was ambiguous, the H and A letter-layer units received sub-maximal input (input < 1). The H received stronger input than A to reflect the fact that the shape of the ambiguous letter was closer to H. As perception of the letter H is mutually exclusive with perception of the letter A, the corresponding letter-layer units had inhibitory connections between them. Similarly, because perception of the letters C, A and T is consistent with perception of the word CAT, excitatory connections were assigned among the corresponding units.

To simulate the limited exposure time used in the experimental task, input was presented to the network for 225 iterations, during which the activity a_i^t of network unit i at time step t built up gradually according to a weighted sum of its inputs:

$$a_i^t = 0.9a_i^{t-1} + 0.1f\left(b_i + \sum_j w_{ij}a_j\right) + n \qquad (5.1)$$

where b_i is the bias to unit i (set to -0.5 for all units), w_{ij} is the connection weight from unit j to unit i, n refers to a normally distributed random noise variable, and $f(x)$ is the sigmoid activation function:

$$f(x) = \frac{1}{1 + e^{-gain \cdot x}}. \qquad (5.2)$$

The parameter *gain* reflected the level of neural gain in the network.

Since the network's task was to reach a decision between perception of the middle character as A and H, two corresponding decision units were added to the network (activity initialized to 0, bias = 0). The decision units shared a mutually inhibitory connection with one another (weight = -1) and a mutually excitatory connection with the respective letter-layer units (weight = 1). Following presentation of the stimulus, the network switched to a 'decision' mode, in which the biases of the letter-layer A and H units were increased from their resting state of -0.5 to 0, simulating the allocation of attention to the letter-decision task (Cohen et al., 1990). Activity was updated using Equation 1 until one of the decision units reached an activity level of 0.9 or 1000 iterations were completed, at which point the probability of choosing the word-forming letter (A) was computed as the activity of the A decision unit divided by the sum of the activity of both decision units.

The strength of the inputs to the H (0.52) and A (0.2) letter-layer units and the level of noise (standard deviation = 0.035) were adjusted so as to make the network equally likely to decide in favor of H or A under conditions of low gain (*gain* = 1).

5.8.1.1 Simulation 1: variations in neural gain

To simulate the effect of neural gain on perception of the ambiguous letter, the gain parameter was varied between 0 and 10. For each level of gain, the procedure described above was repeated 1000 times, and the resulting decisions were averaged.

5.8.1.2 Simulation 2: semantic priming

To simulate the effect of semantic priming on perception of the ambiguous letter, the simulation started with a 'priming' phase in which, prior to presentation of the stimulus input, the CAT word unit received an excitatory 'priming' input (varied between 0 and 1, weight = 1) for 33 iterations. The simulation then proceeded as in Simulation 1. It was repeated 1000 times for each value of "priming" input and each value of neural gain, and the results were averaged.

5.8.1.3 Simulation 3: pupillometric noise

The purpose of this simulation was to determine whether, according to the model, it would be reasonable to use reaction time as a mediating variable in assessing the relationship between gain and letter choice. The motivation for doing so is that it can be assumed that noise corrupts the relationship between pupillometric measurements and actual neural gain within individual participants. However, if gain is more reliably related to reaction time, then the latter might be used as a proxy in assessing the relationship to letter choice. To simulate the effects of noise in measurements of neural gain, the measured level of neural gain was computed as the true level plus randomly distributed noise, the standard deviation of which varied between 0 and 10 (see Figure 5.4A). The simulation proceeded as in Simulation 1. Reaction time was computed as the

iteration in which a decision unit reached an activity level of 0.9. If that did not happen within 1000 iterations then reaction time was counted as 1000. The simulation was repeated 1000 times for each level of noise, and each level of neural gain, with and without priming (priming input = 0 or 1). The correlations between measured gain, reaction time and letter choice were computed for each repetition and averaged.

5.8.2 Ambiguous letters experiment

5.8.2.1 Participants

30 participants (mean age 20.4, age range 18-23, 25 females) performed a preliminary experiment (see below) and 83 participants (mean age 21.6, age range 18-61, 65 females) performed the main experiment. The sample size was chosen based on previous studies of semantic priming effects (Lucas, 2000). Participants were from the Princeton University area, and gave written informed consent before taking part in the study, which was approved by the university's institutional review board. Participants received either monetary compensation ($10) or course credit for participation.

5.8.2.2 Experimental task

Participants were instructed to identify letters regardless of whether they formed words. Participants were presented with 88 3-letter strings, 52 of which included an ambiguous letter that could form a word depending on the way that it was perceived. Half of these letter strings were preceded by subliminal presentation (33 ms) of a semantically related word, while the other half were preceded by subliminal presentation of a similarly sized non-word. Each letter

string was semantically primed in half of the participants. Following the priming stimulus, the 3-letter target stimulus was presented. The target stimulus was flanked by %%% on both sides so as to mask the priming stimulus, which could consist of more than 3 letters. After 225 ms, the 3-letter string disappeared from the screen and an arrow pointed to where the target letter previously appeared. Participants had 5 seconds to choose, out of a list of 4 letters, to which letter the target letter was most similar. The list always included the two letters that the ambiguous letter resembled and two other letters that did not appear in the letter string. Choices of one of the two letters that did not appear in the letter string were infrequent (less than 5% of trials) and were not included in the analysis. Inter-trial interval was varied randomly (uniformly) between 6 s and 10 s. We used a relatively long interval to allow enough time following each trial for the pupil dilation response to resolve (Hoeks & Levelt, 1993).

To avoid the influence of any response biases that may have resulted from conscious awareness of the priming manipulation, participants were asked whether they saw any words appearing immediately before any of the letter strings. Data from 10 participants who reported that they saw such words were excluded from the analysis.

5.8.2.3 Stimuli

We designed 52 ambiguous letters using the Processing programming environment (Reas & Fry, 2007). Each ambiguous letter was created by morphing one letter into a different letter until it looked, in the eyes of the designer, equally similar to the two letters. Each ambiguous letter was embedded in a 3-letter string that could either form or not form a word depending on which letter is perceived. To counteract the contextual effect of the word on perception of the ambiguous letter, ambiguous letters were then slightly morphed toward the letter that does

not form a word (paralleling the assignment of a greater weight to input for that letter in the model). Ambiguous letters were positioned in either the beginning or the end of the letter string (1st or 3rd letter), whereas participants were directed to fixate at the center. This ensured that the distance between the ambiguous letters and the focus of gaze remained constant throughout the experiment, while allowing variability in the location of the ambiguous letter. The words that letter strings could form were all medium-to-high frequency words (above 10 per million; Kucera & Francis, 1967) picked using the MRC Psycholinguistic Database (http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm).

To prime the words that ambiguous letters could form, we used semantically related words, three to seven letters long. To avoid shape-related priming effects, prime words included neither of the two letters that the ambiguous letter resembled, nor other visually confounding letters (e.g., due to visual resemblance F could favor perception of E).

To ensure that participants were paying attention to all three letters of each string and not just to the 'funny looking' letter, we designed 36 additional 3-letter strings in which the letter that participants were asked to identify was not a morphed letter. One of the non-target letters in each such string was somewhat morphed, though not toward any other English letter.

To maximize the ambiguity of the ambiguous letters, we conducted a preliminary experiment, the results of which were used to adjust ambiguous letters so that they appeared equally similar to the word-forming and non-word-forming letters. Four participants performed the task described below. Subsequently, every ambiguous letter that was perceived as one particular letter at least 80% of the time was slightly morphed toward the other letter. This process was iterated 6 times prior to the main experiment, and was reasonably successful, as evidenced by the fact that in the main experiment the word-forming letter was chosen in 51.6% of the trials

in which one of the two relevant letters was chosen (standard deviation across ambiguous letters: 16.5%).

To minimize luminance-related changes in pupil diameter, all stimuli were adjusted to be isoluminant with the background using the flicker-fusion procedure (Lambert et al., 2003) on the display system used in the experiment.

5.8.2.4 Pupillometry

An ASL Series 5000 remote optics eye tracker (Applied Science Laboratories, MA) was used to measure participants' left pupil diameter while they were performing the task. At the beginning of the experiment, a baseline measurement of pupil diameter at rest was taken for a period of 45 s. Pupil-diameter data were processed in MATLAB to detect and remove blinks and other artifacts. For each trial, baseline pupil diameter was computed as the average diameter over a period of 1 s prior to the beginning of the trial (at the end of the inter-trial interval, at which point pupil activity from the trial itself should have subsided). Pupil-dilation response was computed as the difference between the peak diameter recorded during the 4 s that followed the beginning of the trial and the preceding baseline diameter. All pupil dilation responses were normalized by the pre-experiment baseline pupil diameter. Data from 6 participants, who had fewer than 20 trials in which at least half of the baseline pupil diameter and pupil response measurements were free of artifacts, were excluded from the analysis.

5.8.3 Recognition memory experiment

5.8.3.1 Participants

45 participants (mean age 19.8, age range 18-22, 28 females) performed the recognition memory experiment, and received course credit for participation. Participants were Princeton University students who gave written informed consent before taking part in the study, which was approved by the university's institutional review board.

5.8.3.2 Experimental task

To test whether the relationship between neural gain and stimulus processing is also affected by the task-related, *voluntary* direction of attention, participants performed a word recognition memory experiment, in which the learning phase required attention to either visual or semantic aspects of the words (Graf & Ryan, 1990). Participants were presented with 72 words in one of two highly dissimilar fonts, each for a period of 2 s. Half of the words were coupled with a task that focused participants' attention on word shape. Specifically, participants were asked to rate how readable the word was on a scale of 1 (very hard to read) to 4 (very easy to read). The other half of the words were coupled with a semantic task that required processing both a word's shape (so as to read it) and its meaning. Specifically, participants reported for each word whether it refers to something that would exist without humans (for example, trees) or would not exist without humans (for example, buildings). Words were divided into 4 blocks of 18 words, each of which was associated with one of the tasks. Task order was counterbalanced both within and between participants. In order to mitigate primacy and recency effects, each block started and ended with 4 words that were not included in the recognition memory test.

Participants could also indicate that they were not able to read the word, and such words (on average 1.0 ±0.18 words per block) were excluded from further analysis. Words were separated by a (uniformly) random inter-trial interval of 7 s to 9 s. Following an average period of 19.0 ±0.18 minutes, during which participants performed an unrelated roommate decision-making task, we held a word recognition memory test in which half of the words were foils, a quarter of the words had previously appeared in the same font (in the readability or semantic task), and a quarter of the words had previously appeared in a different font. Recognition memory performance was computed in line with signal detection theory as d' (Stanislaw & Todorov, 1999). Performance could not be reliably quantified in participants that reached ceiling performance or that did not recognize any of the target words, and thus data from such participants (2 in the readability task and 16 in the semantic task) was excluded from further analysis. Importantly, excluded participants did not differ in mean pupillary response from the average participant (9.4% ±0.9% vs. 9.3% ±0.5%, $t_{43} = 0.15$, $p = 0.88$). Since we hypothesized that high gain is associated with more selective processing, we expected that recognition memory would be more strongly degraded by font change in participants whose pupillary responses indicated high gain. However, we only expected to see this effect of gain for words from the readability task, which specifically required processing of word shape, and not for words from the semantic task, which required processing of both word shape and meaning.

5.8.3.3 Stimuli

176 words, each 5 to 7 letters long, of medium-to-high frequency (above 10 per million; Kučera & Francis, 1967) were randomly assigned for each participant to different blocks or used as foils. Words were presented using an isoluminant color in capital letters in one of two fonts,

Old English Text MT or Matura MT Script, which were chosen since they are highly dissimilar.

5.8.3.4 Pupillometry

A desk-mounted SMI RED 120Hz eye-tracker (SensoMotoric Instruments Inc., MA) was used to measure participants' left and right pupil diameters at a rate of 60 samples per second while they were performing the experiment with their head fixed on a chinrest. Pupil diameter data were processed as for the ambiguous letters experiment. Mean pupil dilation response was computed separately for the readability and the semantic tasks.

5.8.3.5 Statistical analysis

Analyses were carried out using MATLAB. All correlation values reported are Pearson correlation coefficients. Averaging of correlation coefficients was preceded by Fisher r-to-z transformation and followed by Fisher's z-to-r transformation, so as to mitigate the problem of the non-additivity of correlation coefficients (Fisher, 1921). Group-level significance of within-participant correlations was computed using a one-tailed one-sample Student's t-test on the vector of correlation coefficients following Fisher r-to-z transformation. Significance of Pearson correlation coefficients was computed using the Student's t-distribution. Interactions between pupil response and experimental conditions were computed using ANCOVA. All statistical tests were two tailed.

Chapter 6

Neural gain and decision making biases

If neural gain affects how we process information, it is bound to have a substantial impact on the decisions we make. For instance, gain may affect the degree to which our decisions are susceptible to being biased. The results presented in the previous chapter suggest that high gain is associated with greater susceptibility to a perceptual bias induced by subliminal priming. However, in that specific case priming biased perception by focusing it on a particular feature of the stimulus. Thus, a more focused mode of processing was conducive in that case to a stronger bias. In contrast, in more complex decisions, for instance those that involve personal preference, biases are thought to emerge from the integration of information (Usher at al., 2013; Busemeyer et al., 2006). If that is the case, high gain, by limiting integration, could lead to decisions that are *less* biased. Here we test the relationship between neural gain, as indexed by pupillometry, and biases in a variety of decision making scenarios.

6.1 Introduction

In some well-described scenarios, human decision making exhibits systematic deviations from rational behavior. For instance, a particular action could be more or less likely to be chosen

depending on how it is framed, even though the information provided is in both cases equivalent (Tversky et al., 1981). However, the classic decision making biases that have been described in the literature are typically induced by peripheral aspects of the decision problem, exerting a relatively weak effect that is only detectable when the behavior of dozens or even hundreds of participants is averaged (e.g., Levin et al., 1998; Kühberger, 1998). Based on our findings, we hypothesize that such biases only manifest in decision makers that process information in a broad, integrative manner, which takes into account both central and peripheral aspects of the problem. Indeed, recent theoretical work suggests that many decision biases, such as sensitivity to framing and other contextual and attentional effects, arise from the gradual integration of information that takes place as the different aspects of a decision problem are examined (Usher at al., 2013; Busemeyer et al., 2006; Krajbich & Rangel, 2011). Accordingly, models involving gradual integration of information have been used to explain many of the peculiarities that characterize human decision making (Usher & McClelland, 2004; Busemeyer & Townsend, 1993; Diederich, 1997; Roe et al., 2001; Johnson & Busemeyer, 2005).

We have seen that pupil diameter indices of locus coeruleus-norepinephrine function and neural gain (Servan-Schreiber et al., 1990; Aston-Jones & Cohen, 2005) track the degree to which information processing is narrowly focused on the most strongly represented stimulus features, or conversely, is broadly integrative of both weakly and strongly represented features. An increase in gain can be thought of as an increase in contrast between weakly and strongly active neural units that further focuses information processing on the strongest representations (Figure 1.2). Thus, if the manifestation of decision making biases depends on the integration of weakly-represented aspects of the problem, low gain should be associated with more robust biases, whereas high gain may diminish them.

6.2 Anchoring

One of the simplest decision biases, which directly arises from integration of information over time, is that of the anchoring of estimations to arbitrary values considered in preceding questions (Tversky & Kahneman, 1974). In line with previous studies of anchoring, we asked participants to indicate whether seven different quantities (e.g., the height of the Eiffel tower) was higher or lower than some arbitrary value, and then estimate the quantity (Jacowitz & Kahneman, 1995). Anchoring was measured as the degree to which a participant's estimation deviated towards the arbitrary value that the participant was asked to consider, as compared to other participants' estimations. In addition, we used pupil dilation in response to task stimuli as an inverse index of gain.

We divided participants into tertiles of low, medium and high mean pupil dilation, and computed the mean anchoring effect for each group. All groups of participants exhibited a significant anchoring effect, regardless of pupillary response (low: $t_{12} = 3.7$, $p < 0.005$; medium: $t_{13} = 3.3$, $p < 0.01$; high: $t_{12} = 4.2$, $p < 0.005$; Figure 6.1), and the trend towards stronger anchoring with higher pupillary response (indicating lower gain) was not significant (low vs. high: $t_{24} = 0.81$, $p = 0.43$).

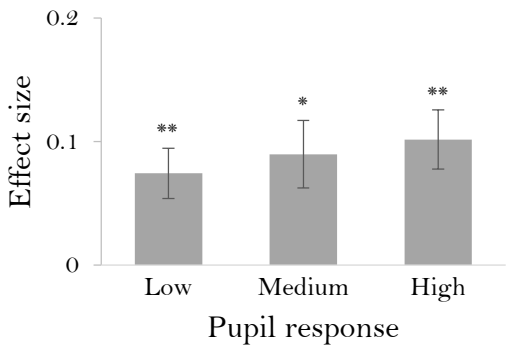

Figure 6.1. Anchoring effect. Deviation of participants' estimates towards the arbitrary anchors which they were asked to consider. Estimates were normalized to the range of 0 to 1. n = 40 participants, *: $p < 0.01$, **: $p < 0.005$, error bars: across-participant s.e.m.

6.3 Persistence of belief

The persistence of beliefs formed early in an experiment in the face of contradictory evidence that is presented later (Peterson & DuCharme, 1967) is thought to arise from the effect of previously gathered information on the perception of new information (Lord et al., 1979). Therefore, this bias too may depend on integration of information over time. To test persistence of belief, we presented participants with a series of colored balls while asking them which of two urns the balls are more likely to be coming from. The two urns differed in the proportion of balls of each color, and thus, in the probability of being the source of the series of balls (Figure 6.2A). The order of the balls presented was predetermined so as to initially favor one urn (first 30 balls), and then the other (last 60 balls). We quantified persistence of belief bias by the degree to which participants continued to favor the initially favored urn during the second part of the sequence. Only participants with high pupillary responses (indicating low gain) continued to prefer the initially supported urn ($t_{11} = 2.8$, $p < 0.05$). In contrast, participants with low pupillary responses updated their estimates to a similar extent in the first and second part of the experiment ($t_{22} = 2.4$, $p < 0.05$; Figure 6.2B).

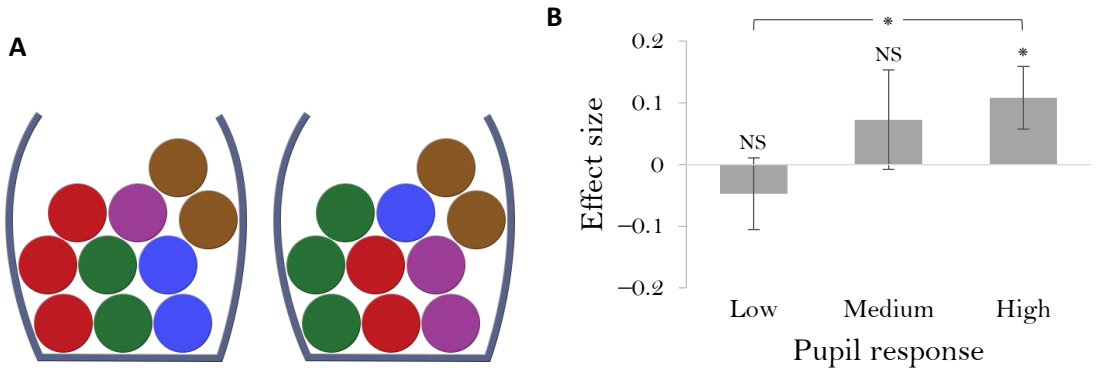

Figure 6.2. Persistence of belief. (A) The two urns contained different proportions of balls of different colors. (B) Preference of the initially supported urn during the last 60 balls, which were suggestive of the other urn. An optimal observer would be indifferent on average. Preferences were mapped to a scale between -1 and 1. $n = 35$ participants, NS: $p > 0.1$, *: $p < 0.05$, error bars: across-participant s.e.m.

6.4 Framing

Many decision biases arise not from the temporal structure of the problem, but rather from the integration of its multiple attributes. We tested two types of biases that characterize decision problems involving the evaluation of items with multiple attributes: framing effects and sample-size neglect. Considered by many as a prime example of irrational behavior, "framing effects" refer to the often-replicated finding that logically equivalent descriptions of a problem can lead to systematically different decisions (Levin et al., 1998). We tested three different types of framing effects previously reported in the literature: attribute framing, risky choice framing and task framing.

6.4.1 Attribute framing

Participants evaluated items of three different types (ground beef, student exam performance and gambles), whose attributes were framed either positively or negatively (Levin et al., 1985).

For example, student exam performance could be described in terms of % correct (positive frame) or % incorrect (negative frame). Positive framing invoked significantly higher evaluations than negative framing only in participants with high pupillary responses (indicating low gain; $t_{13} = 2.17$, $p < 0.05$; Figure 6.3).

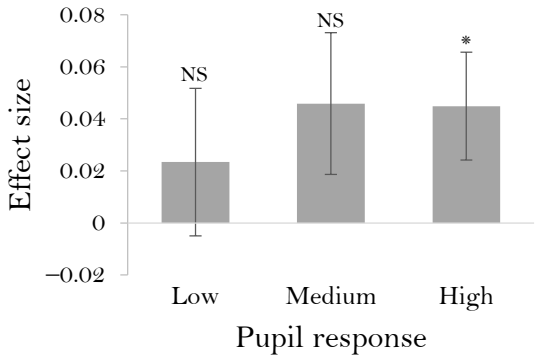

Figure 6.3. Attribute framing effect. Difference in evaluation of items framed positively rather than negatively. Item rating were mapped to a scale between 0 and 1. $n = 43$ participants, NS: $p > 0.1$, *: $p < 0.05$, error bars: across-participant s.e.m.

6.4.2 Risky choice framing

In this task, participants chose between a certain and an uncertain outcome, both framed either as gains or as losses (Tversky et al., 1981; Van Schie & Van Der Pligt, 1995). For example, the outcome of a treatment program could be described as '200 people (out of 600) will be saved' or as '400 people (out of 600) will die'. This manipulation builds on people's previously-established tendency to be risk averse in the domain of gains, but risk seeking in the domain of losses (Kahneman & Tversky, 1979). Framing outcomes as gains rather than losses evoked more risk averse preferences only in participants with high pupillary responses ($t_{13} = 2.34$, $p < 0.05$; Figure 6.4).

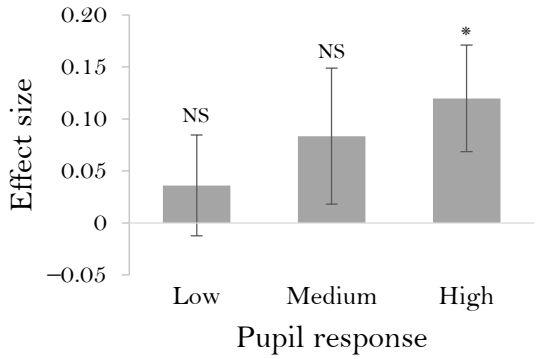

Figure 6.4. Risky choice framing effect. Increase in risk aversion when outcomes were described in terms of gain rather than losses. Preferences were mapped to a scale between -1 and 1. $n = 42$ participants, NS: $p > 0.2$, *: $p < 0.05$, error bars: across-participant s.e.m.

6.4.3 Task framing

Participants were asked to either accept or reject one of two options. One option – the enriched option – had more positive as well as more negative dimensions than the other, impoverished, option (Shafir, 1993). It has been shown that people are biased to select the enriched option regardless of whether they are accepting it or rejecting it, presumably because the enriched option provides good reasons to do either. Participants with medium or high pupillary responses, but not with low pupillary responses, showed a significant preference for the enriched option across the "accept" and "reject" tasks (medium: $t_{13} = 3.5$, $p < 0.005$; high: $t_{13} = 2.2$, $p < 0.05$; Figure 6.5).

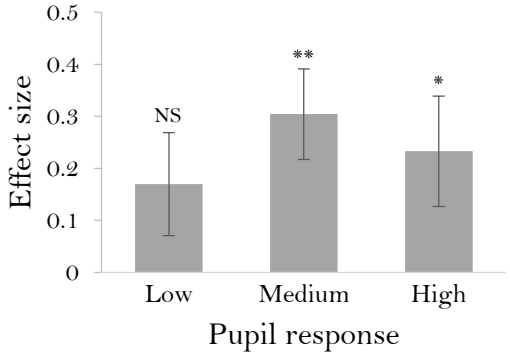

Figure 6.5. Task framing effect. Preference to both accept and reject the enriched option more than the impoverished option. Preferences were mapped to a scale between -1 and 1. $n = 42$ participants, NS: $p > 0.1$, *: $p < 0.05$, **: $p < 0.005$, error bars: across-participant s.e.m.

6.5 Sample-size neglect

A different bias that may result from multi-attribute integration is evident in people trying to determine whether a coin is biased to heads or tails, given the number of head and tail outcomes. People's certainty about the coin's bias is typically determined by the ratio between heads and tails (Griffin & Tversky, 1992). However, ratios such as 9 to 7 or 6 to 5 are unlikely to be computed precisely, and thus, the decision process has to involve the integration of both raw numbers. In contrast, an optimal judge can simply reduce the numbers of heads and tails to the difference between them, which is easily computable, and then make a decision based on this single attribute. This simpler, single-attribute strategy is optimal since it does not fail to take into account the sample size (i.e., the total number of outcomes; see Methods).

Thus, we asked participants how certain they were that a coin was biased in favor of heads given different sets of outcomes. Participants with medium and high pupil responses exhibited significant sample-size neglect (medium: $t_{12} = 6.3$, $p < 10^{-5}$; high: $t_{11} = 4.1$, $p < 0.005$), whereas those with low pupil responses exhibited only a trend-level effect ($t_{11} = 2.1$, $p = 0.06$) that was significantly weaker than in the other groups (vs. medium: $t_{23} = 2.5$, $p < 0.05$; vs. high: $t_{22} = 2.6$, $p < 0.05$; Figure 6.6). Accordingly, responses of participants with low pupil responses reflected precise inference more than responses in the other groups (vs. medium: $t_{23} = 2.7$, $p < 0.05$; vs. high: $t_{22} = 3.5$, $p < 0.005$). In addition, the *difference* between heads and tails predicted the estimates of participants with low pupil responses better than the *ratio* between heads and tails (beta difference 0.56 ±0.13, $t_{11} = 4.3$, $p < 0.005$), but this was not true for the estimates of participants with medium (beta difference 0.10 ±0.09, $t_{12} = 1.1$, $p = 0.31$; difference from low group: $t_{23} = 3.0$, $p < 0.01$) and high (beta difference -0.04 ±0.14, $t_{11} = -0.2$, $p = 0.81$; difference

from low group: $t_{22} = 3.1$, $p < 0.005$) pupil responses, indicating that only low-pupil-response participants primarily relied on the difference, not the ratio, between heads and tails.

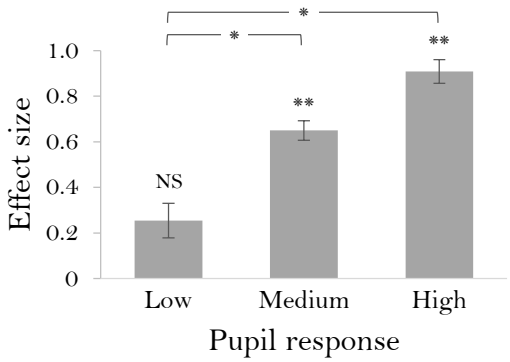

Figure 6.6. Sample-size neglect. Measured as the overweighting of the ratio between heads and tails relative to the weight given to the optimal inferences. $n = 37$ participants, NS: $p > 0.05$, *: $p < 0.05$, **: $p < 0.005$, error bars: across-participant s.e.m.

6.6 Overall susceptibility to biases

Although differences in decision making biases between participants with low and high pupil responses were, for the most part, not statistically significant, they were highly consistent – in each one of the six tasks, participants with high pupil responses were more strongly biased. Thus, we compared the average normalized effect size across all experiments using a permutation test. Participants with low pupillary responses were significantly less biased overall (Figure 6.7), suggesting that pupillary response indexed general susceptibility to decision making biases in our experiments.

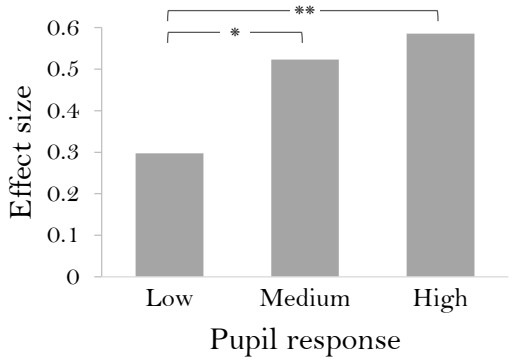

Figure 6.7. Overall susceptibility to biases. Average normalized effect size across all experiments. The values 0 and 1 correspond to the minimal and maximal effect sizes exhibited by any selection of 14 participants in each experiment. $n = 44$ participants, *: $p < 0.01$, **: $p < 0.0005$, permutation test.

6.7 Computational model

We next used Usher & McClelland's (2004) decision model (Figure 6.8A) to illustrate how high gain may weaken the expression of a bias (e.g., framing effect) in a multi-alternative, multi-attribute decision problem, using. Consider a choice between two items, one attribute of which favors the first item and a second attribute favors the second item. The model assumes that on each time step, one of the attributes is selected at random, and the evidence it provides is accumulated at the decision layer. A framing bias can be implemented in the model either as a tendency to select one of the attributes more frequently, or as a selective increase in the strength of evidence provided by one of the attributes. Either way, over a large number of time steps even a small bias may consistently determine the result of the decision process. In contrast, with a low number of time steps, the decision would be determined by whichever attribute happened to be selected more often so far. Increasing gain strengthens the effect of evidence on the decision units, and consequently, fewer time steps are required to reach a decision. As a result, the bias is diminished (Figure 6.8B). With minor modification, this model may explain the weakening of any of the three framing effects tested here. We also note that

our specific choice of decision model is inconsequential, since similar results should be obtained using any decision model, as long as it relies on a gradual integration or diffusion process.

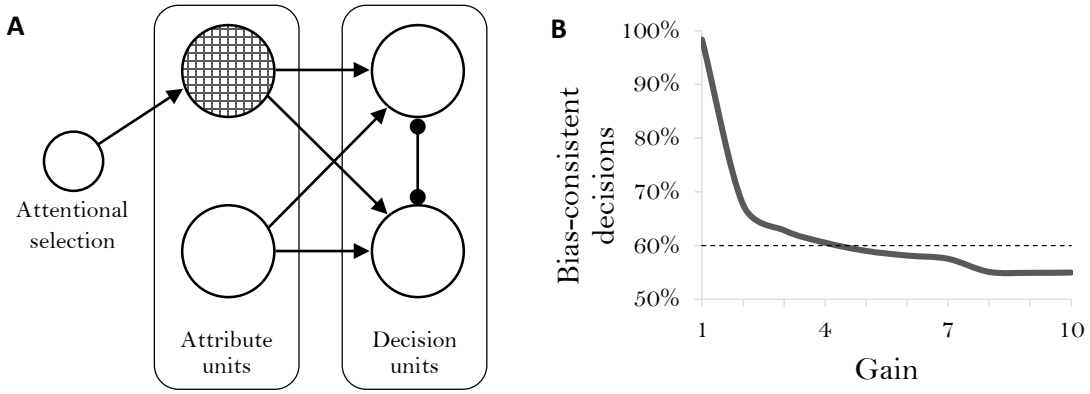

Figure 6.8. A model of the effect of gain on the manifestation of a decision bias. (A) Structure of the model. On each time step, one attribute is randomly selected, and the evidence in favor of each item is accumulated by the competing decision units. A bias was implemented as a tendency to select one of the attributes more frequently. (B) Proportion of bias-consistent decisions made by the model as a function of gain. With higher gain decisions are less biased. The decision process was simulated 100,000 times with each level of gain. The dashed line indicates the proportion that would be needed to detect a statistically-significant bias given a sample of 100 decisions (binomial test).

6.8 The cost of weaker biases

While high gain seems to be associated with weaker biases, this should come at the cost of limited integration, leading to potential decisions that are based on fewer samples and are thus less certain. Uncertainty concerning a potential decision is thought to affect the likelihood of executing the decision (Daw et al., 2005), and thus we may expect that high gain would be associated with a higher likelihood of indecision (e.g., expressing indifference between available options). Indeed, in the multi-alternative decision problems, participants with low pupil

responses decided not to decide more often than participants with high pupil responses (Figure 6.9A; $p < 0.05$, permutation test).

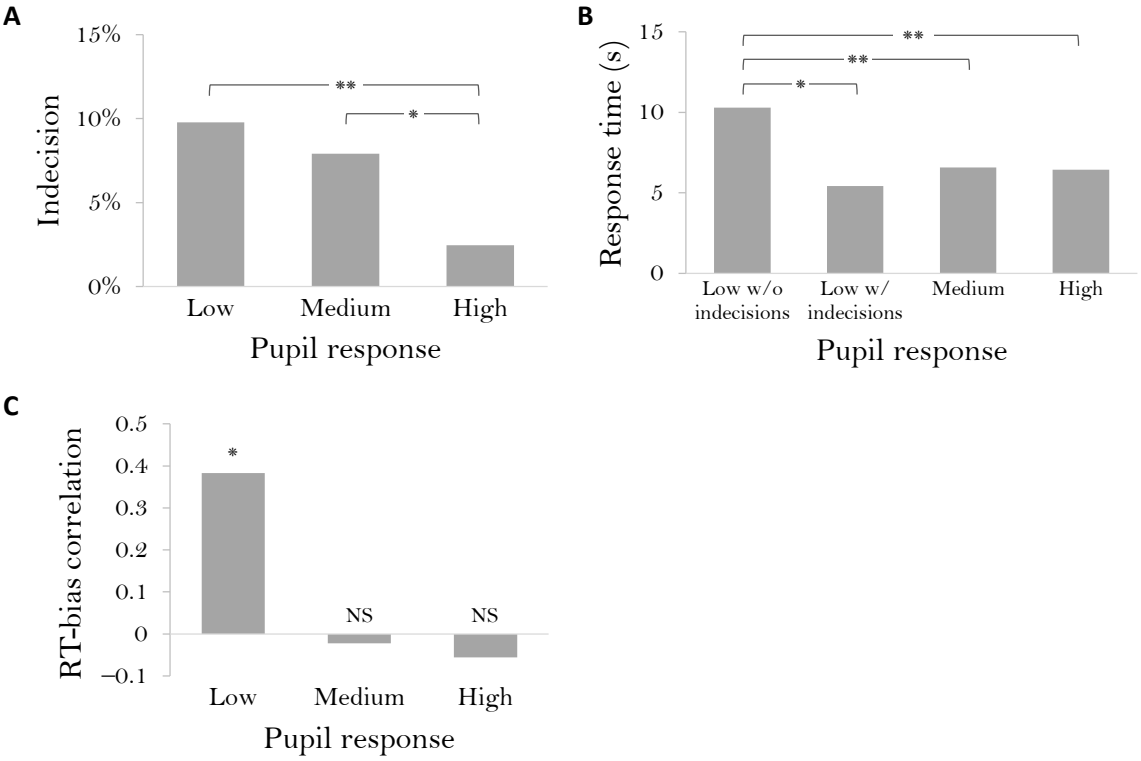

Figure 6.9. Indecision and response time (RT). (A) Proportion of decisions expressing indifference between available options. Data include experiments that involved choice between two alternatives (i.e., risky choice framing, task framing and persistence of belief experiments). (B) Response time in two-alternative decision problems in which decision time was measureable (i.e., risky choice framing, task framing). The low pupil response group was further divided into participants that exhibited indecisions and those that did not. (C) Mean across-participant correlation between average response time and bias effect size. Data include experiments in which the source of bias was continuously available (i.e., the framing and sample-size neglect experiments). $n = 44$ participants, NS: $p > 0.2$, *: $p < 0.05$, **: $p < 0.005$, permutation test.

Alternatively, failure to reach a certain, integrative decision may induce a deliberative, time-consuming process whose purpose is to increase decision certainty (Glöckner & Betsch, 2008; Daw et al., 2005). In line with this view, participants with low pupillary responses that avoided indecision took more time to make decisions than those that did not avoid indecision ($p < 0.05$,

permutation test), as well as than those with stronger pupillary responses ($p < 0.005$, Figure 6.9B). Moreover, taking more time seems to have restored some level of evidence integration, since in participants with low pupillary responses longer response times were associated with stronger biases ($p < 0.05$, permutation test; Figure 6.9C).

6.9 Discussion

We investigated the relationship between a pupillary index of neural gain and individual differences in decision making biases. Low pupil responses, which are consistent with high gain, were associated with weaker biases across six different tasks involving integration of information over time and over problem attributes. We used a computational model to illustrate how the latter type of biases may be diminished with high gain as a result of diminished integration. Diminished biases, however, came at the cost of indecisiveness, or alternatively, longer deliberation time. In contrast, participants with high pupil responses showed remarkably consistent biases.

Typically, to demonstrate a decision making bias, dozens or even hundreds of participants are needed. It is thus notable that a small group of participants, at most 15 in number, consistently exhibited statistically-significant biases across six different experiments. This underscores the role of information-processing mode, broadly integrative or narrowly focused, in explaining inter-individual differences in decision making.

Throughout this chapter, we adopted the view that decision biases emerge from a process of integration. While this view is supported by various experimental and theoretical works (Usher at al., 2013; Busemeyer et al., 2006), it is by no means fully established. Moreover, different

biases could arise from different mechanisms. Our findings, however, lend further support to the role of integration in a diverse set of decision making biases, by showing that susceptibility to biases can be predicted by a pupillary index of gain, which we have previously linked to behavioral and neural markers of integration.

Our findings may seem to suggest that high gain is generally associated with more optimal decisions. However, classic decision making biases, such as the ones tested here, are specifically designed to exploit people's tendency to integrate irrelevant cues into the decision process. In more complex and real-life like tasks, in which integration of information is paramount (Usher et al., 2011; Rusou et al., 2013), we predict that high gain would in fact be associated with poorer decision making.

6.10 Appendix: Methods

6.10.1 Experimental methodology

6.10.1.1 Participants

44 Princeton University students (mean age 19.5, age range 18-23, 28 females) performed the experiment. Participants gave written informed consent before taking part in the study, which was approved by the university's institutional review board. Participants received course credit for participation.

6.10.1.2 Stimuli

Stimuli were generated using the Processing programming environment (Reas & Fry, 2007). To minimize luminance-related changes in pupil diameter, stimuli were made isoluminant with the background by adjusting their colors using the flicker-fusion procedure (Lambert et al., 2003) on the display system that was used in the experiment. Stimuli were presented on a computer screen using MATLAB software (MathWorks) and the Psychophysics Toolbox (Brainard, 1997).

6.10.1.3 Anchoring experiment

Participants answered two questions about each of 7 quantities (e.g., the height of the Eiffel tower). They first indicated whether the quantity was greater or less than an anchor value. Next, they estimated the quantity. Each quantity was coupled with a low anchor for half of the participants and with a high anchor for the other half. Each participant was presented with a low anchor for half (3 or 4) of the quantities, and with a high anchor for the other half. Quantities and calibrated anchor values were taken from a previous study (Jacowitz & Kahneman, 1995), including: length of the Mississippi river, population of Chicago, number of babies born per day in the US, height of mount Everest, pounds of meat an American eats per day, year the telephone was invented, and maximum speed of a house cat. Anchoring effect was quantified by the deviation of an estimate from the group mean estimate in the direction of the anchor, normalized to the group estimates' range. Data from 3 participants whose estimates were clear outliers (i.e., whose distance from others' estimates was more than ten times the range of others' estimates) and 1 participant with fewer than two valid (i.e., mostly artifact free) pupil response measurements were excluded from the analysis.

6.10.1.4 Persistence of belief experiment

Participants were presented with two urns filled with colored balls (Figure 6.2A), and with a sequence of 90 balls, which they were told were sampled with replacement from one of the urns (Peterson & DuCharme, 1967). Every 5 balls, participants indicated using a sliding bar which urn they thought the sequence was sampled from. The precise position of the bar indicated degree of certainty. The first 3 participants performed a preliminary version of the experiment in which they responded after every ball. This was changed to make the experiment faster and more engaging, and thus, 41 participants performed the final version of the experiment. One urn contained 3 red balls, 2 green balls, 2 blue balls, 2 brown balls and 1 purple ball, and the other urn contained 2 red balls, 3 green balls, 1 blue ball, 2 brown balls and 2 purple balls. The sequence of balls was set up so that the first 30 balls favored one of the urns as their source with a probability of 0.95, and the next 60 balls favored the other urn to a similar degree (per 30 balls). Therefore, it was optimal to favor one urn after 30 balls, be indifferent after 60 balls, and favor the other urn after 90 balls (Figure 6.10). Accordingly, an optimal observer would be indifferent on average during the last 60 balls. Thus, persistence-of-belief effect was quantified by the degree to which participants' average response during the last 60 balls favored the initially-favored urn. The initially-favored urn was counterbalanced between participants. Data from 4 participants who did not favor the correct urn during the first 30 balls and 2 participants with fewer than two valid pupil response measurements were excluded from the analysis.

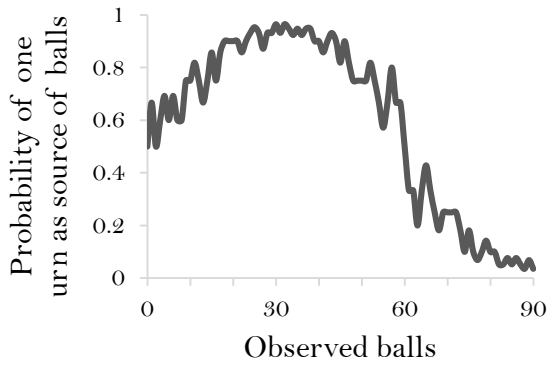

Figure 6.10. Probability of one urn being the source of the sequence of balls as the sequence progressed, determined by the relative likelihood of each of the balls coming out of the urn, given the contents of both urns.

6.10.1.5 Attribute framing experiment

Participants used a sliding bar to rate ground beef products, gambles, and students' performance, whose attributes were framed either positively or negatively (Levin et al., 1985). In the ground beef task, participants were asked to imagine that they were having a friend over for dinner and they were about to make their favorite lasagna dish with ground beef. They were then asked to rate how satisfied they would be purchasing each of 4 ground beef products, described in terms of price per pound ($2.7 and $3.3), and either percentage lean (80% and 90%, positive frame) or percentage fat (20% and 10%, negative frame). In the gambles task, participants were asked to imagine that they started out with $10 and they can either keep the $10 and not play the gamble or pay the $10 to take the gamble. They were then asked to rate how likely they were to take each of 3 gambles, described in terms of amount to be won ($50, $100 and $200) and either probability of wining (20%, 10% and 5%, positive frame) or probability of losing (80%, 90% and 95%, negative frame). In the student performance task, participants were asked to evaluate each of 2 students on the basis of midterm exam and final exam performance, described in terms of either % correct (50% and 70%, positive frame) or % incorrect (50% and 30%, negative frame). Each item was framed positively in half of the

participants, and negatively in the other half. For a given participant, all items of a particular type were similarly framed (i.e., either positively or negatively), so as to minimize awareness of the framing manipulation, but framing was varied within participants across item types. Framing effect was quantified for each item type by the deviation of a participant's mean rating from the group mean rating in the direction of the frame (i.e., upwards for positive frames, and downwards for negative frames). Data from 1 participant with fewer than two valid pupil response measurements were excluded from the analysis.

6.10.1.6 Risky choice framing experiment

Participants faced two different scenarios, a medical scenario and a fire scenario, and indicated using a sliding bar which of two available actions they would choose in each scenario. One action had a certain outcome and the other an uncertain outcome, both of which were framed in terms of either gains or losses. Scenarios were described in full as done previously (Van Schie & Van Der Pligt, 1995). In the medical scenario, which concerned the treatment of a deadly disease at an island inhabited with 600 inhabitant, participants chose between the gain-framed outcomes '300 people will be saved' and 'a 50% chance that 600 people will be saved and a 50% chance that none of the people will be saved', or between the loss-framed outcomes '300 people will die' and 'a 50% chance that 600 people will die and a 50% chance that none of the people will die'. In the fire scenario, which concerned the treatment of fires threatening 9000 acres of forest, participants chose between the gain-framed outcomes '3000 acres of forest will be saved' and 'a 60% chance that 5000 acres will be saved and a 40% chance that no forest under threat will be saved', or between the loss-framed outcomes '6000 acres of forest will be lost' and 'a 60% chance that 4000 acres will be lost and a 40% chance that 9000 acres will be lost'. Framing

effect was quantified as the deviation of a participant's preferences from the group mean in the direction of the frame (i.e., towards the certain outcome in the gain frame, and towards the uncertain option in loss frame). Data from 2 participants with no valid pupil response measurements were excluded from the analysis.

6.10.1.7 Task framing experiment

Participants faced 5 different problems, concerning various subjects such as child custody, vacation choice, ice-cream choice and gambling. Each problem involved one option that had more positive and negative attributes (the "enriched" option) and one option that had fewer positive and negative attributes (the "impoverished" option). In each problem, half of the participants were asked to choose one of the options, and the other half were asked to reject one of the options. For example, in one problem participants were asked to imagine that they served on the jury of an only-child sole-custody case following a relatively messy divorce, and they decided to base their decision entirely on the following few observations. Parent A: average income, average health, average working hours, reasonable rapport with the child, relatively stable social life. Parent B: above-average income, very close relationship with the child, extremely active social life, lots of work-related travel, minor health problems. Half of the participants were asked to which parent they would award sole custody of the child, while the other half were asked which parent they would deny sole custody of the child. Framing effect was quantified by the degree to which across tasks (i.e., award and reject) participants preferred the enriched option (i.e., Parent A) more frequently than the impoverished option (i.e., Parent B). Full description of the other problems can be found elsewhere (Shafir, 1993; problems 1, 2,

4, 5 and 6). Data from 2 participants with fewer than two valid pupil response measurements were excluded from the analysis.

6.10.1.8 Sample-size neglect experiment

Participants were told to imagine that they were spinning a biased coin, and they recording how often the coin landed heads and how often the coin landed tails. They knew that the coin tended to land on one side 3 out of 5 times, but they did not know if this bias is in favor of heads or in favor of tails. Participants were then presented with 10 different sets of results (number of heads and number of tails), in which the heads always outnumbered the tails, and they indicated using a sliding bar how certain they were given each set that the coin was biased in favor of heads. Sets of results were similar to those used previously (Griffin & Tversky, 1992).

The probability that the coin was biased in favor of heads was inferred as:

$$p(H|D) = e^{(h-t)\log\frac{3}{2}} \qquad (6.1)$$

where h is the number of heads and t is the number of tails. This expression is equivalent to

$$p(H|D) = e^{n\frac{(h-t)}{n}\log\frac{3}{2}} \qquad (6.2)$$

which depends on the sample size (i.e., the number of outcomes, n) and on the ratio between heads and tails ($\frac{h-t}{n}$). It was previously found that people tend to overweigh the ratio component at the expense of the sample size component (sample-size neglect; Griffin & Tversky, 1992). To measure this bias, we regressed participants' estimates against the real probabilities (Eq. 6.1), and then regressed the residuals against the ratio component alone

($e^{\frac{h-t}{n}\log\frac{3}{2}}$). The resulting regression coefficients captured overweighing of the ratio component at the expense of the sample size. In addition, to test whether the difference between heads and tails explained participants' estimates better than the ratio between them, we reversed the steps. That is, we first regressed participants' estimates against the ratio component, and then the residuals against the real probabilities, which reflect the difference between heads and tails (Eq. 6.1). We then compared the resulting regression coefficients to the coefficients produced when the regressions were performed in the reverse order. All inputs to regression analyses were z scored so as to produce normalized coefficients. 7 participants who were more certain that the coin was biased in favor of heads given 3 heads and 2 tails, than given 7 heads and 2 tails, were excluded from the analysis, as we suspected that they mistakenly looked for a ratio that best matched 3 to 2.

6.10.1.9 Eye tracking

A desk-mounted SMI RED 120Hz eye-tracker (SensoMotoric Instruments Inc., MA) was used to measure participants' left and right pupil diameters at a rate of 60 samples per second while they were performing the behavioral tasks with their head fixed on a chinrest. At the beginning of the experiment, a baseline measurement of pupil diameter at rest was taken for a period of 45 s. Pupil-diameter data were processed in MATLAB to detect and remove blinks and other artifacts. For each trial, baseline pupil diameter was computed as the average diameter over a period of 1 s prior to the beginning of the trial (at the end of the inter-trial interval, at which point pupil activity from the trial itself should have subsided). Pupil-dilation response was computed as the difference between the peak diameter recorded during the 4 s that followed the beginning of the trial and the preceding baseline diameter. All pupil dilation responses were

normalized by the pre-experiment baseline pupil diameter. Pupil dilation responses in which more than half of the measurements were affected by artifacts were considered invalid and excluded from the analysis.

6.10.1.10 Statistical analysis

Analyses were carried out using MATLAB. Permutation tests were performed by sampling 100,000 random permutations of the coupling between pupillary and behavioral individual data sets. Results based on the permuted data served as null distributions to which actual results were compared. Correlation values reported are Spearman correlation coefficients. Averaging of correlation coefficients was preceded by Fisher *r*-to-*z* transformation and followed by Fisher's *z*-to-*r* transformation, so as to mitigate the problem of the non-additivity of correlation coefficients (Fisher, 1921). All statistical tests were two tailed.

6.10.2 Computational model

We modeled decision between two items, each with two attributes, using a leaky competing accumulator model (Usher & McClelland, 2001). The model consisted of two competing accumulators, one for each item (Figure 6.8A). Every time step, activity a_i of accumulator i was updated to reflect evidence in favor of the respective item by:

$$\Delta a_i = 0.1\left(-a_i + g(I_i - \left|a_j\right|^+) + \epsilon\right) \qquad (6.3)$$

where g reflects the level of gain, I_i is the evidence-based excitatory input to accumulator i, j is the competing accumulator whose positive component ($|\,|^+$) provided inhibitory input, and ϵ is normally-distributed noise with a standard deviation of 0.01. As in previous models of multi-

attribute multi-item decisions (Usher & McClelland, 2004), on each time step, one attribute was selected at random, and excitatory input was determined accordingly. One of the attributes favored one item, and thus generated input of 1.25 to one accumulator and 0.75 to the other accumulator (plus normally-distributed random noise with standard deviation of 0.01). The other attribute favored the other item, and thus generated the same input but reversed. A decision was reached once one of the accumulators reached a value of 1. A bias was implemented by either setting the likelihood of selecting one of the attributes to 0.55 (instead of 0.5), or by increasing the input to one of the accumulators by 0.05. The two implementations gave similar results, and thus only the results of the first are shown. We conducted 1,000,000 simulations with each level of gain between 1 and 10.

Chapter 7

Theoretical and practical implications

In this chapter, I will attempt to integrate the different results presented in this thesis into a coherent Bayesian perspective on the effects of neural gain. I will then argue that high levels of gain may provide a more complete account of the behavioral and biological features of autism than previous theories. Finally, I will highlight some of the questions that are left open and potential future directions.

7.1 A Bayesian perspective

I started this thesis by examining neural gain at the cellular level, and from there proceeded to explore, via a set of mechanistic network models, system-level neural and behavioral effects of gain. Thus, in term of Marr's (1982) levels of explanation, my approach proceeded from the implementational to the algorithmic. We will now take another step up Marr's ladder, and attempt to understand the effects of gain from a computational, normative Bayesian perspective.

In Chapter 4, we saw that increased gain focuses learning on stimulus features to which one is predisposed to attend. In that context, individual predisposition can be understood as a prior on

the types of features that could be relevant. Such a prior is necessary to guide learning in a multidimensional environment, in which exploration of all features is infeasible (Wilson & Niv, 2011). Thus, the results of the learning experiment may lead us to suggest that high gain is associated with stronger (i.e., narrower) priors. However, in later chapters we encountered results that could lead us to the opposite view. In 0, we saw that high gain was associated with a weaker effect of prior semantic knowledge on perception of ambiguous letters (in the no-priming condition). In Chapter 6, we saw that high gain was associated with prior information having a weaker biasing effect (in the anchoring and persistence-of-belief tasks). Both of these results suggest that high gain is associated with weaker priors. Thus, the effects of gain on priors were not consistent across experiments.

We can gain a better understanding of the computational implications of high gain if we start from the algorithmic level. At that level, the effects of high gain all seem to reflect reduced representational breadth, and thus, reduced integration. Breadth and integration, however, are fundamental to Bayesian inference, which requires representation of probability distributions in order to integrate incoming evidence with prior expectations. Thus, the effects of high gain can be understood as narrowing distributions – that is, overweighing high probabilities events at the expense of low probabilities – and thus compromising the correct integration of evidence and priors. This could amount, in some cases, to overweighing evidence (Figure 7.1A) and in other cases to overweighing priors (Figure 7.1B), depending on which is stronger (i.e., more precise) in a specific situation.

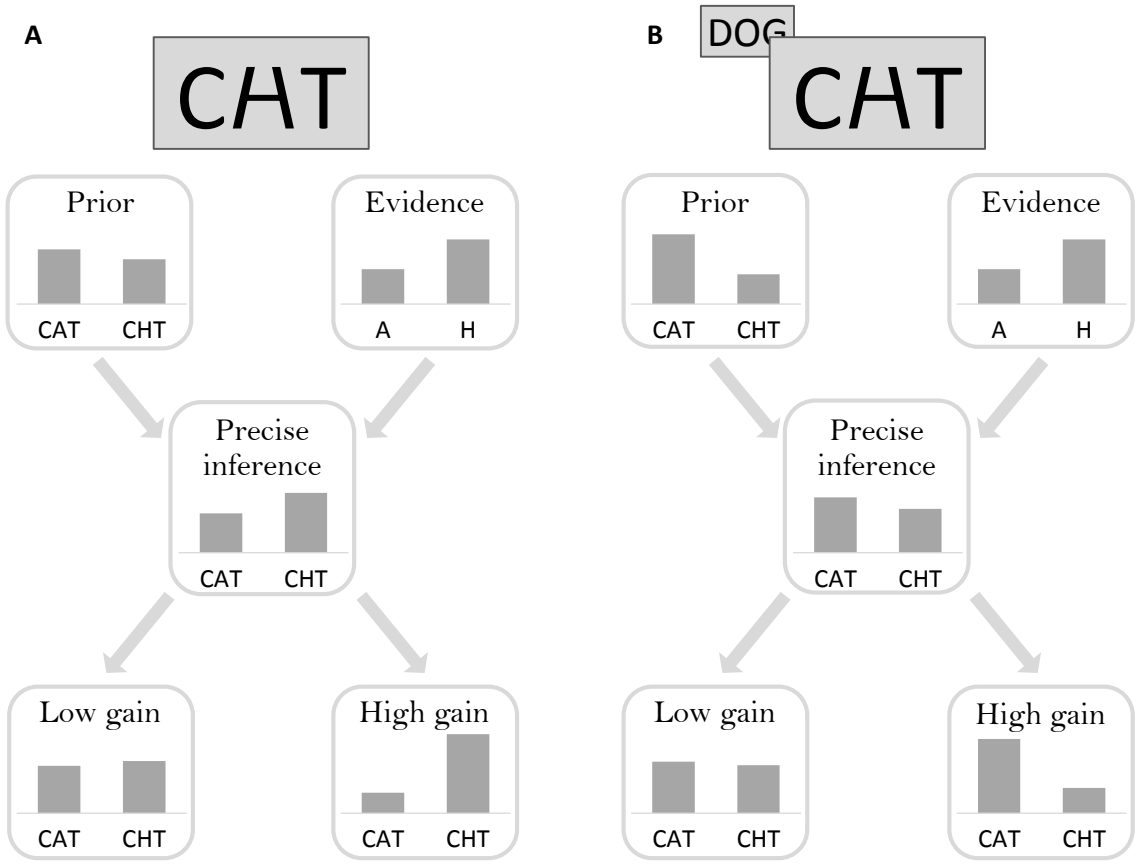

Figure 7.1. Integration of prior expectation and evidence, with low and high gain. (A) The visual evidence in favor of H is stronger than the prior expectation to see A in between C and T. Thus, precise inference favors H, and this preference is amplified by high gain and reduced by low gain. In this case, high gain results in overweighing the evidence. (B) Semantic priming strengthens the prior expectation to see CAT, which is now stronger than the evidence. Thus, precise inference favors A, and this preference too is amplified by high gain and reduced by low gain. In this case, high gain results in overweigh the prior. Note that low levels of gain also compromise precise inference, but in favor of the less probable option.

It could be argued that variations in gain do not, in fact, compromise inference, but rather, gain is optimally adjusted so as to produce precise inference. For instance, in a highly predictable environment in which few potential outcomes are probable, inference should yield narrow distributions, and thus high gain may be suitable. Conversely, in a less predictable environment, low gain may allow the representation of the larger number of probable outcomes. Such a view, however, would be hard to reconcile with the considerable inter-

individual differences that we observed in measures of gain and behavior in a simple perceptual task. Instead, we propose that variations in gain do lead to deviations from optimal inference, but these deviations are in fact advantageous in certain circumstances. Specifically, in stressful situations, which are known to activate the LC-NE system (Korf et al., 1973; Abercrombie & Jacobs, 1987) and thus presumably increase gain, immediate coherent action is often required. In such situations, it could be disadvantageous to consider the multiple low-probability possibilities that optimal inference typically involves. Rather, it is often best to focus one's resources on the single, most immediately-relevant possibility. Conversely, in low-stake situations, when there is no pressing need for action, and no significant outcome is imminent, it can be advantageous to explore a wider range of possibilities, including less probable ones, thus gaining information about the environment and improving one's ability to maximize utility (Kaelbling et al., 1996). In such typically low-arousal circumstances, low gain may enhance the representation of low-probability possibilities and thereby facilitate their exploration. Thus, while variations in gain likely compromise optimal inference, they may help optimize utility.

7.2 A neural gain account of autism

Our account of the effect of gain on the integration between prior and evidence points to a potential solution to a long-standing controversy concerning the key cognitive dysfunction underlying autism spectrum disorders (ASD). Early cognitive theories of autism, such as the weak central coherence theory (Frith & Happé, 1994), the theory of mind theory (Baron-Cohen et al., 1985), and the executive dysfunction theory (Hughes et al., 1994), helped conceptualize the wide-range of behavioral findings in autism in terms of few fundamental characteristics. These theories, however, often relied on imprecise theoretical concepts, were focused on

particular clusters of symptoms, and did not provide mechanistic explanations for the emergence of the deficits that they posited.

More recently, the discussion shifted towards Bayesian formulations of the problem. Weaker priors (Pellicano & Burr, 2012), or alternatively, stronger evidence (Brock, 2012), have been suggested to underlie perceptual atypicalities in autism such as more accurate perception (Ropar & Mitchell, 2002), failure to use prior information (Becchio et al., 2010), and the sense of being overwhelmed by sensory information (Bogdašina, 2005). However, autistic individuals sometimes show inflexible and perseverating behavior (Ciesielski & Harris, 1997), which implies excessively strong priors. In addition, they do not have a problem forming priors in a simple sensory oddball task (Ferry et al., 2003; Kujala et al., 2007). This lead Van de Cruys et al. (2013) to propose a predictive coding (Bastos et al., 2012) account of autism, according to which the disorder is caused by chronically strong sensory prediction errors (i.e., strong evidence), which lead to the formation of priors that are too strong to be applicable in all but the simplest cases.

The predictive coding hypothesis is appealing since it predicts overweighting of priors in simple cases, and underweighting of priors in more complex circumstances, as seems to be the case in autism. However, it falls short on two accounts. First, if evidence signals are uniformly strong then multiple stimuli presented simultaneously should all elicit an equally strong response (or if signals are normalized, an equally weak response), and should thus be processed equally, in a manner that is even more balanced than in typically developing individuals, at least in those cases in which priors are inapplicable. Evidence, however, suggests the opposite. Autistic individuals spontaneously attend to fewer cues than typically developing individuals (Lovaas et al., 1971; Liss et al., 2006; Ciesielski & Harris, 1997). Even in a simple, highly

structured task, autistic participants did not benefit from congruent information simultaneously provided by two dimensions of a stimulus, in contrast to typically developing participants (Plaisted et al., 1999). Autistic children could only learn to discriminate multiple-cue complexes when taught to use a sequential strategy (Koegel & Schreibman, 1977). In fact, a recent meta-analysis found that the most consistent sensory symptom in autism is not hyper-responsivity, but rather reduced responsivity, which presumably reflects selective attention directed elsewhere (Ben-Sasson et al., 2009). Thus, over-selective attention seems to be a primary feature of autism, and it is not consistent with the predictive coding account.

In addition, the predictive coding account, like previous theories, fails to provide a parsimonious mechanistic explanation for the dysfunction that it proposes. In predictive coding, selective attention is implemented by the modulation of sensory prediction error signals, so that error signals evoked by attended stimuli are enhanced, and error signals evoked by unattended stimuli are inhibited (Van de Cruys et al., 2013). Thus, since autistic individuals are capable of attending selectively, they must be capable (according to the predictive coding account) of modulating prediction errors signals. The dysfunction in autism, then, cannot be in a primary mechanism that is required for modulation of error signals, but rather, it has to be in a system that determines when, where and to what extent error signals are modulated, or in other words, in the system that controls attention. However, areas that are involved in regulating attention, mostly located in frontal and parietal lobes (Corbetta & Shulman, 2002), are thought to be involved in multiple functions, and thus, isolated dysfunction in the modulation of attention seems unlikely. Moreover, there is little evidence that attention areas are structurally or biochemically distinct from nearby cortical areas. Thus, the prospect of

identifying biological factors that specifically affect the attention system does not seem promising.

I argue that high neural gain provides a more parsimonious and complete account of the symptomatology of autism, and moreover, it is tightly linked to biological factors that are known to be affected in autism. The high gain account of autism subsumes the key predictions made by the predictive coding account, including stronger evidence and stronger priors, either of which could dominate in any given case. However, in the high gain account, these predictions are inherently coupled with reduced attentional breadth and an inability to integrate information, as shown throughout this thesis. Thus, high gain naturally explains autistic individuals' over-selective attention, as well as their difficulties with multisensory integration (Smith & Bennetto, 2007), integration of context (Wang et al., 2006), and integration of prior knowledge (Becchio et al., 2010). In addition, since high gain is associated with more locally-clustered functional connectivity, it can explain the stronger local and weaker global connectivity that is thought to develop in autism (Courchesne & Pierce, 2005; Barttfeld et al., 2011), as well as the association between autism and focal epilepsy (Levisohn, 2007).

In keeping with my proposal, autism has already been associated with abnormalities that may increase brain-wide levels of gain. LC-NE activity, which modulates neural gain throughout the brain, has been suggested to play a central role in the pathophysiology of autism, due to the role that the LC plays in generating fever and the ameliorating effect fever has in autism (Mehler & Purpura, 2009). Moreover, drugs that oppose the effect of NE, such as β-adrenergic antagonists and α_2 agonists, have been shown to improve cognitive and social functioning in autistic individuals (Beversdorf et al., 2008; Beversdorf et al., 2011; Ming et al., 2008; Ratey et

al., 1987). Autism has also been associated with increases in tonic pupil diameter (Anderson & Colombo, 2009), skin conductance (Hirstein et al., 2001), heart rate (Ming et al., 2005), blood pressure (Ming et al., 2005), and respiratory rate (Zahn et al., 1987), all of which could reflect increased LC-NE activity. Finally, many of the genetic abnormalities that underlie susceptibility to autism are involved in synaptic growth, and thus, may have a direct impact on neural gain (Bourgeron, 2009).

In sum, I argue that a neural gain account of autism improves on previous theories by providing a computationally precise theory that explains a wider range of the behavioral and biological characteristics of the disorder.

7.3 Open questions

Previous findings suggested that the effects of NE follow an inverted-U-shaped relationship, both at the cellular level (Devilbiss & Waterhouse, 2000) and at the level of behavior (Baldi & Bucherelli, 2005). In contrast, in our experiments pupillary indices of LC-NE function were monotonically related to behavioral and neural measures. However, due to the sluggish pupillary response, all of our experiments involved relatively long inter-trial intervals that presumably induced boredom and decreased arousal. Thus, it is possible that participants' LC-NE activity levels were relatively low, and thus, the results only reflected the first half of the inverted U. Indeed, the effects of NE reuptake inhibitors was opposite to that associated with pupillary indices of high gain. It is unclear however whether the difference between the pharmacological and pupillary results reflected differences in NE levels, or other differences that exist between pharmacological and physiological NE stimulation. Further research is

necessary to clarify whether the effects that we found to be correlated with pupillary indices of gain follow an inverted U-shaped relationship.

Finally, throughout this work, we used fMRI to investigate the whole-brain effects of gain on neural activity and connectivity. However, the BOLD signal does not exclusively reflect neural activity (Maier et al., 2008; Sirotin & Das, 2009; Logothetis, 2008), and it may be affected by any factor that affects blood flow. This is especially concerning because NE is known to affect cerebral arteries (Toda & Fujita, 1973). Thus, the imaging work presented here should be complemented by methods that do not rely on measurements of blood flow. For instance, electrocorticography would be particularly suitable for replicating our connectivity analyses due to its high spatial and temporal resolution.

7.4 Future directions

7.4.1 Information processing styles

The effects of neural gain in our experiments mostly manifested in differences between individuals in the way in which they processed information. These differences may reflect in part stable tendencies of particular individuals to utilize different modes of information processing to different extents. Such information processing styles could arise from genetic variations in NE signaling and metabolism, as well as from differences in genes that affect synaptic efficacy, such as those that have been implicated in autism. As a first step, to test for the existence of stable information processing styles, the same participants can be re-tested at different times and in different tasks, for instance using the tasks presented in this theses.

7.4.2 Dynamic interactions between gain and information processing

Although gain may be determined in part by stable individual traits, we have seen that it also varies within participants over the course of an experiment. This opens up the possibility for exploring the dynamic interactions between gain and information processing. For example, in a risky situation some individuals may become more stressed because they pay more attention to potential negative outcomes than to potential positive outcomes. Stress may then lead to higher LC-NE activity which should increase gain and thus make these individuals more narrowly focused on the potential negative outcomes. This, in turn, can make them even more stressed, and thus a vicious cycle may ensue.

7.5 Conclusion

I have shown that the idea of gain modulation, although originally derived from the study of single neurons, and measured here indirectly using pupillometry, successfully accounts for a wide range of system-level neural and behavioral phenomena. Neurally, gain modulation can explain brain-wide changes in fMRI activity and connectivity. Specifically, our findings suggest that increased gain is associated with a shift from a globally distributed mode of neural communication to a locally clustered mode. Paralleling this shift in neural function is a behavioral shift, from a broad, integrative mode of information processing to a narrowly focused mode. This behavioral effect, which we explored in four different experiments, was manifested in the way participants perceived and remembered visual input, learned from outcomes, and made decisions. The close association of the brain's gain control system with the "fight-or-flight" sympathetic system suggests a normative explanation for the effect of gain

modulation: in high-stake situations, arousal and gain increase, thus focusing resources on the most immediately-relevant possibility. Conversely, in low-stake situations, low gain may facilitate attention to, and exploration of, a wider range of possibilities. Finally, the neural, behavioral and computational effects of high gain examined here suggest a novel, parsimonious account of many of the disturbances found in autism spectrum disorders.

Bibliography

Alexander, J. K., Hillier, A., Smith, R. M., Tivarus, M. E., & Beversdorf, D. Q. Beta-adrenergic modulation of cognitive flexibility during stress. *Journal of Cognitive Neuroscience* 19, 468-478 (2007).

Andersen, R. A., Essick, G. K., & Siegel, R. M. Encoding of spatial location by posterior parietal neurons. *Science* 230, 456-458 (1985).

Anderson, C. J., & Colombo, J. Larger tonic pupil size in young children with autism spectrum disorder. *Developmental psychobiology* 51, 207-211 (2009).

Aston-Jones, G., & Bloom, F. E. Activity of norepinephrine-containing locus coeruleus neurons in behaving rats anticipates fluctuations in the sleep-waking cycle. *The Journal of Neuroscience* 1, 876-886 (1981a).

Aston-Jones, G., & Bloom, F. E. Nonrepinephrine-containing locus coeruleus neurons in behaving rats exhibit pronounced responses to non-noxious environmental stimuli. *The Journal of Neuroscience* 1, 887-900 (1981b).

Aston-Jones, G., & Cohen, J. D. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annual Review of Neuroscience* 28, 403-450 (2005).

Aston-Jones, G., Foote, S. L., & Bloom, F. E. Anatomy and physiology of locus coeruleus neurons: functional implications. *Frontiers of clinical neuroscience* 2, 92-116 (1984).

Aston-Jones, G., Foote, S. L., & Segal, M. Impulse conduction properties of noradrenergic locus coeruleus axons projecting to monkey cerebrocortex. *Neuroscience* 15, 765-777 (1985).

Aston-Jones, G., Rajkowski, J., Kubiak, P., & Alexinsky, T. Locus coeruleus neurons in monkey are selectively activated by attended cues in a vigilance task. *The Journal of Neuroscience* 14, 4467-4480 (1994).

Aston-Jones, G., Rajkowski, J., & Kubiak, P. Conditioned responses of monkey locus coeruleus neurons anticipate acquisition of discriminative behavior in a vigilance task. *Neuroscience* 80, 697-715 (1997).

Aston-Jones, G., Shipley, M. T., & Grzanna, R. The locus coeruleus, A5 and A7 noradrenergic cell groups. *The rat nervous system* 2, 183-213 (1995).

Baldi, E., & Bucherelli, C. The inverted" u-shaped" dose-effect relationships in learning and memory: modulation of arousal and consolidation. *Nonlinearity in Biology, Toxicology, and Medicine* 3, 9-21 (2005).

Baron-Cohen, S., Leslie, A. M., & Frith, U. Does the autistic child have a "theory of mind"? *Cognition* 21, 37-46 (1985).

Barttfeld, P., Wicker, B., Cukier, S., Navarta, S., Lew, S., & Sigman, M. A big-world network in ASD: dynamical connectivity analysis reflects a deficit in long-range connections and an excess of short-range connections. *Neuropsychologia* 49, 254-263 (2011).

Bassett, D. S., Wymbs, N. F., Porter, M. A., Mucha, P. J., Carlson, J. M., & Grafton, S. T. Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences* 108, 7641-7646 (2011).

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. Canonical microcircuits for predictive coding. *Neuron* 76, 695-711 (2012).

Beaudet, A., & Descarries, L. The monoamine innervation of rat cerebral cortex: synaptic and nonsynaptic axon terminals. *Neuroscience* 3, 851-860 (1978).

Becchio, C., Mari, M., & Castiello, U. Perception of shadows in children with autism spectrum disorders. *PloS one* 5, e10582 (2010).

Ben-Sasson, A., Hen, L., Fluss, R., Cermak, S. A., Engel-Yeger, B., & Gal, E. A meta-analysis of sensory modulation symptoms in individuals with autism spectrum disorders. *Journal of autism and developmental disorders* 39, 1-11 (2009).

Beversdorf, D. Q., Carpenter, A. L., Miller, R. F., Cios, J. S., & Hillier, A. Effect of propranolol on verbal problem solving in autism spectrum disorder. *Neurocase* 14, 378-383 (2008).

Beversdorf, D. Q., Saklayen, S., Higgins, K. F., Bodner, K. E., Kanne, S. M., & Christ, S. E. Effect of propranolol on word fluency in autism. *Cognitive and Behavioral Neurology* 24, 11-17 (2011).

Bogdašina, O. *Communication issues in autism and Asperger syndrome: do we speak the same language?* London, UK: Jessica Kingsley Publishers (2005).

Bönisch, H., & Brüss, M. The norepinephrine transporter in physiology and disease. In *Neurotransmitter Transporters* (pp. 485-524). Berlin, Heidelberg, Germany: Springer (2006).

Bouret, S., & Sara, S. J. Reward expectation, orientation of attention and locus coeruleus-medial frontal cortex interplay during learning. *European Journal of Neuroscience* 20, 791-802 (2004).

Bourgeron, T. A synaptic trek to autism. *Current opinion in neurobiology* 19, 231-234 (2009).

Brainard, D. H. The psychophysics toolbox. *Spatial vision* 10, 433-436 (1997).

Brock, J. Alternative Bayesian accounts of autistic perception: comment on Pellicano and Burr. *Trends in cognitive sciences* 16, 573-574 (2012).

Bullmore, E., & Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience* 10, 186-198 (2009).

Bullmore, E., & Sporns, O. The economy of brain network organization. *Nature Reviews Neuroscience* 13, 336-349 (2012).

Busemeyer, J. R., Jessup, R. K., Johnson, J. G., & Townsend, J. T. Building bridges between neural models and complex decision making behaviour. *Neural Networks* 19, 1047-1058 (2006).

Busemeyer, J. R., & Townsend, J. T. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review* 100, 432 (1993).

Campbell, H. L., Tivarus, M. E., Hillier, A., & Beversdorf, D. Q. Increased task difficulty results in greater impact of noradrenergic modulation of cognitive flexibility. *Pharmacology Biochemistry and Behavior* 88, 222-229 (2008).

Ciesielski, K. T., & Harris, R. J. Factors related to performance failure on executive tasks in autism. *Child Neuropsychology* 3, 1-12 (1997).

Chance, F. S., Abbott, L. F., & Reyes, A. D. Gain modulation from background synaptic input. *Neuron* 35, 773-782 (2002).

Clayton, E. C., Rajkowski, J., Cohen, J. D., & Aston-Jones, G. Phasic activation of monkey locus ceruleus neurons by simple decisions in a forced-choice task. *The Journal of neuroscience* 24, 9914-9920 (2004).

Coffield, F., Moseley, D., Hall, E., & Ecclestone, K. *Learning styles and pedagogy in post-16 learning: A systematic and critical review.* London, UK: Learning and Skills Research Centre (2004).

Cohen, J. D., Dunbar, K., & McClelland, J. L. On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological review* 97, 332 (1990).

Corbetta, M., & Shulman, G. L. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience* 3, 201-215 (2002).

Courchesne, E., & Pierce, K. Why the frontal cortex in autism might be talking only to itself: local over-connectivity but long-distance disconnection. *Current opinion in neurobiology* 15, 225-230 (2005).

Cragg, S. J., & Rice, M. E. DAncing past the DAT at a DA synapse. *Trends in neurosciences* 27, 270-277 (2004).

Dayan, P., & Yu, A. Norepinephrine and neural interrupts. *Advances in neural information processing systems* 18, 243 (2006).

Devilbiss, D. M., & Waterhouse, B. D. Norepinephrine exhibits two distinct profiles of action on sensory cortical neuron responses to excitatory synaptic stimuli. *Synapse* 37, 273-282 (2000).

Devilbiss, D. M., & Waterhouse, B. D. The effects of tonic locus ceruleus output on sensory-evoked responses of ventral posterior medial thalamic and barrel field cortical neurons in the awake rat. *The Journal of neuroscience* 24, 10773-10785 (2004).

Dias-Ferreira, E., Sousa, J. C., Melo, I., Morgado, P., Mesquita, A. R., Cerqueira, J. J., Costa, R. M., & Sousa, N. Chronic stress causes frontostriatal reorganization and affects decision-making. *Science* 325, 621-625 (2009).

Diederich, A. Dynamic stochastic models for decision making under time constraints. *Journal of Mathematical Psychology* 41, 260-274 (1997).

Easterbrook, J. A. The effect of emotion on cue utilization and the organization of behavior. *Psychological review* 66, 183 (1959).

Eguíluz, V. M., Chialvo, D. R., Cecchi, G. A., Baliki, M. & Apkarian, A.V. Scale-free brain functional networks. *Physical Review Letters*. 94, 018102 (2005).

Einhäuser, W., Stout, J., Koch, C., & Carter, O. Pupil dilation reflects perceptual selection and predicts subsequent stability in perceptual rivalry. *Proceedings of the National Academy of Sciences* 105, 1704-1709 (2008).

Elam, M., Yao, T., & Svensson, T. H. Hypercapnia and hypoxia: chemoreceptor-mediated control of locus coeruleus neurons and splanchnic, sympathetic nerves. *Brain research* 222, 373-381 (1981).

Elam, M., Yoa, T., Svensson, T. H., & Thoren, P. Regulation of locus coeruleus neurons and splanchnic, sympathetic nerves by cardiovascular afferents. *Brain research* 290, 281-287 (1984).

Eldar, E., Cohen, J. D., & Niv, Y. The effects of neural gain on attention and learning. *Nature neuroscience* 16, 1146-1153 (2013).

Felder, R. M., & Silverman, L. K. Learning and teaching styles in engineering education. *Engineering education* 78, 674-681 (1988).

Felder, R. M., & Spurlin, J. Applications, reliability and validity of the index of learning styles. *International Journal of Engineering Education* 21, 103-112 (2005).

Ferri, R., Elia, M., Agarwal, N., Lanuzza, B., Musumeci, S. A., & Pennisi, G. The mismatch negativity and the P3a components of the auditory event-related potentials in autistic low-functioning subjects. *Clinical Neurophysiology* 114, 1671-1680 (2003).

Fisher, R. A. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* 1, 3-32 (1921).

Foote, S. L., Aston-Jones, G., & Bloom, F. E. Impulse activity of locus coeruleus neurons in awake rats and monkeys is a function of sensory stimulation and arousal. *Proceedings of the National Academy of Sciences* 77, 3033-3037 (1980).

Foote, S. L., Freedman, R., & Oliver, A. P. Effects of putative neurotransmitters on neuronal activity in monkey auditory cortex. *Brain research* 86, 229-242 (1975).

Frances, A., First, M. B., Pincus, H. A. *DSM-IV guidebook*. Washington, DC, American Psychiatric Press (1995).

Freedman, R., Hoffer, B. J., Woodward, D. J., & Puro, D. Interaction of norepinephrine with cerebellar activity evoked by mossy and climbing fibers. *Experimental neurology*, 55, 269-288 (1977).

Frith, U., & Happé, F. Autism: Beyond "theory of mind". *Cognition* 50, 115-132 (1994).

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience* 10, 252-269 (2010).

Grant, S. J., Aston-Jones, G., & Redmond D. E. J. Responses of primate locus coeruleus neurons to simple and complex sensory stimuli. *Brain research bulletin* 21, 401-410 (1988).

Griffin, D., & Tversky, A. The weighing of evidence and the determinants of confidence. *Cognitive psychology* 24, 411-435 (1992).

Haider, B., & McCormick, D. A. Rapid neocortical dynamics: cellular and network mechanisms. *Neuron* 62, 171-189 (2009).

Hajós, M., Fleishaker, J. C., Filipiak-Reisner, J. K., Brown, M. T., & Wong, E. H. The selective norepinephrine reuptake inhibitor antidepressant reboxetine: pharmacological and clinical profile. *CNS drug reviews* 10, 23-44 (2004).

Hasselmo, M. E., Linster, C., Patil, M., Ma, D., & Cekic, M. Noradrenergic suppression of synaptic transmission may influence cortical signal-to-noise ratio. *Journal of Neurophysiology* 77, 3326-3339 (1997).

Hill, S. A., Taylor, M. J., Harmer, C. J., & Cowen, P. J. Acute reboxetine administration increases plasma and salivary cortisol. *Journal of Psychopharmacology* 17, 273-275 (2003).

Hirstein, W., Iversen, P., & Ramachandran, V. S. Autonomic responses of autistic children to people and objects. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268, 1883-1888 (2001).

Hobson, J. A., McCarley, R. W., & Wyzinski, P. W. Sleep cycle oscillation: reciprocal discharge by two brainstem neuronal groups. *Science* 189, 55-58 (1975).

Hoeks, B., & Levelt, W. J. Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods, Instruments, & Computers* 25, 16-26 (1993).

Hoffer, B. J., Siggins, G. R., & Bloom, F. E. Studies on norepinephrine-containing afferents to Purkinje cells of rat cerebellum. II. Sensitivity of Purkinje cells to norepinephrine and related substances administered by microiontophoresis. *Brain Research* 25, 523-534 (1971).

Hoffer, B. J., Siggins, G. R., Oliver, A. P., & Bloom, F. E. Activation of the pathway from locus coeruleus to rat cerebellar purkinje neurons: pharmacological evidence of noradrenergic central inhibition. *Journal of Pharmacology and Experimental Therapeutics* 184, 553-569 (1973).

Hughes, C., Russell, J., & Robbins, T. W. Evidence for executive dysfunction in autism. *Neuropsychologia* 32, 477-492 (1994).

Hurley, L. M., Devilbiss, D. M., & Waterhouse, B. D. A matter of focus: monoaminergic modulation of stimulus coding in mammalian sensory networks. *Current opinion in neurobiology* 14, 488-495 (2004).

Jacowitz, K. E., & Kahneman, D. Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin* 21, 1161-1166 (1995).

Jepma, M., & Nieuwenhuis, S. Pupil diameter predicts changes in the exploration–exploitation trade-off: evidence for the adaptive gain theory. *Journal of cognitive neuroscience* 23, 1587-1596 (2011).

Johnson, J. G., & Busemeyer, J. R. A dynamic, stochastic, computational model of preference reversal phenomena. *Psychological Review* 112, 841 (2005).

Jones, B. E., Halaris, A. E., McIlhany, M., & Moore, R. Y. Ascending projections of the locus coeruleus in the rat. I. Axonal transport in central noradrenaline neurons. *Brain research* 127, 1-21 (1977).

Jones, B. E., & Moore, R. Y. Ascending projections of the locus coeruleus in the rat. II. Autoradiographic study. *Brain Research* 127, 23-53 (1977).

Jones, B. E., & Yang, T. Z. The efferent projections from the reticular formation and the locus coeruleus studied by anterograde and retrograde axonal transport in the rat. *Journal of Comparative Neurology* 242, 56-92 (1985).

Kaelbling, L. P., Littman, M. L., & Moore, A. W. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research* 4 (1996).

Kahneman, D., & Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society* 263-291 (1979).

Kitzbichler, M. G., Henson, R. N., Smith, M. L., Nathan, P. J., & Bullmore, E. T. Cognitive effort drives workspace configuration of human brain functional networks. *The Journal of Neuroscience* 31, 8259-8270 (2011).

Koss, M. C. Pupillary dilation as an index of central nervous system α_2-adrenoceptor activation. *Journal of pharmacological methods* 15, 1-19 (1986).

Krajbich, I., & Rangel. A. A multi-alternative drift diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *PNAS* 108,13852-13857 (2011).

Kucera, H., & Francis, W. N. *Computational Analysis of Present-Day American English.* Providence, RI: Brown University Press (1967).

Kühberger, A. The influence of framing on risky decisions: A meta-analysis. *Organizational behavior and human decision processes* 75, 23-55 (1998).

Kujala, T., Aho, E., Lepistö, T., Jansson-Verkasalo, E., Nieminen-von Wendt, T., von Wendt, L., & Näätänen, R. Atypical pattern of discriminating sound features in adults with Asperger syndrome as reflected by the mismatch negativity. *Biological psychology* 75, 109-114 (2007).

Lambert, A., Wells, I., & Kean, M. Do isoluminant color changes capture attention? *Perception & psychophysics* 65, 495-507 (2003).

Levisohn, P. M. The autism-epilepsy connection. *Epilepsia* 48, 33-35 (2007).

Levin, I. P., Johnson, R. D., Russo, C. P., & Deldin, P. J. Framing effects in judgment tasks with varying amounts of information. *Organizational Behavior and Human Decision Processes* 36, 362-377 (1985).

Levin, I. P., Schneider, S. L., & Gaeth, G. J. All frames are not created equal: A typology and critical analysis of framing effects. *Organizational behavior and human decision processes* 76, 149-188 (1998).

Liss, M., Saulnier, C., Fein, D., & Kinsbourne, M. Sensory and attention abnormalities in autistic spectrum disorders. *Autism* 10, 155-172 (2006).

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

Lord, C. G., Ross, L., & Lepper, M. R. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* 37, 2098 (1979).

Lovaas, O. I., Schreibman, L., Koegel, R., & Rehm, R. Selective responding by autistic children to multiple sensory input. *Journal of Abnormal Psychology* 77, 211 (1971).

Lucas, M. Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review* 7, 618-630 (2000).

Luce, R. & Perry, A. A method of matrix analysis of group structure. *Psychometrika* 14, 95-116 (1949).

Maier, A., Wilke, M., Aura, C., Zhu, C., Ye, F. Q., & Leopold, D. A. Divergence of fMRI and neural signals in V1 during perceptual suppression in the awake monkey. *Nature neuroscience* 11, 1193-1200 (2008).

Massaro, D. W. (1979). Letter information and orthographic context in word perception. *Journal of Experimental Psychology: Human Perception and Performance* 5, 595-609.

Massaro, D. W., & Cohen, M. M. (1991). Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology* 23, 558-614.

Marr, D. *Vision: A computational investigation into the human representation and processing of visual information.* New York, NY: Henry Holt and Co. Inc. (1982).

McAdams, C. J., & Maunsell, J. H. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *The Journal of Neuroscience* 19, 431-441 (1999).

McClelland, J. L., & Rumelhart, D. E. An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological review* 88, 375 (1981).

Mehler, M. F., & Purpura, D. P. Autism, fever, epigenetics and the locus coeruleus. *Brain research reviews* 59, 388-392 (2009).

Millan, M. J., Gobert, A., Lejeune, F., Newman-Tancredi, A., Rivet, J. M., Auclair, A., & Peglion, J. L. S33005, a novel ligand at both serotonin and norepinephrine transporters: I. Receptor binding, electrophysiological, and neurochemical profile in comparison with venlafaxine, reboxetine, citalopram, and clomipramine. *Journal of Pharmacology and Experimental Therapeutics* 298, 565-580 (2001).

Ming, X., Julu, P. O., Brimacombe, M., Connor, S., & Daniels, M. L. Reduced cardiac parasympathetic activity in children with autism. *Brain and Development* 27, 509-516 (2005).

Miyahara, S., & Oomura, Y. Inhibitory action of the ventral noradrenergic bundle on the lateral hypothalamic neurons through alpha-noradrenergic mechanisms in the rat. *Brain research* 234, 459-463 (1982).

Moises, H. C., Woodward, D. J., Hoffer, B. J., & Freedman, R. Interactions of norepinephrine with Purkinje cell responses to putative amino acid neurotransmitters applied by microiontophoresis. *Experimental neurology* 64, 493-515 (1979).

Moises, H. C., Waterhouse, B. D., & Woodward, D. J. Locus coeruleus stimulation potentiates Purkinje cell responses to afferent input: the climbing fiber system. *Brain research* 222, 43-64 (1981).

Moises, H. C., Waterhouse, B. D., & Woodward, D. J. Locus coeruleus stimulation potentiates local inhibitory processes in rat cerebellum. *Brain Research Bulletin* 10, 795-804 (1983).

Moises, H. C., & Woodward, D. J. Potentiation of GABA inhibitory action in cerebellum by locus coeruleus stimulation. *Brain Research* 182, 327-344 (1980).

Moore, R. Y., & Bloom, F. E. Central catecholamine neuron systems: anatomy and physiology of the norepinephrine and epinephrine systems. *Annual review of neuroscience* 2, 113-168 (1979).

Murphy, P. R., Robertson, I. H., Balsters, J. H., & O'connell, R. G. Pupillometry and P3 index the locus coeruleus–noradrenergic arousal function in humans. *Psychophysiology* 48, 1532-1543 (2011).

Nakai, Y., & Takaori, S. Influence of norepinephrine-containing neurons derived from the locus coeruleus on lateral geniculate neuronal activities of cats. *Brain research* 71, 47-60 (1974).

Necker, L. A. Observations on some remarkable optical phenomena seen in Switzerland; and on an optical phenomenon which occurs on viewing a figure of a crystal or geometrical solid. *The London and Edinburgh Philosophical Magazine and Journal of Science* 1, 329–337 (1832).

Nicol, R. M., Chapman, S. C., Vértes, P. E., Nathan, P. J., Smith, M. L., Shtyrov, Y., & Bullmore, E. T. Fast reconfiguration of high-frequency brain networks in response to surprising changes in auditory input. *Journal of Neurophysiology* 107, 1421-1430 (2012).

Olpe, H. R., Glatt, A., Laszlo, J., & Schellenberg, A. Some electrophysiological and pharmacological properties of the cortical, noradrenergic projection of the locus coeruleus in the rat. *Brain Research* 186, 9-19 (1980).

Paap, K. R., Newsome, S. L., McDonald, J. E., & Schvaneveldt, R. W. An activation–verification model for letter and word recognition: The word-superiority effect. *Psychological review* 89, 573 (1982).

Palmer, T. E. The effects of contextual scenes on the identification of objects. *Memory & Cognition* 3, 519-526 (1975).

Papadatou-Pastou, M., Miskowiak, K. W., Williams, J. M. G., Harmer, C. J., & Reinecke, A. Acute antidepressant drug administration and autobiographical memory recall: A functional magnetic resonance imaging study. *Experimental and Clinical Psychopharmacology* 20, 364 (2012).

Pellicano, E., & Burr, D. When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trends in cognitive sciences* 16, 504-510 (2012).

Pellicano, E., & Rhodes, G. Holistic processing of faces in preschool children and adults. *Psychological Science* 14, 618-622 (2003).

Peterson, C. R., & DuCharme, W. M. A primacy effect in subjective probability revision. *Journal of Experimental Psychology* 73, 61 (1967).

Phillis, J. W., Tebēcis, A. K., & York, D. H. The inhibitory action of monoamines on lateral geniculate neurones. *The Journal of physiology* 190, 563-581 (1967).

Plaisted, K., Swettenham, J., & Rees, L. Children with autism show local precedence in a divided attention task and global precedence in a selective attention task. *Journal of child psychology and psychiatry* 40, 733-742 (1999).

Rajkowski, J., Kubiak, P., & Aston-Jones, G. Correlations between locus coeruleus (LC) neural activity, pupil diameter and behavior in monkey support a role of LC in attention. *Society for Neuroscience Abstracts* 19, 974 (1993).

Rajkowski, J., Kubiak, P., Ivanova, S., & Aston-Jones, G. State-related activity, reactivity of locus ceruleus neurons in behaving monkeys. *Advances in Pharmacology* 42, 740-744 (1997).

Rasmussen, K., Morilak, D. A., & Jacobs, B. L. Single unit activity of locus coeruleus neurons in the freely moving cat: I. During naturalistic behaviors and in response to simple and complex stimuli. *Brain research* 371, 324-334 (1986).

Ratey, J. J., Bemporad, J., Sorgi, P., Bick, P., Polakoff, S., O'Driscoll, G., & Mikkelsen, E. Brief report: open trial effects of beta-blockers on speech and social behaviors in 8 autistic adults. *Journal of autism and developmental disorders* 17, 439-446 (1987).

Reas, C. & Fry, B. *Processing : a programming handbook for visual designers and artists*. Cambridge, MA: MIT Press (2007).

Reicher, G. M. Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of experimental psychology* 81, 275 (1969).

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological review* 108, 370 (2001).

Rogawski, M. A., & Aghajanian, G. K. Norepinephrine and serotonin: opposite effects on the activity of lateral geniculate neurons evoked by optic pathway stimulation. *Experimental neurology* 69, 678-694 (1980).

Ropar, D., & Mitchell, P. Shape constancy in autism: The role of prior knowledge and perspective cues. *Journal of Child Psychology and Psychiatry* 43, 647-653 (2002).

Rumelhart, D. E., & McClelland, J. L. *Parallel distributed processing: explorations in the microstructure of cognition. Volume 1. Foundations*. Cambridge, MA: MIT Press (1986).

Sacchetti, G., Bernini, M., Bianchetti, A., Parini, S., Invernizzi, R. W., & Samanin, R. Studies on the acute and chronic effects of reboxetine on extracellular noradrenaline and other monoamines in the rat brain. *British journal of pharmacology* 128, 1332-1338 (1999).

Salinas, E. Fast remapping of sensory stimuli onto motor actions on the basis of contextual modulation. *The Journal of neuroscience* 24, 1113-1118 (2004).

Salinas, E., & Abbott, L. F. Transfer of coded information from sensory to motor networks. *The journal of Neuroscience* 15, 6461-6474 (1995).

Salinas, E., & Bentley, N. M. (2009). Gain Modulation as a Mechanism for Switching Reference Frames, Tasks, and Targets. In *Coherent Behavior in Neuronal Networks* (pp. 121-142). Springer, New York (2009).

Salinas, E., & Thier, P. Gain modulation: a major computational principle of the central nervous system. *Neuron* 27, 15-21 (2000).

Sato, H., & Kayama, Y. Effects of noradrenaline applied iontophoretically on rat superior collicular neurons. *Brain research bulletin* 10, 453-457 (1983).

Schwabe, L., Höffken, O., Tegenthoff, M., & Wolf, O. T. Preventing the stress-induced shift from goal-directed to habit action with a β-adrenergic antagonist. *The Journal of Neuroscience* 31, 17317-17325 (2011).

Schwabe, L., Tegenthoff, M., Höffken, O., & Wolf, O. T. Concurrent glucocorticoid and noradrenergic activity shifts instrumental behavior from goal-directed to habitual control. *The Journal of Neuroscience* 30, 8190-8196 (2010).

Schwabe, L., & Wolf, O. T. Stress-induced modulation of instrumental behavior: from goal-directed to habitual control of action. *Behavioural brain research* 219, 321-328 (2011).

Segal, M., & Bloom, F. E. The action of norepinephrine in the rat hippocampus. I. Iontophoretic studies. *Brain Research* 72, 79-97 (1974a).

Segal, M., & Bloom, F. E. The action of norepinephrine in the rat hippocampus. II. Activation of the input pathway. *Brain Research* 72, 99-114 (1974b).

Segal, M., & Bloom, F. E. The action of norepinephrine in the rat hippocampus. IV. The effects of locus coeruleus stimulation on evoked hippocampal unit activity. *Brain Research* 107, 513-525 (1976).

Seguela, P., Watkins, K. C., Geffard, M., & Descarries, L. Noradrenaline axon terminals in adult rat neocortex: an immunocytochemical analysis in serial thin sections. *Neuroscience* 35, 249-264 (1990).

Servan-Schreiber, D., Printz, H., & Cohen, J. D. A network model of catecholamine effects: gain, signal-to-noise ratio, and behavior. *Science* 249, 892-895 (1990).

Servan-Schreiber D, Bruno R, Carter C & Cohen JD. Dopamine and the mechanisms of cognition. Part I: A neural network model predicting dopamine effects on selective attention. *Biological Psychiatry* 43, 713-722 (1998a).

Servan-Schreiber D, Carter C, Bruno R & Cohen JD. Dopamine and the mechanisms of cognition. Part II: D-Amphetamine effects in human subjects performing a selective attention task. *Biological Psychiatry* 43, 723-729 (1998b).

Sessler, F. M., Cheng, J. T., & Waterhouse, B. D. Electrophysiological actions of norepinephrine in rat lateral hypothalamus. I. Norepinephrine-induced modulation of LH

neuronal responsiveness to afferent synaptic inputs and putative neurotransmitters. *Brain research* 446, 77-89 (1988).

Shafir, E. Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition* 21, 546-556 (1993).

Sirotin, Y. B., & Das, A. Anticipatory haemodynamic signals in sensory cortex not predicted by local neuronal activity. *Nature* 457, 475-479 (2009).

Smith, E. G., & Bennetto, L. Audiovisual speech integration and lipreading in autism. *Journal of Child Psychology and Psychiatry* 48, 813-821 (2007).

Sompolinsky, H., Crisanti, A., & Sommers, H. J. Chaos in random neural networks. *Physical Review Letters* 61, 259-262 (1988).

Staal, M. A. Stress, cognition, and human performance: A literature review and conceptual framework. *NaSA technical memorandum, 212824* (2004).

Stone, T. W. Pharmacology of pyramidal tract cells in the cerebral cortex. *Naunyn-Schmiedeberg's Archives of Pharmacology* 278, 333-346 (1973).

Svensson, T. H. Peripheral, autonomic regulation of locus coeruleus noradrenergic neurons in brain: putative implications for psychiatry and psychopharmacology. *Psychopharmacology* 92, 1-7 (1987).

Szabo, S. T., & Blier, P. Effect of the selective noradrenergic reuptake inhibitor reboxetine on the firing activity of noradrenaline and serotonin neurons. *European Journal of Neuroscience* 13, 2077-2087 (2001).

Toda, N., & Fujita, Y. Responsiveness of isolated cerebral and peripheral arteries to serotonin, norepinephrine, and transmural electrical stimulation. *Circulation research* 33, 98-104 (1973).

Treue, S., & Trujillo, J. C. M. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* 399, 575-579 (1999).

Tversky, A., & Kahneman, D. Judgment under uncertainty: Heuristics and biases. *science* 185, 1124-1131 (1974).

Tversky, A., Kahneman, D., & Choice, R. The framing of decisions. *Science* 211, 453-458 (1981).

Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., & Aston-Jones, G. The role of locus coeruleus in the regulation of cognitive performance. *Science* 283, 549-554 (1999).

Usher, M., & McClelland, J. L. The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review* 108, 550 (2001).

Usher, M., & McClelland, J. L. Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological review* 111, 757 (2004).

Usher, M., Tsetsos, K., Erica, C. Y., & Lagnado, D. A. Dynamics of decision-making: from evidence accumulation to preference and belief. *Frontiers in psychology* 4 (2013).

Van de Cruys, S., de-Wit, L., Evers, K., Boets, B., & Wagemans, J. Weak priors versus overfitting of predictions in autism: Reply to Pellicano and Burr (TICS, 2012). *i-Perception* 4, 95 (2013).

Van Schie, E. C., & Van Der Pligt, J. Influencing risk preference in decision making: The effects of framing and salience. *Organizational Behavior and Human Decision Processes* 63, 264-275 (1995).

Videen, T. O., Daw, N. W., & Rader, R. K. The effect of norepinephrine on visual cortical neurons in kittens and adult cats. *The Journal of neuroscience* 4, 1607-1617 (1984).

Wang, A. T., Lee, S. S., Sigman, M., & Dapretto, M. Neural basis of irony comprehension in children with autism: the role of prosody and context. *Brain* 129, 932-943 (2006).

Waterhouse, B. D., Moises, H. C., & Woodward, D. J. Noradrenergic modulation of somatosensory cortical neuronal responses to lontophoretically applied putative neurotransmitters. *Experimental neurology* 69, 30-49 (1980).

Waterhouse, B. D., Sessler, F. M., Jung-Tung, C., Woodward, D. J., Azizi, S. A., & Moises, H. C. New evidence for a gating action of norepinephrine in central neuronal circuits of mammalian brain. *Brain research bulletin* 21, 425-432 (1988).

Waterhouse, B. D., & Woodward, D. J. Interaction of norepinephrine with cerebrocortical activity evoked by stimulation of somatosensory afferent pathways in the rat. *Experimental neurology* 67, 11-34 (1980).

Wilson, R. C., & Niv, Y. Inferring relevance in a changing world. *Frontiers in human neuroscience* 5, 189 (2011).

Woodward, D. J., Moises, H. C., Waterhouse, B. D., Hoffer, B. J., & Freedman, R. Modulatory actions of norepinephrine in the central nervous system. *Federation proceedings* 38, 2109 (1979).

Zahn, T. P., Rumsey, J. M., & Van Kammen, D. P. Autonomic nervous system activity in autistic, schizophrenic, and normal men: Effects of stimulus significance. *Journal of Abnormal Psychology* 96, 135 (1987).