Title
On objective and subjective visual perception

Author
Knotts, Jeffrey David

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

On objective and subjective visual perception

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Psychology

by

Jeffrey David Knotts

2019

ABSTRACT OF THE DISSERTATION

On objective and subjective visual perception

by

Jeffrey David Knotts

Doctor of Philosophy in Psychology University of California, Los Angeles, 2019

Professor Hakwan Lau, Chair

GOAL: The goal of this dissertation is to investigate three major questions in the field of conscious visual perception. First, to what extent do objective and subjective perception dissociate in normal observers? Second, is prefrontal cortex necessary for conscious awareness? Third, is phenomenology rich or sparse, and what would an effective operational approach to this question look like? These questions are examined in three sets of experiments with human subjects, summarized by the following aims.

AIM 1: Investigating the impacts of binocular suppression and monocular pattern masking on the relationship between objective and subjective perception in normal observers. There is considerable disagreement in the literature about whether or not normal human observers can perform forced-choice perceptual tasks unconsciously (e.g., Kolb & Braun, 1995; Morgan, Mason, & Solomon, 1997; Peters & Lau, 2015). In the first study of this dissertation, we examined whether any of four commonly used visual suppression techniques can facilitate unconscious forced-choice orientation discrimination. In three initial experiments we looked for

differences in the relationship between objective and subjective perception under different pairs of monocular and binocular suppression techniques by comparing each pair directly in an unbiased two-interval forced choice paradigm. In a fourth experiment, we examined whether continuous flash suppression can facilitate absolute unconscious perception using a similar two-interval forced choice task.

AIM 2: Using decoded fMRI neurofeedback to investigate the role of prefrontal cortex in the conscious perception of color. A second debate in consciousness science concerns whether prefrontal cortex is critical for conscious visual perception. Some metacognitive theories hypothesize that the frontoparietal activity underlying perceptual confidence judgments is indeed critical for consciousness. In the second study of this dissertation, we decoded multivariate functional magnetic resonance imaging (fMRI) patterns corresponding to perceptual confidence. We then used neurofeedback to test a potential causal relationship between activation of decoded confidence patterns and the subjective experience of color.

AIM 3: Investigating the richness of phenomenology by looking for subjective inflation effects in dot motion discrimination. A third major debate concerns whether visual phenomenology is rich or sparse. In the third part of this dissertation we review this debate and argue that subjective inflation, an effect in which peripheral or minimally attended perception appears to be subjectively richer than would be expected based on objective performance, may provide an intermediate answer: phenomenology is sparse, but it is subjectively inflated such that it feels rich. We then use a series of psychophysical experiments to test the extent to which subjective inflation occurs for random dot motion discrimination at different retinal eccentricities, and discuss the results in terms of the richness debate.

The dissertation of Jeffrey Knotts is approved.

Hongjing Lu

Martin Monti

Megan Peters

Hakwan Lau, Committee Chair

University of California, Los Angeles

2019

I dedicate this dissertation to my friends and family, both academic and not. I have been very lucky on both fronts.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This research was supported by the following:

UCLA Distinguished University Fellowship

National Science Foundation Graduate Research Fellowship Program

UCLA Dissertation Year Fellowship

As the three main chapters of this manuscript are versions of the following articles, I would like to thank the co-authors for their contributions:

Knotts, J. D., Lau, H., & Peters, M. A. K. (2018). Continuous flash suppression and monocular pattern masking impact subjective awareness similarly. *Attention, Perception & Psychophysics*. https://doi.org/10.3758/s13414-018-1578-8

Knotts, J.D., Odegaard, B., & Lau, H. (2018). The Key to Consciousness May Not be Under the Streetlight. *Current Biology,* 28 (13), R749-R752. https://doi.org/10.1016/j.cub.2018.05.033

Knotts, J.D., Cortese, A., Taschereau-Dumouchel, V., Kawato, M., & Lau, H. (Submitted). Multivoxel patterns for perceptual confidence are associated with false color detection.

Knotts, J. D., Odegaard, B., Lau, H., & Rosenthal, D. (2018). Subjective inflation: phenomenology's get-rich-quick scheme. *Current Opinion in Psychology*, *29*, 49–55.

I would also like to sincerely thank the following mentors, colleagues, and friends for their invaluable support and feedback on the work herein:

Graduate advisor, Prof. Hakwan Lau; dissertation committee, Prof. Megan Peters, Prof. Hongjing Lu, Prof. Martin Monti; Prof. Mitsuo Kawato, Prof. David Rosenthal, Prof. Brian Odegaard, Dr. Caitlin Howe, Prof. Ladan Shams, Dr. Vincent Taschereau-Dumouchel, Dr.

Aurelio Cortese, Lisa Lee, Cheryl Polfus, Prof. Alicia Izquierdo, Prof. Dean Buonomano, Prof. Dobromir Rahnev, Prof. Keith Holyoak, Prof. Frank Krasne, Prof. Alan Lee, Prof. Zili Liu, Prof. H. Tad Blair, Dr. Brian Maniscalco, Dr. Jorge Morales, Dr. Gennady Erlikhman, Dr. Pamela Kennedy, Dr. Jared Wong, Dr. Vanessa Rodriguez Barrera, Dr. Toshinori Chiba, Dr. Jesse Taylor, Kaori Nakamura, Mieko Hirata, Freya Vaughn, Hiroki Matsumura, Pablo Maceira, Dr. Piercesare Grimaldi, Andrew Eastwick, Mouslim Cherkaoui, Cody Cushing, Sivananda Rajananda, Marlene Berke, Matthias Michel, Eugene Ruby, Nathan Giles, Michael Miuccio, Ria Bhatt, Cristian Giron, Yasha Mouradi, Chloe Chow, Eric Harvey, Sarah Gonzales, Dr. Nina Lichtenberg, Dr. Yujia Peng, Nick Baker, Akila Kadambi, Andrew Frane, Mary Flaim, Shiyun Wang, McKenzie Koch, Toru Fiberesima, & Sarah Nosseir.

VITA

Education

BS, Physiology and Neuroscience 2010
University of California, San Diego

Published Manuscripts

Knotts, J.D., Odegaard, B., Lau, H., & Rosnethal, D. (2018). Subjective inflation: phenomenology's get-rich-quick scheme. *Current Opinion in Psychology*, 29, 49-55. doi: 10.1016/j.copsyc.2018.11.006

Knotts, J.D., Lau, H., & Peters, M.A.K. (2018). Continuous flash suppression and monocular pattern masking impact subjective awareness similarly. *Attention, Perception, & Psychophysics*, 80, 1974-1897. https://doi.org/10.3758/s13414-018-1578-8

Knotts, J.D., Odegaard, B., & Lau, H. (2018). The Key to Consciousness May Not be Under the Streetlight. *Current Biology,* 28 (13), R749-R752. https://doi.org/10.1016/j.cub.2018.05.033

Giron, C., Lau, H., & **Knotts, J.D.** (2018). Are Open Interviews Superior to Button Presses? A Commentary on Haun et al. (2017). *Symposium on the Brains Blog* [Blog post]. Retrieved from http://philosophyofbrains.com/2018/04/13/symposium-on-haun-tononi-koch-and-tsuchiya-are-we-underestimating-the-richness-of-visual-experience.aspx

Taschereau-Dumouchel, V., Cortese, A., Chiba, T., **Knotts, J. D.**, Kawato, M., & Lau, H. (2018). Towards an unconscious neural reinforcement intervention for common fears. *Proceedings of the National Academy of Sciences*, 201721572. http://doi.org/10.1073/pnas.1721572115

Peters, M.A.K., Fesi, J., Amendi, N., **Knotts, J.D.**, Lau, H., & Ro, T. (2017). Transcranial magnetic stimulation to visual cortex induces suboptimal introspection. *Cortex*, 93, 119–132.

Knotts, J.D., & Shams, L. (2016). Clarifying signal detection theoretic interpretations of the Muller–Lyer and sound-induced flash illusions, *Journal of Vision, 16*, 1–4. http://doi.org/10.1167/16.11.18.doi

Conference Presentations

Knotts, J.D., Peters, M.A.K., Lee, A., & Lau, H, **(Talk)**. Subjective confidence judgments for motion direction discrimination are centrally biased despite matched objective performance in the periphery. *Vision Sciences Society, St. Petersburg, FL, May 2019.*

Knotts, J.D., Cortese, A, Kawato, M, & Lau, H, **(Poster)**. Using decoded fMRI neurofeedback to identify multivariate patterns for illusory color perception. *Computational Neuroimaging and Neuroengineering Symposium, Riverside, CA, November 2018.*

Knotts, J.D., Cortese, A, Kawato, M, & Lau, H, **(Talk)**. Using decoded fMRI neurofeedback to identify multivariate patterns for illusory color perception. *Association for the Scientific Study of Consciousness Annual Meeting, Krakow, Poland, June 2018.*

Knotts, J.D., Cortese, A, Kawato, M, & Lau, H, **(Talk)**. Using decoded fMRI neurofeedback to identify multivariate patterns for illusory color perception. *Hong Kong Cognitive Science Meeting, Hong Kong, December 2017.*

Knotts, J.D., Lau, H, & Peters, MAK **(Talk)**. Human subjects have optimal introspective access even under varieties of masking conditions. *Association for the Scientific Study of Consciousness Annual Meeting,* Buenos Aires, Argentina, June 2016.

Invited Talks

"Using Psychophysics and Decoded Neurofeedback to Investigate Conscious (and Unconscious) Visual Processing". *Center of Functionally Integrative Neuroscience, Aarhus University, Denmark, May 2017.*

Awards and Fellowships

First place poster, Computational Neuroimaging and Neuroengineering Symposium, University of California, Riverside	2018
UCLA Graduate Division Dissertation Year Fellowship	2018
Real-time Functional Imaging and Neurofeedback Conference Travel Award	2017
National Sciences Foundation Graduate Research Fellowship	2015
UCLA Graduate Summer Research Mentorship Award	2015
UCLA Distinguished University Fellowship	2014
Academic Achievement Award for highest GPA in Physiology and Neuroscience major (UCSD)	2010
Phi Beta Kappa Honor Society, UCSD chapter	2009

Teaching Experience

Teaching Assistant, Fundamentals of Learning — Spring 2019
University of California, Los Angeles, Los Angeles, CA
Overall Student Rating: TBD

Teaching Assistant, Behavioral Neuroscience Laboratory — Winter 2019
University of California, Los Angeles, Los Angeles, CA
Overall Student Rating: 8.8/9.0

SAT Math Tutor, Acceso Academy, Los Angeles, CA — Winter 2019

Teaching Assistant, **I**ntroduction to Behavioral Neuroscience — Winter 2017
University of California, Los Angeles, Los Angeles, CA
Overall Student Rating: 7.4/9.0 (Average of 2 Sections)

Teaching Assistant, Introduction to Genetics — Fall 2009, Summer 2010
UC San Diego, La Jolla, CA, Department of Biological Sciences

VI. General Introduction

Background

The extent to which objective and subjective perception dissociate in normal human vision is an open question. Hereafter, objective perception refers to an observer's ability to perform a forced choice task like indicating whether a grating is tilted to the left or to the right of vertical. Subjective perception on the same task would refer to the observer's awareness of their objective performance (e.g., as measured by a confidence rating) or their rating of the visibility of the grating stimulus. Evidence that these two types of perception can dissociate comes from clinical cases of blindsight, where patients with damage to visual cortex report no subjective awareness of a target stimulus, such as a vertically or horizontally oriented line, despite performing well above chance when forced to guess the line's orientation (Weiskrantz, 1986; Weiskrantz, Warrington, Sanders, & Marshall, 1974). However, it is currently debated whether such dissociations can occur in normal observers (Peters, Kentridge, Phillips, & Block, 2017; Ian Phillips, 2017; Ian Phillips & Block, 2016)

This is a critical issue in the scientific study of consciousness, because when attempting to isolate visual consciousness, which is generally considered a subjective measure, objective measures of perception should be treated as a potential confounding factor (Lau, 2008; Lau & Rosenthal, 2011). Importantly, dissociations between objective and subjective perception can occur in two different ways. The weaker of these is a relative dissociation, also known as relative blindsight, in which objective performance is matched between two conditions but subjective awareness differs (Koizumi, Maniscalco, & Lau, 2015; Lau & Passingham, 2006; Maniscalco, Peters, & Lau, 2016; Odegaard, Chang, Lau, & Cheung, 2018; Rahnev et al., 2011;

Solovey, Graney, & Lau, 2015). The stronger case, referred to hereafter as either absolute blindsight or unconscious perception, is that in which normal observers can perform forced choice perceptual tasks with no subjective awareness. While relative blindsight is well supported by empirical evidence (see above), evidence for absolute blindsight has been controversial (Kolb & Braun, 1995; Morgan, Mason, & Solomon, 1997; Peters & Lau, 2015). Therefore, to deepen the field's understanding of the relationship between objective and subjective perception, it is important to investigate A) the extent to which relative blindsight generalizes across different psychophysical contexts, and B) whether, using rigorously controlled behavioral tasks, we can find any evidence for absolute blindsight in normal observers.

Another prominent debate in the study of consciousness regards whether prefrontal cortex (PFC) is critical for conscious experience (Boly et al., 2017; Odegaard, Knight, & Lau, 2017). While some advocates of first order theories of consciousness suggests that PFC activation reflects not perceptual content but only post-perceptual reporting mechanisms (Koch, Massimini, Boly, & Tononi, 2016; Tsuchiya, Wilke, Frässle, & Lamme, 2015), others have pointed out a wealth of support in the literature for perceptual content being encoded in PFC activity (Cortese, Amano, Koizumi, Kawato, & Lau, 2016; Hebart, Schriever, Donner, & Haynes, 2014; Mante, Sussillo, Shenoy, & Newsome, 2013; Panagiotaropoulos, Deco, Kapoor, & Logothetis, 2012; M. Wang, Arteaga, & He, 2013), and that previous failures to identify this relationship are likely due to insufficient sensitivity in analytical approach. To resolve this debate, it is therefore essential to further examine the relationship between PFC activity, perceptual content, and subjective awareness using the most sensitive recording and analysis methods that are currently available.

Finally, there is another longstanding debate about whether visual phenomenology is rich or sparse (Fazekas & Overgaard, 2018). Those who favor the Rich view argue that, taking changes in receptive field size and general representational capacity outside of the fovea into account, phenomenology is rendered in high resolution across the visual field independent of attention (e.g., Block, 1995, 2007, 2014; Lamme 2003, 2010). Those who favor the Sparse view argue that attention impacts phenomenology: where there is less attention there is sparser or more compressed phenomenology (e.g., Dehaene, Naccache, & Sergent, 2006; Cohen & Dennett, 2011). While many have argued that this debate is fundamentally irresolvable [e.g., Kouider & de Gardelle, 2010; Phillips, 2011, 2017; Overgaard & Fazekas 2016], others have argued that the debate can still benefit from the process of inference to the best explanation (Block, 2007). However, to effectively do inference to the best explanation, there is a current need for more operational definitions of rich and sparse phenomenology that can, in principle, lead to less ambiguous and more easily testable hypotheses.

Dissertation Overview

This dissertation is divided into three main chapters (VII - IX). Chapter VII examines the extent to which objective and subjective perception dissociate for central orientation discrimination judgments in a series of psychophysical experiments using different visual suppression techniques. These include forward and backward masking, interocular suppression, backward masking alone, and continuous flash suppression. Both relative and absolute blindsight are considered.

Chapter VIII begins with an overview of the debate about prefrontal cortex and consciousness (Boly et al., 2017; Odegaard et al., 2017). This overview focuses specifically on the argument

that null findings for associations between prefrontal cortex and visual consciousness may be mostly attributable to underpowered or otherwise insensitive methodological approaches. A recent study that involved waking rats from anesthesia via pharmacological stimulation of the prefrontal cortex (Pal & Mashour, 2018) is highlighted, and a consequently modified version of the mesocircuit model of consciousness (Schiff, 2010) is considered. The rest of the chapter covers a real-time fMRI neurofeedback experiment in which we rewarded the simultaneous activation of decoded multivoxel patterns for color in visual cortex and decoded multivoxel patterns for perceptual confidence in frontoparietal cortex to see if this could boost unconscious color content in spontaneous brain activity into conscious awareness.

Chapter IX starts with a review of the relevant empirical evidence in the debate about the richness of phenomenology. We focus on studies that have observed an effect known as subjective inflation, in which subjective ratings for peripheral or otherwise minimally attended stimuli appear to be overestimated based on objective performance. We argue that subjective inflation provides a helpful operational scheme for understanding and testing the richness debate. We then put this scheme to use in a series of psychophysical experiments that examine objective and subjective awareness in central and peripheral dot motion discrimination. We conclude by discussing these experiments in the theoretical frameworks of both subjective inflation and blindsight in normal observers: important similarities and differences are highlighted, and future directions that may clarify our understanding of both phenomena are suggested.

VII. The Search for Blindsight in Normal Observers: Continuous Flash Suppression and Pattern Masking

Abstract

Peters and Lau (2015) found that when criterion bias is controlled for, there is no evidence for unconscious visual perception in normal observers, in the sense that they cannot directly discriminate a target above chance without knowing it. One criticism of that study is that the visual suppression method used, forward and backward masking (FBM), may be too blunt in the way it interferes with visual processing to allow for unconscious forced-choice discrimination. To investigate this question, we compared FBM directly to continuous flash suppression (CFS) in a two-interval forced-choice task. Although CFS is popular, and may be thought of as a more powerful visual suppression technique, we found no difference in the degree of perceptual impairment between the two suppression types. To the extent that CFS impairs perception, both objective discrimination and subjective awareness are impaired to similar degrees under FBM. This pattern was consistently observed across three experiments in which various experimental parameters were varied. In a fourth experiment, we further found no evidence for absolute blindsight under CFS. These findings provide evidence for an ongoing debate about unconscious perception: normal observers cannot perform forced-choice orientation discrimination tasks unconsciously.

Introduction

Whether normal observers can perform forced-choice discrimination tasks unconsciously, i.e., whether thresholds for objective performance and subjective awareness in normal observers can dissociate in such tasks, is controversial (Peters et al., 2017; Ian Phillips, 2017; Ian Phillips

& Block, 2016). This is an important issue to resolve, because a reliable means of demonstrating unconscious perception would be an invaluable tool for studying the neural correlates of consciousness while controlling for unconscious signal processing confounds (Lau, 2008; Lau & Rosenthal, 2011).

While the dissociation of objective and subjective thresholds is typically thought to occur in blindsight (Weiskrantz, 1986; but see Phillips, 2017 for an opposing view), evidence for the same dissociation in normal observers has been contentious. Some studies have failed to replicate (e.g., Kolb & Braun, 1995; Morgan et al., 1986) while others (Hesselmann, Hebart, & Malach, 2011; Sahraie, Weiskrantz, & Barbur, 1998; Salti, Monto, Charles, & King, 2015; Vlassova, Donkin, & Pearson, 2014) have been potentially subject to the well-known confound of criterion bias (Eriksen, 1960; Hannula, Simons, & Cohen, 2005; Lloyd, Abrahamyan, & Harris, 2013; Merikle, Smilek, & Eastwood, 2001). With respect to criterion bias, the specific worry is that participants may be overly-conservative when making subjective ratings, such that on trials in which they see a small portion or a noisy gist of a target stimulus, they will still rate that target stimulus as “unseen” or “invisible” because the internal experience of the stimulus does not surpass a conservative internal criterion.

A recent study by Peters & Lau (2015) showed that when such criterion bias is controlled for by collecting subjective ratings via a two-interval forced choice (2IFC) task, there is no evidence for unconscious forced-choice discrimination in normal observers. Briefly, in their study, participants performed a left/right grating orientation discrimination task in each of two stimulus intervals and indicated in which of the two intervals they felt more confident in their orientation judgment. In each interval a series of forward and backwards masks (sometimes referred to as sandwich

masking, but hereafter referred to as FBM) were presented, but, unbeknownst to participants, on every trial, one of the two intervals lacked a target grating. Results showed that as soon as participants performed above chance in discriminating the orientation of the physically-presented target grating, they were also above chance in meaningfully assigning their confidence judgments to the interval that contained that target grating. These results suggest that when the potential for criterion bias is minimized, objective performance and subjective awareness thresholds for orientation discrimination do not dissociate -- i.e., there is no unconscious orientation discrimination -- under FBM.

A concern regarding the Peters & Lau (2015) study is that FBM may interfere at too early a stage in visual processing to facilitate unconscious perception. It could be argued then that if the authors had used a visual suppression method that interferes at a later stage, such as metacontrast masking or continuous flash suppression (CFS) (Breitmeyer, 2015), then unconscious perception would have been observed.

We addressed this concern in the current study by directly comparing different visual suppression methods in an adapted version of the 2IFC paradigm used in Peters & Lau (2015). Critically, instead of varying the presence/absence of a target stimulus between the two intervals, we presented a target stimulus in both intervals, and instead varied the suppression method used to reduce target visibility between intervals. Specifically, on each trial, a left- or right-tilted target grating in one interval was masked by a monocular pattern masking method (FBM in Experiments 1.1 & 1.2, backward masking (BM) in Experiment 1.3), while a left- or right-tilted target grating in the other interval was masked by a binocular rivalry-based method (CFS in Experiments 1.1 & 1.3, interocular suppression (IS) in Experiment 1.2).

Using this setup, if one suppression method is in fact more permissive of unconscious processing than the other, then when subjective awareness of the target grating is matched between suppression methods, there should be higher objective discrimination performance under the more permissive method. Similarly, when left/right discrimination performance under the two methods is matched near perceptual threshold, subjective awareness of the target grating should be relatively reduced under the more permissive method. In other words, we should find a difference in the magnitude of any dissociation between objective and subjective discrimination thresholds, or relative blindsight (Lau & Passingham, 2006), between the two suppression methods.

We chose to first compare masking types directly, instead of simply attempting a replication of the method in Peters & Lau (2015) with different suppression techniques, because theoretically it should be easier to find a relative dissociation between objective and subjective thresholds (e.g., Lau & Passingham, 2006) than it is to find an absolute dissociation [e.g., the failure to find such a dissociation in Peters & Lau (2015)]. A test of relative blindsight should therefore have higher sensitivity in terms of being able to detect meaningful differences in the relative positioning of objective and subjective thresholds between different visual suppression methods. We tested the hypothesis that such differences exist using a different pair of suppression methods in each of three psychophysical experiments. To anticipate, we found no such evidence. However, a lack of relative blindsight does not preclude the presence of absolute blindsight. Therefore, in a fourth experiment we replicated the task used in Peters & Lau (2015), but used CFS instead of forward-backward masking. Again, we found no evidence for unconscious perception.

Experiment 1.1: Continuous Flash Suppression Versus Forward and Backward Masking

Methods

Participants

In order to determine the number of participants that would provide sufficient power for detecting an unconscious forced-choice discrimination effect in each of the present experiments, we first thought to base our predicted effect size on previous studies by Hesselman et al. (2011) [4AFC percent correct on unseen trials = 43%, SD = 14%, Cohen's d = 1.29] and Salti et al. (2015) [8AFC percent correct on unseen trials = 50.4%, SD = 22.2%, Cohen's d = 1.71]. However, because these effect sizes are relatively large -- thus requiring relatively few subjects -- in order to increase the likelihood of detecting a more moderate unconscious effect, we instead set our predicted effect size to a more conservative level of Cohen's d = 0.8, a standard value for a large effect. To further increase the odds of finding an unconscious perception effect, we set our desired level of power to $1-\beta = 0.90$ at $\alpha = 0.05$ rather than the standard $1-\beta = 0.80$. Based on these parameters, and assuming a two-tailed one sample t-test for hypothesis testing, a power analysis showed that the necessary sample size was 19 participants.

Twenty-six participants (7 female, ages 19-39, 1 left-handed, 10 left-eye dominant, 3 experienced), including the first author, gave written informed consent to participate in Experiment 1.1. Three of these participants, noted as "experienced" above, participated in Experiments 1.2 and 1.3 prior to Experiment 1.1. All participants had normal or corrected-to-normal eyesight and normal stereo vision, and all were either paid $10 USD or given course credit for their participation. The data of five participants were removed due to failure to pass the adaptive staircasing stage (see Procedure section below). The data of one

additional participant were removed after they disclosed that they began pushing buttons at random during the main experiment. Therefore, 20 total participants (6 female, ages 20-39, 1 left-handed, 8 left-eye dominant, 3 experienced) were included in the analyses for Experiment 1.1. This and all subsequent experiments were conducted in accordance with the Declaration of Helsinki and were approved by the UCLA Institutional Review Board.

Apparatus and Stimuli

All stimuli were generated with custom Matlab R2013a (Natuck, MA) scripts using PsychToolbox 3.0.12 on a gamma-corrected Dell E773c CRT monitor with a resolution of 1024 x 768 pixels and a 75Hz refresh rate. To achieve binocular rivalry, all stimuli were viewed through a ScreenScope Desktop stereoscope. Target stimuli were sinusoidal gratings with a spatial frequency of .025 cycles/pixel tilted 45° to either the left or the right of vertical. Gratings were 153 pixels in diameter and were viewed through a circular annulus of the same diameter with a Gaussian hull spatial constant of 100. The viewing distance was 33 cm, making grating stimuli approximately 6.5 visual degrees in diameter. Mask stimuli were colored Mondrian patterns of the same dimensions as target stimuli, and were created in Matlab as previously described [(Stein, Hebart, & Sterzer, 2011)]. Target and mask stimuli were presented centered within two square-shaped boxes or “fusion contours” (one for each eye, diameter 7.4°), each side of which was composed of eleven 17x17 pixel squares, alternating between black and white (see Figure 1a). By default, fusion contours were horizontally centered within each left/right half of the screen (11.2 degrees from the midline each) and vertically centered on the screen. At the beginning of each session participants were allowed to shift the on-screen location of the left fusion contours by button press (one pixel per press in any of the cardinal directions), so as to achieve optimal fusion when viewing the screen through the stereoscope. Eight of 20 participants included in the main

analyses used this function (mean ± SD shifts = 3.58° ± 1.87° leftward and 0.77° ± 1.31° downward).

Each trial of the main experiment contained two stimulus intervals: one in which the target was masked by FBM, and the other in which the target was masked by CFS (Figure 1b,c). Each stimulus interval had a total length of 533.3 ms. In both FBM and CFS intervals, a series of 5 different masks was presented to one eye. In the FBM interval, all stimuli were presented to the non-dominant eye. Each mask was presented for 53.3 ms and separated from the next mask by a 53.3 ms blank interval, with the exception of the interval between the 2nd and 3rd masks, in the middle of which the target appeared for 26.7 ms (Figure 1b). The dominant eye was presented with nothing during the FBM interval. In the CFS interval, masks were presented to the dominant eye with the same temporal profile as in the FBM interval. To the non-dominant eye, the target was presented at a range of contrast levels, which started at 0, and ramped up linearly to a peak contrast level over the course of 173 ms. The target stayed at peak contrast for 26.7 ms, and then ramped back down to zero linearly over the course of 173 ms. The last 159.6 ms of the stimulus interval for the non-dominant eye were blank (Figure 1b). Target offset occurred prior to mask offset to prevent image aftereffects that were identified in pilot experiments and have also been identified previously (Tsuchiya & Koch, 2005). The side of target presentation was thus fixed across all trials for each participant to the side of the non-dominant eye. Eye dominance was determined using the Miles test (Miles, 1930). Timing of all stimulus presentations was validated using a Tektronix TDS 3014B oscilloscope.

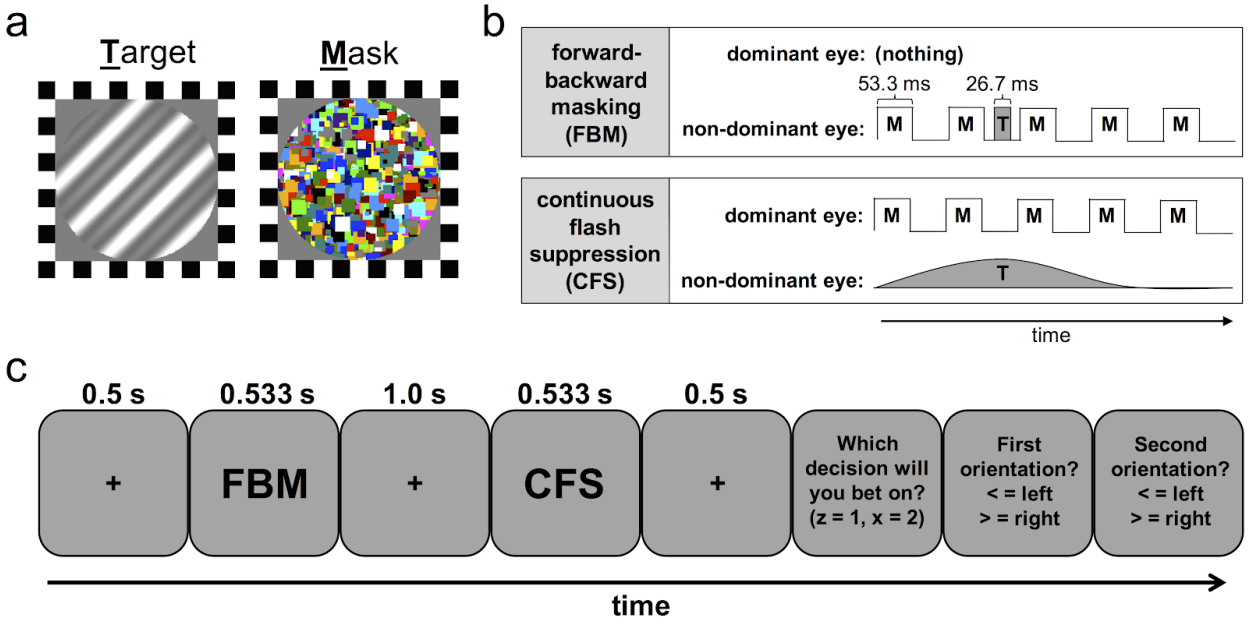

Figure 1. Continuous flash suppression versus forward and backward masking stimuli and procedure. **a)** Examples of target grating and mask stimuli used in all experiments. **b)** Temporal dynamics of stimuli in Experiment 1.1. Masks were presented for 53.3 ms each, with intervening blank gaps of the same length. In FBM intervals, masks and target were presented to the non-dominant eye, with the target appearing for 26.7 ms evenly between masks 2 and 3. In CFS intervals, masks were presented to the dominant eye, while the target grating was presented to the non-dominant eye. The contrast of the target ramped up linearly from zero to peak contrast over a period of 173 ms, remained at peak contrast for 26.7 ms, then ramped down linearly to zero over another 173 ms. **c)** procedure. FBM and CFS stimuli were presented in pseudo random order, separated by a 1.0 s interstimulus interval. Following presentation of stimuli, participants were instructed to bet on the interval in which they felt more confident in their ability to judge the orientation of the target grating. They were then asked to judge the orientations (left or right) of the target gratings in intervals 1 and 2, in that order.

Procedure

The trial structure in the main experiment extends the two-by-two forced-choice (2x2FC) paradigm first introduced by Nachmias and Weber (1975). This method was subsequently used to explore the relationship between detection and identification (Thomas, Gille, & Barker, 1982; A. B. Watson & Robson, 1981), and has more recently been applied to research on perceptual confidence (Barthelmé & Mamassian, 2009, 2010; de Gardelle & Mamassian, 2014). The participant's task was to discriminate the orientation of a masked target grating (left or right) in

each interval (Type 1 decision) and to indicate the interval in which they felt more confident about their orientation judgment (Type 2 decision; Figure 1c). Each trial started with the presentation of a white fixation cross (0.34° diameter) for 0.5 s. This was followed by the two stimulus intervals (described above – 0.5 s each) separated by a 1.0 s inter-stimulus interval containing another white fixation cross. The second stimulus interval was followed by a 0.5 s blue fixation cross to signal the upcoming response period. Participants were then presented with three response prompts, always in the same order, all of which were responded to by button press on a regular computer keyboard. First, participants were asked to make the Type 2 judgment by choosing the interval in which they felt more confident in their orientation judgment. Then participants were asked to make the Type 1 orientation judgments for the targets in the first and second intervals, respectively (Figure 1c). The confidence judgment was placed before the orientation judgments to prevent participants from factoring their reaction times on the orientation task into their confidence judgments. There was no time limit for response, and speed was never emphasized to participants. Participants were also informed that there would be several intervals in which they may not subjectively feel they saw the target, and that for these intervals, they should give their best guess as to the target's orientation.

Prior to the main experiment, participants completed 42 practice trials. Practice trial structure was identical to that in the main experiment, except for the addition of trial-by-trial feedback about the accuracy of both orientation and confidence responses. A confidence response was considered accurate if the participant bet on a correct orientation judgment. In the first 12 trials, target Michelson contrast was 100% under both suppression conditions. In the first 6 trials, stimuli were displayed at half speed. For the last 30 trials, target contrast was varied independently under each suppression condition according to an adaptive staircasing procedure

[QUEST (Andrew B. Watson & Pelli, 1983)] set to estimate the target stimulus contrast at which orientation discrimination accuracy would be 75% correct. It should be noted, however, that the function of these 30 trials was only to familiarize participants with the task under gradually more difficult conditions. Threshold contrast values were not estimated from practice session data.

Following the practice trials, participants performed another adaptive staircasing procedure to actually estimate the target contrast values at which orientation discrimination accuracy would be matched at 75% correct for both suppression methods. This procedure consisted of 4 blocks of 40 trials each, where the trial structure was identical to that of the main experiment (Figure 1c), with the exception that participants were not asked to make a confidence judgment. Staircases for CFS and FBM target gratings were independent, and a threshold contrast value was estimated for each suppression method in each block (4 estimates total per suppression method). The median of these threshold contrast estimates for each suppression method was then multiplied by five different proportions, varied slightly from subject to subject by the experimenters, in order to target orientation discrimination performance values across the range of 60-90% correct, or, roughly speaking, d' = 0.5-2.5. Proportions used to determine FBM and CFS contrast values were as follows: $\text{Proportions}_{\text{FBM}}$ = 0.51 ± 0.17, 0.75 ± 0.10, 0.95 ± 0.08, 1.09 ± 0.19, 1.35 ± 0.14; $\text{Proportions}_{\text{CFS}}$ = 0.35 ± 0.11, 0.56 ± 0.11, 0.79 ± 0.09, 1.00 ± 0.16, 1.25 ± 0.26. Notably, the proportions used for CFS stimuli were lower than those used for FBM stimuli to account for the fact that the staircasing procedure had a greater tendency to overestimate threshold contrast values for CFS stimuli compared to FBM stimuli. Furthermore, the QUEST procedure tended to overestimate the 75% correct threshold for both CFS and FBM; it was because this fact was only gradually revealed to the experimenters as more participants were included that the proportions of the median threshold estimate that were used to set

experimental contrast levels were ultimately varied slightly between participants. Furthermore, to minimize potential ceiling effects that could arise from perceptual learning during the main experiment, staircasing threshold estimates over 75% contrast were excluded from the median threshold contrast calculation.

Additionally, if participants did not have threshold contrast estimates less than or equal to 75% contrast in at least two blocks for each suppression method, they repeated the same staircasing procedure (i.e., they performed an additional four staircasing blocks). If a participant repeated the staircasing procedure, threshold estimates from only the second staircasing procedure were used to determine the contrast values used in the main experiment, and threshold estimates up to 100% were included in the median threshold calculation. As long as a participant in the second staircasing procedure had at least one threshold contrast estimate under 100% for each suppression method, they were allowed to proceed to the main experiment. Otherwise, they were told that the experiment was finished and were excluded from participating in the main experiment. Five participants were excluded in this way. Notably, all five failed the QUEST procedure only for CFS stimuli, suggesting that the CFS task was, on average, more difficult to learn than the FBM task.

For the main experiment, a full factorial design was used in which all combinations of suppression method order (2), target orientation (2x2), and target contrast level for each suppression method (5x5) were presented, leading to a total of 200 unique trials. Each unique trial was presented twice, making for a total of 400 trials, which were randomized over eight 50-trial blocks. At the end of each block, participants were allowed to take a break with no time limit. At this time they were also given a score corresponding to their performance on the

previous block, which was computed according to the following rules: one point was added or subtracted for each correct or incorrect orientation judgment, respectively. An additional point was either added or subtracted for each trial in which they correctly or incorrectly, respectively, discriminated the target orientation in the interval in which they indicated higher confidence. Participants were given a bonus of $10 USD if their final score exceeded that of the previous participant.

After participants completed the main experiment they were asked verbally by the experimenter whether, across the main experiment, they noticed any differences between the two stimulus intervals beyond basic differences in difficulty. This question was important in determining whether there may have been decisional or other cognitive response biases influencing subjects' confidence responses. For example, if a participant could consistently distinguish between the FBM and CFS intervals, they might have consciously associated one of the two with higher confidence and, consequently, bet on that interval more frequently.

Data Analysis

The main question that was investigated in each of the current studies was whether or not we could find a difference in the relationship between subjective awareness and objective performance between two visual suppression methods. To get at this question, we used orientation discrimination d' (Green & Swets, 1966) as an index of objective performance and confidence judgments as an index of subjective awareness (Fleming & Lau, 2014; Lau & Rosenthal, 2011).

For each subject, data were collapsed across target orientation order (Left-Left, Left-Right, Right-Left, Right-Right) and mask order (FBM-CFS, CFS-FBM) for each combination of contrast levels (5 FBM contrasts x 5 CFS contrasts = 25 combinations) in each trial. Orientation discrimination d' was calculated for each suppression method for each of these contrast combinations. Type 1 hits were defined as trials in which the target had a left tilt and the subject chose left. Type 1 false alarms were defined as trials in which the target had a right tilt and the subject chose left. To adjust for values of infinite d' in all experiments we used a standard correction that converts hit rates and false alarm rates of 1 and 0 to 1 - 1/2N and 1/2N, respectively, where N is the number of trials used in the calculation of d' (MacMillan & Creelman, 2004).

We then plotted, for each of the 25 contrast combinations for each subject, the proportion of trials in which the CFS interval was rated with higher confidence against the difference in d' between the CFS and FBM intervals (see Figure 2b). Individual psychometric curves were then generated by fitting the resulting 25 data points with a cumulative normal distribution function with free parameters α (threshold) and β (slope), and fixed parameters γ (lapse rate) = 0 and δ (guess rate) = 0, using the Palamedes Toolbox (Kingdom & Prins, 2010).

If there is no difference in the relationship between subjective awareness and objective performance between two given suppression methods, then we should expect the point of subjective equality (PSE), or the difference in d' at which participants are equally likely to bet on the two suppression methods, to be zero. Similarly, the point of objective equality (POE), or the likelihood of betting on the CFS interval when $d'_{CFS} - d'_{FBM} = 0$, should be 50%. If, on the other hand, the relationship between subjective awareness and objective performance is significantly

different between the two suppression methods, then the psychometric function should shift such that the PSE and POE should be significantly different from zero and 50%, respectively. Therefore, in each of the following experiments, the first two major tests of interest were one-sample t-tests ($\alpha = .05$, two-tailed) conducted on the PSEs and POEs obtained from the individually-fitted psychometric functions, with the null hypothesis being that the mean PSE and POE across participants are equal to zero and 50%, respectively. In the case of the POE analysis, since d' is matched between suppression methods, subjective awareness is operationally defined in this case in line with Giles, Lau, & Odegaard (2016), as the difference in Type 2 responding when Type 1 performance is matched. Analyses for each experiment were conducted in Matlab R2013a (Natuck, MA), with the exception of repeated measures ANOVAs, which were conducted in SPSS v22 (IBM, Armonk, NY, USA), and TOST equivalence tests, which were conducted in R using the TOSTone.bf function provided in Lakens et al. (2018). All repeated measures ANOVAs were adjusted for violations of the assumption of sphericity with the Greenhouse-Geisser correction when necessary.

Additionally, we used the 'two one-sided tests' (TOST) approach as described by Lakens et al. (2018) to more rigorously test whether or not we can reject the presence of an unconscious perception effect as small as Cohen's d = 0.8 (see power analysis description in 'Participants' section above). This approach, as it sounds, entails performing two one-sided t-tests against lower and upper bounds that are determined by a given minimum effect size of interest. In this case the null hypothesis is now that the true effect (here, a shift in either the PSE or POE) is at least as large as this smallest effect size of interest. If we are able to reject this null hypothesis at each bound at alpha = 0.05, then we can conclude that the true effect in the population is smaller than the smallest effect size of interest with a type 1 error rate of 5% (Lakens et al.,

2018). We therefore present two sets of t- and p-values for each TOST analysis, along with the 90% confidence intervals for POE and PSE shifts.

Results & Interim Discussion

A repeated measures ANOVA with within-subjects factors contrast (5 levels) and suppression method (FBM or CFS) revealed the expected main effect of contrast on orientation discrimination d' [$F(1.92,36.56) = 79.62$, $p < 0.001$] (Figure 2a), i.e. that increased contrast led to higher performance. The ANOVA also showed no main effect of suppression method [$F(1,19) = 1.73$, $p = 0.20$], but a significant interaction between contrast and suppression method [$F(2.40,45.58) = 4.68$, $p = 0.010$]. Figure 2a suggests that this interaction is driven by the sudden divergence in d' between suppression methods at the highest contrast level. This was confirmed by post hoc Bonferroni corrected two-tailed paired t-tests [at contrast level 5: $t(19) = 4.39$, $p < 0.001$, whereas p-values for contrast levels 1-4 were all > 0.32]. Because, by design, contrast levels for the main experiment were selected on a subject-by-subject basis with the goal of optimally matching d' between suppression methods, this result is mostly attributable to experimenter error. Nonetheless, the lack of a main effect of suppression method on discrimination d' in the ANOVA indicates that, across contrast levels, Type 1 performance was matched between FBM and CFS. However, to check for any potential biasing of the PSE and POE analyses that could have resulted from the difference in d' between suppression methods at the highest contrast level, we conducted the PSE and POE analyses once with all contrast levels included, and once using only contrast levels 1-4.

As for the main analyses, looking across all contrast levels, PSE and POE values were -0.28 ± 0.21 and 51.5% ± 2.1%, respectively. Two-tailed paired t-tests indicated insufficient evidence to

reject the null hypotheses that the PSE is equal to zero [t(19) = -1.35, p = 0.19, 95% CI = (-0.72, 0.15)] and the POE is equal to 50% [t(19) = 0.73, p = 0.47, 95% CI = (47.2%, 55.8%)] (Figure 2b). When we excluded the highest contrast level, there was still insufficient evidence to reject the null hypothesis in each case [PSE: t(19) = -1.29, p = 0.21, 95% CI = (-0.80, 0.1926); POE: t(19) = 1.04, p =0.31, 95% CI = (47.9%, 56.4%)].

TOST equivalence tests for both PSE and POE suggested that we can reject the hypothesis that there is an unconscious perception effect in the study sample as small as Cohen's d = 0.8 [PSE: t1(19) = 2.22, p1 = 0.019, t2(19) = -4.93, p2 < 0.001, 90% CI for PSE shift = (-0.64, 0.08); POE: t1(19) = 4.31, p1 < 0.001, t2(19) = -2.84, p2 = 0.005, 90% CI for POE shift = (-2.1%, 5.1%)]. We emphasize again that the minimum effect size selected for our power and TOST analyses of Cohen's d = 0.8, which is, generally speaking, a large effect size, is relatively conservative here given multiple previously reported unconscious forced-choice discrimination effect sizes greater than Cohen's d = 1.2 (see Experiment 1.1 Methods). However, given the more conservative controls for criterion bias in the present paradigm, we might naturally expect smaller effect sizes than those found in earlier studies. The same pattern of results held when excluding the highest contrast level [PSE: t1(19) = 2.29, p1 = 0.017, t2(19) = -4.87, p2 < 0.001, 90% CI for PSE shift = (-0.71, 0.10); POE: t1(19) = 4.61, p1 < 0.001, t2(19) = -2.54, p2 = 0.010, 90% CI for POE shift = (-1.41%, 5.61%);].

Overall, the PSE results suggest that when subjective awareness is matched, there is no difference in the level of objective performance under FBM and CFS. Similarly, the POE analysis suggests that when Type 1 performance is matched, there is no difference in participants' subjective awareness of the target stimulus between the two suppression methods.

Importantly, all participants responded in the negative when asked, after the main experiment, if on any trials they noticed differences between the two intervals beyond difficulty level. This suggests that participants' confidence judgements were not subject to decisional biases based on explicit knowledge about the difference between FBM and CFS stimuli.

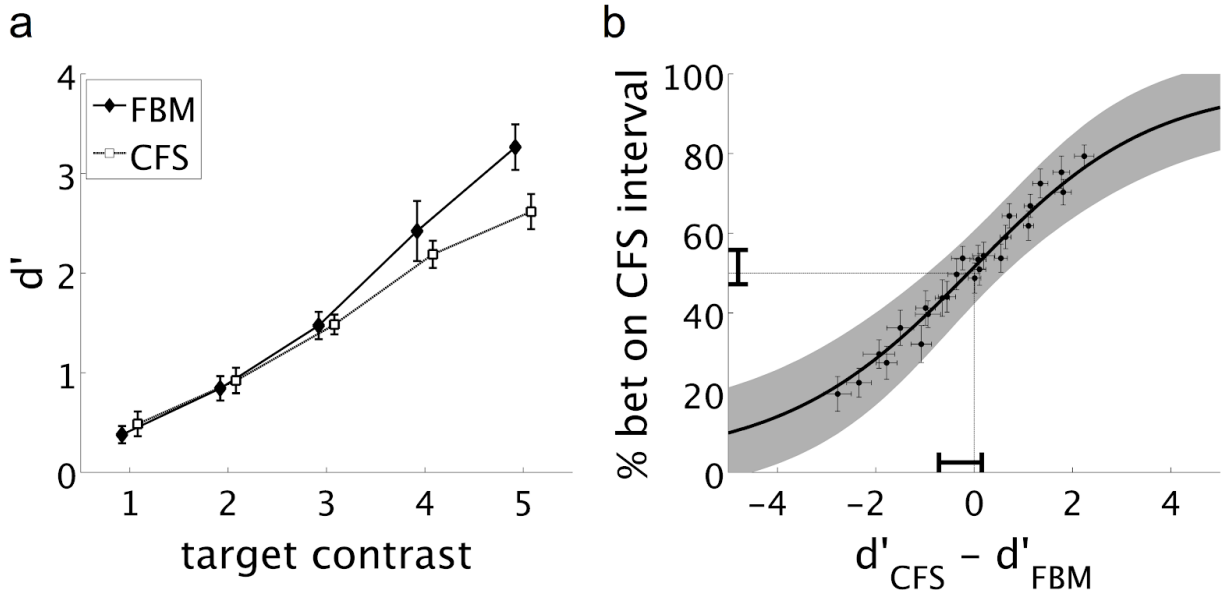

Figure 2. Continuous flash suppression versus forward and backward masking results (Experiment 1.1) results. **a)** Orientation discrimination performance (d') at increasing target contrast under FBM (solid line) and CFS (dashed line). Error bars indicate ± 1 SEM. **b)** Average psychometric curve. For each participant, the proportion of trials in which they bet on the CFS interval was plotted as a function of the difference in d' between CFS and FBM intervals for each of the 25 combinations of stimulus contrast levels that could occur in a single trial (shown are group means ± 1 SEM). A cumulative normal function was then fit to each participant's data, with mean and slope as free parameters. Plotted is the mean of the individual participant fits (black line) ± 1 SD (gray). The 95% confidence interval for the estimated PSE and POE group means are shown by the black bars sitting near the x- and y-axes, respectively. A significant rightward shift of the psychometric curve, such that the confidence interval for the PSE were to fall above of zero, would suggest that when subjective awareness is matched between CFS and FBM, d' is significantly higher under CFS than under FBM. Similarly, if a rightward shift of the psychometric curve makes it such that the confidence interval for the POE falls below 50%, then it would suggest that subjective awareness of the target stimulus is higher under FBM when d' is matched. This would indicate relative blindsight. The opposite interpretations would hold if the confidence interval for PSE were below zero and the confidence interval for POE were above 50%. The fact that zero falls within the observed PSE confidence interval suggests that when

subjective awareness of the target was matched between CFS and FBM, there was no significant difference in discrimination d' between the two suppression methods. Similarly, the fact that 50% falls within the observed POE confidence interval suggests no evidence for relative blindsight.

One concern is that the gaps between masks in the CFS condition, which lead to a collective 156.9 ms in which the target is presented to one eye with no mask presented to the other eye (Figure 1b), may minimize the degree to which the CFS condition elicits a true binocular rivalry effect. If this is the case, then, presumably, it should also minimize mechanistic differences underlying the disruption of visual processing between the two suppression methods, thereby reducing our chances of rejecting the null hypothesis.

One potential piece of evidence that FBM and CFS use different mechanisms to disrupt visual processing is that target contrast values were significantly lower for CFS stimuli (23.42 ± 2.94% Michelson contrast) than they were for FBM stimuli (35.18 ± 1.17% Michelson contrast) [two-tailed, paired t-test: $t(19) = 4.89$, $p < 0.001$]. Importantly, this result holds when excluding the highest contrast level [two-tailed, paired t-test: $t(19) = 5.12$, $p < 0.001$]. However, an alternative interpretation is that the lower contrast thresholds found in the CFS condition are simply driven by the longer presentation times for CFS target stimuli relative to FBM target stimuli. Disambiguating these hypotheses is critical for establishing that the FBM and CFS conditions induce mechanistically different visual suppression effects. We address this issue directly in Experiment 1.2.

Experiment 1.2: Continuous Flash Suppression Versus Interocular Suppression

Methods

Participants

Twenty participants (9 female, ages 18-39, 3 left-handed, 8 left-eye dominant, 7 experienced), including the first author, gave written informed consent to participate. Six of the 20 participants in Experiment 1.2 had previously participated in Experiment 1.1. One participant had previously participated in Experiment 1.3. One participant (inexperienced) was excluded due to reporting incomplete fusion of binocular stimuli on many trials during the main experiment. Therefore, 19 total participants (8 female, ages 21-39, 2 left-handed, 8 left-eye dominant, 7 experienced) were included in the analyses for Experiment 1.2. All participants had normal or corrected-to-normal eyesight and normal stereo vision, and all were either paid $10 USD or given course credit for their participation.

Apparatus and Stimuli

Apparatus and stimuli in Experiment 1.2 were the same as in Experiment 1.1, except for the following. Instead of CFS, we used a binocular rivalry technique conventionally referred to as interocular suppression (IS) (Breitmeyer, 2015; Izatt et al., 2014). The sole difference between the CFS condition in Experiment 1.1 and the IS condition in Experiment 1.2 is that the target grating no longer had its contrast ramped up from and down to zero. Instead, the IS target grating had the same duration as the FBM target grating (26.7 ms), and its contrast was constant (Figure 3a). Furthermore, in each interval the target had an equal probability of appearing between either masks 2 and 3 or masks 3 and 4. The randomization was independent in the two intervals such that in approximately half of all trials (48% ± 3%) the target appeared between the same mask numbers (e.g., 2 and 3) in each interval, while in the

remainder of trials the target appeared between different mask numbers in each interval (e.g., between masks 2 and 3 for FBM and masks 3 and 4 for IS). This manipulation was introduced to minimize the degree to which participants could anticipate the timing of target onset. Such anticipation, whether conscious or unconscious, could potentially minimize visual processing differences between the two masking conditions. Eight participants included in the analyses shifted the left fusion contours at the beginning of the experiment by 3.36° ± 1.53° leftward and 2.51° ± 1.96° downward.

Therefore, across the entire experiment, the only difference between FBM and IS stimuli was ocularity (Figure 3a). It follows that if we observe differences in contrast thresholds and stimulus contrast values at matched d' between FBM and IS similar to those found between FBM and CFS in Experiment 1.1, then these differences should be attributed to the difference in ocularity between the two conditions. This result would provide additional evidence for the presence of a binocular rivalry-based suppression effect in our original CFS condition.

We also reasoned, based on previous evidence for a higher degree of subliminal priming under FBM than IS (Breitmeyer, 2015; Izatt, Dubois, Faivre, & Koch, 2014), that IS may allow a greater degree of unconscious orientation discrimination than FBM. If true, we would expect a leftward shift in the psychometric function such that at the PSE there would be significantly higher discrimination d' under FBM than under IS, and at the POE there would be a significantly higher tendency to bet on the IS interval.

Procedure and Data Analyses

The procedure in Experiment 1.2 was the same as that in Experiment 1.1 except for the following. Different proportions of the median threshold contrast estimate from staircasing were used to determine target contrast values for the main experiment ($Proportions_{FBM}$ = 0.50 ± 0.14, 0.69 ± 0.11, 0.95 ± 0.11, 1.09 ± 0.13, 1.35 ± 0.19; $Proportions_{IS}$ = 0.35 ± 0.13, 0.56 ± 0.09, 0.79 ± 0.04, 1.02 ± 0.08, 1.25 ± 0.19). Again, the proportions used for IS were lower than those used for FBM to account for the tendency of the staircasing procedure to overestimate threshold contrast values to a greater extent for IS stimuli than for FBM stimuli.

Additionally, 40 catch trials, in which the contrast of the target grating in one of the two intervals (counterbalanced between suppression methods) was at 100%, were randomly interleaved among the 400 main experiment trials. This made for a total of 440 trials in the main experiment, which were divided into eight 55-trial blocks. These catch trials were added both to help participants maintain perceptual templates of the left- and right-tilted target gratings, and to keep participants motivated throughout what is otherwise a very difficult and, according to anecdotal evidence from some participants following Experiment 1.1, sometimes demoralizing task.

Analysis procedures followed those conducted in Experiment 1.1.

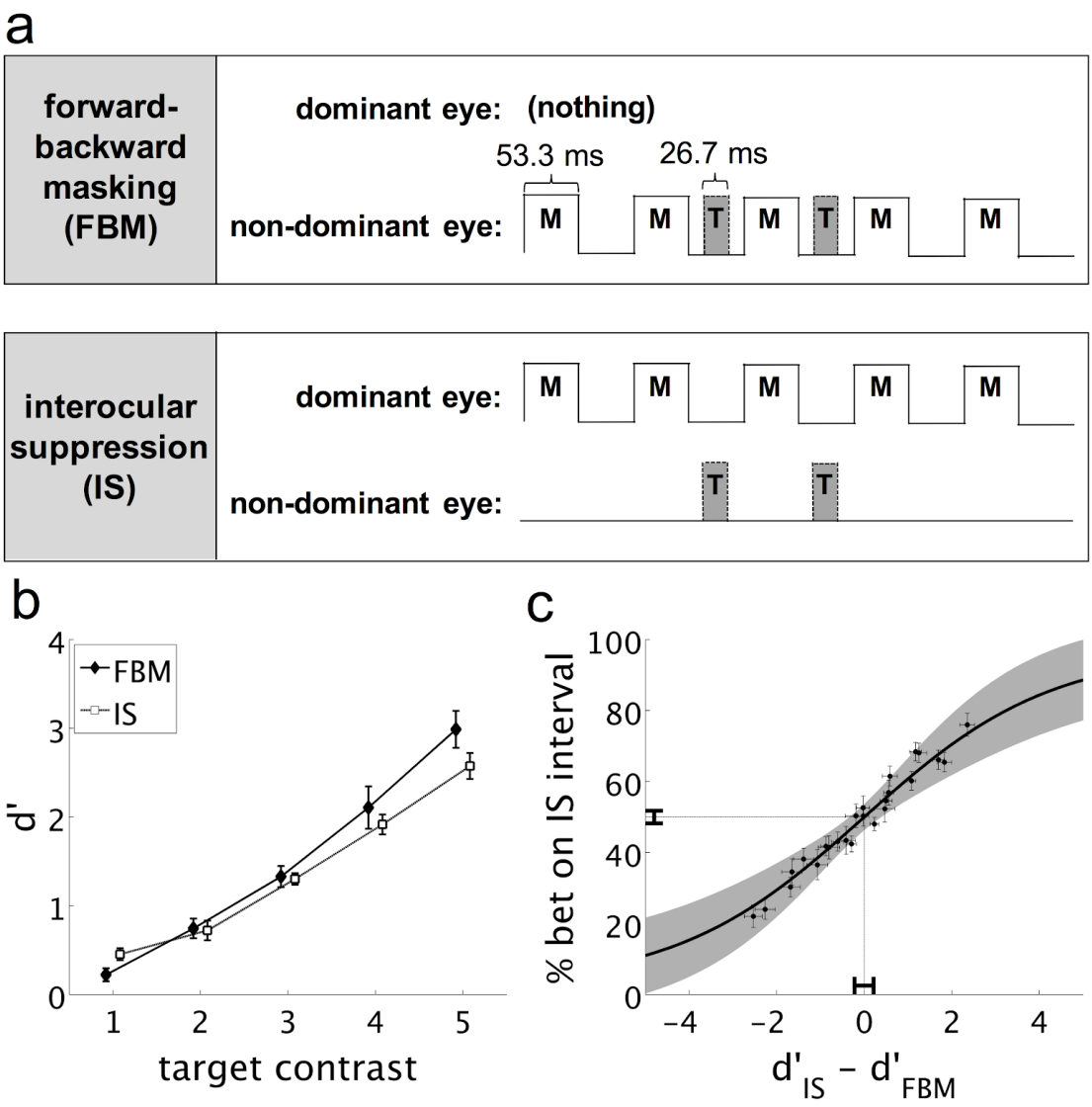

Figure 3. Experiment 1.2 procedure and results. **a)** Temporal dynamics of stimuli from Experiment 1.2. Mask stimuli had the same temporal profile as those in Experiment 1.1. Target stimuli in the IS interval appeared abruptly at peak contrast instead of ramping up and down in contrast as in Experiment 1.1. In each interval target stimuli were presented pseudo randomly between either the second and third or third and fourth masks. **b)** Orientation discrimination performance (d') at increasing target contrast under FBM (solid line) and IS (dashed line). **c)** Average psychometric curve and 95% confidence intervals for estimated PSE and POE group means, calculated and shown the same way as in Experiment 1.1 (see methods, Figure 2). Because the PSE confidence interval contains the point $d'_{difference} = 0$ and the POE confidence interval contains the point at which subjects were 50% likely to bet on either suppression method, these results suggest that there was no evidence for a difference in the relationship between objective and subjective thresholds between FBM and IS. Error bars in B & C indicate ± 1 SEM. Gray region in C indicates ± 1 SD of psychometric fits.

Results & Interim Discussion

A repeated measures ANOVA with within-subjects factors contrast (5 levels), suppression method (FBM or IS), and target timing (between masks 2 and 3 or between masks 3 and 4) again showed the expected main effect of contrast on orientation discrimination d' [$F(2.74,49.38) = 159.85$, $p < 0.001$; Figure 3a]. As in Experiment 1.1, there was no main effect of suppression method [$F(1,18) = 0.17$, $p = 0.68$], suggesting again that, overall, performance was matched between the FBM and IS conditions. Unlike Experiment 1.1, however, there was no interaction between contrast and suppression method [$F(2.66,47.85) = 2.26$, $p = 0.10$], suggesting that discrimination d' was matched effectively between the two suppression methods across contrast levels.

Interestingly, there was a main effect of target timing [$F(1,18) = 25.41$, $p < 0.001$], such that discrimination d' was significantly higher when the target stimulus was presented between masks 2 and 3 than when it was presented between masks 3 and 4. This effect of stimulus timing on objective performance may be attributable to rhythmic attentional sampling (Landau & Fries, 2012) set by visual cues preceding the onset of the target stimulus (e.g., the initial fixation cross or the onset of the first mask).There was no interaction between either target timing and contrast [$F(4,72) = 2.01$, $p = 0.10$], or target timing and suppression method [$F(1,18) = 0.65$, $p = 0.432$]. There was no significant 3-way interaction [$F(4,72) = 1.26$, $p = 0.30$].

As in Experiment 1.1, we did not find evidence to reject the null hypothesis that d' is matched between suppression methods at the PSE [$t(18) = -0.07$, $p = 0.95$, 95% CI = (-0.22, 0.21); Figure 3b]. Similarly, we did not find evidence to reject the null hypothesis that subjects are equally likely to bet on each suppression method at the POE (i.e., there was no evidence for

relative blindsight) [t(18) = -0.004, p > 0.99, 95% CI = (48.2%, 51.8%)]. TOST equivalence tests similarly suggested that the effect size of any POE or PSE shift in the population is no larger than Cohen's d = 0.8 [PSE: t1(18) = 3.42, p1 = 0.002, t2(18) = -3.55, p2 = 0.001, 90% CI for PSE shift = (-0.19, 0.17); POE: t1(18) = 3.48, p1 = 0.001, t2(18) = -3.49, p2 = 0.001, 90% CI for POE shift = (-1.46%, 1.46%)].

It was also verified that target contrast values (across all levels) were again lower under IS (15.99 ± 1.78%) than they were under FBM (30.52 ± 2.51%) [t(18) = 6.19, p < 0.001]. This provides evidence for a difference in the mechanism of visual suppression between FBM and the binocular conditions in both Experiments 1.1 and 1.2, despite the absence of the hypothesized difference in the relationship between objective performance and subjective awareness.

Also consistent with Experiment 1.1, no participants indicated noticing a difference between FBM and IS intervals when questioned after the main experiment. Furthermore, participants were 98.8% ± 0.64% correct when discriminating catch trial target stimuli with 100% contrast. Betting accuracy on catch trials was similarly high (97.5 ± 0.94% correct, where a correct bet is defined as a bet on an interval in which the orientation judgment was correct), suggesting that participants were maintaining attention throughout the experiment.

Given that the main results in Experiments 1.1 and 1.2 both suggest that there is no difference in the relative positions of subjective and objective perceptual thresholds between the respective monocular and binocular suppression methods, we next turned to backward masking (BM) as an alternative to FBM. Previous evidence suggests that BM, but not CFS, allows for the

subliminal priming with non-manipulable objects (Almeida, Mahon, Nakayama, & Caramazza, 2008). It has also been suggested that, relative to FBM, the visual signal under BM may benefit from an increased signal-to-noise ratio when performance is matched (Breitmeyer, 2015; Harris, Wu, & Woldorff, 2011; Macknik & Livingstone, 1998). We therefore hypothesized that BM may allow for a greater degree of unconscious processing than CFS, and that, in our 2IFC paradigm, we may therefore see the psychometric function shift so as to show higher discrimination d' under BM at the PSE, and a higher tendency to bet on the CFS interval at the POE.

Experiment 1.3: Continuous Flash Suppression Versus Backward Masking

Methods

Participants

Twenty participants (8 female, ages 21-39, 2 left-handed, 10 left-eye dominant, 14 experienced), including the first author, gave written informed consent to participate. Three of the 20 participants in Experiment 1.3 had previously participated in only Experiment 1.1, seven had previously participated in only Experiment 1.2, and four had previously participated in both Experiments 1.1 and 1.2. One participant was removed due to failure to pass the adaptive staircasing stage. Therefore, 19 participants (7 female, ages 21-39, 2 left-handed, 10 left-eye dominant, 14 experienced) were included for analysis. All participants had normal or corrected-to-normal eyesight and normal stereo vision, and all were either paid $10 USD or given course credit for their participation.

Apparatus and Stimuli

Apparatus and stimuli in Experiment 1.3 were the same as in Experiment 1.1, except for the following. For both BM and CFS conditions, mask stimuli were shifted later in time by 26.7 ms.

In the BM interval, the first mask was preceded by the target, which had a duration of 26.7 ms, meaning target offset coincided with mask onset. In the CFS interval, target onset coincided with the onset of the first mask and returned to the same ramping dynamics used in Experiment 1.1 (Figure 4a). Nine of the 19 participants included in the analyses shifted the left fusion contours at the beginning of the experiment by 2.81° ± 1.95° leftward and 2.65° ± 1.89° downward .

Procedure

The procedure in Experiment 1.3 was the same as that in Experiment 1.2 except for the use of different proportions of the median threshold contrast estimate to determine target contrast values for the main experiment ($Proportions_{BM}$ = 0.15 ± 0.10, 0.28 ± 0.12, 0.48 ± 0.09, 0.70 ± 0.08, 0.96 ± 0.09; $Proportions_{CFS}$ = 0.19 ± 0.18, 0.38 ± 0.14, 0.62 ± 0.08, 0.87 ± 0.05, 1.18 ± 0.13). Interestingly, median threshold contrast estimates were, on average, more overestimated under BM than they were under CFS. As a result, the proportions used to determine target contrast levels for the main experiment were lower for BM stimuli than they were for CFS stimuli.

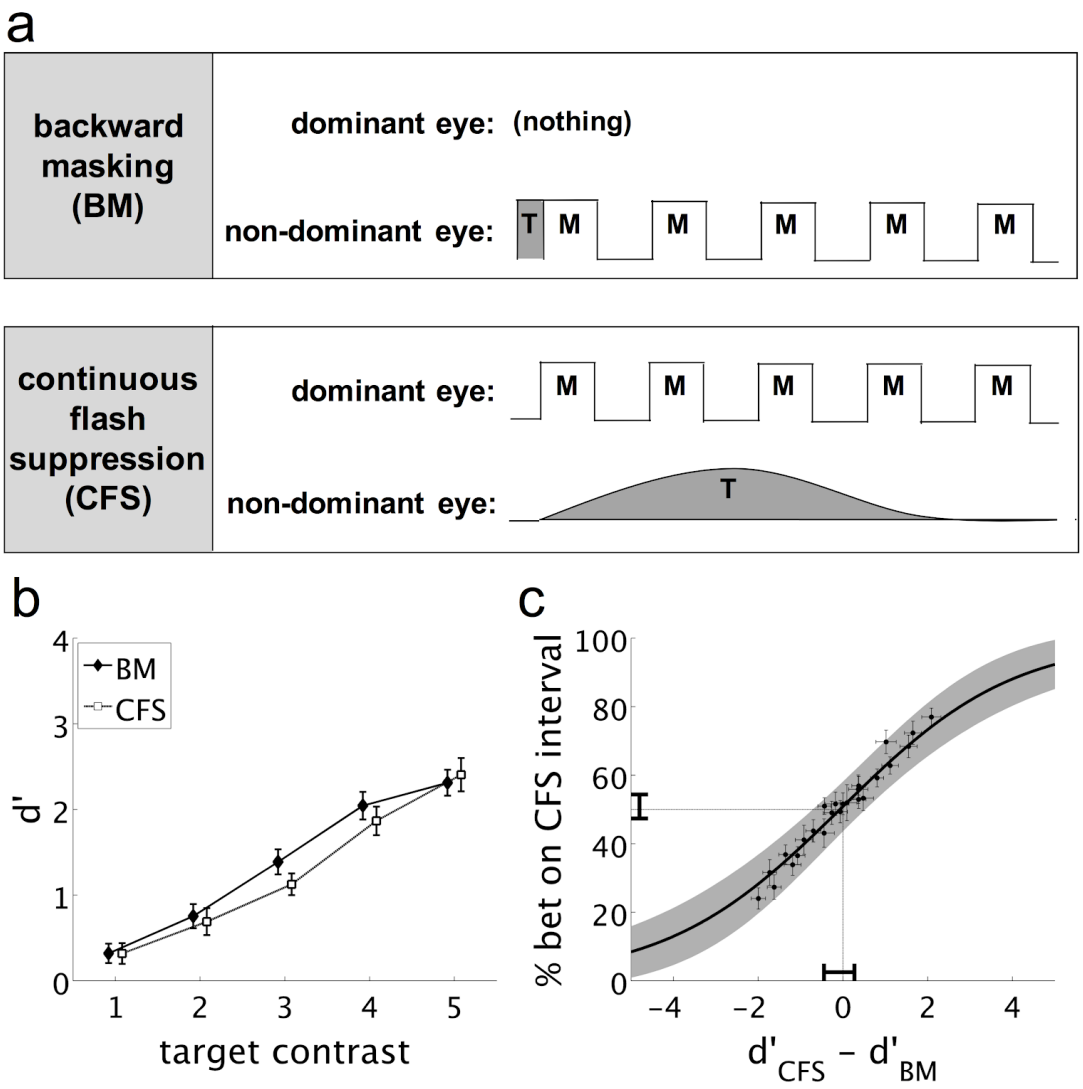

Figure 4. Experiment 1.3 procedure and results. **a)** Temporal dynamics of stimuli from Experiment 1.3. In BM intervals the target was first presented for 26.7 ms and was immediately followed by the first mask. Five masks were presented for 53.3 ms each, with intervening blank gaps of the same length. The offset of the last mask was followed by a blank gap of 26.7 ms. Masks in the CFS interval had the same temporal profile as those in the BM interval. The onset of the target stimulus in the CFS interval occurred simultaneously with the onset of the first mask and otherwise had the same temporal ramping profile as the target stimulus in Experiment 1.1 (see methods, Figure 1). **b)** Orientation discrimination performance (d') at increasing target contrast under BM (solid line) and CFS (dashed line). **c)** Average psychometric curve and 95% confidence intervals for estimated PSE and POE group means, calculated and shown the same way as in Experiment 1.1 (see methods, Figure 2). Because the PSE confidence interval contains the point $d'_{difference} = 0$ and the POE confidence interval contains the point at which subjects were 50% likely to bet on either suppression method, these results suggest that there was no evidence for a difference in the relationship between objective and subjective thresholds between BM and CFS. Error bars in B & C indicate ± 1 SEM. Gray region in C indicates ± 1 SD of psychometric fits.

Results & Interim Discussion

Consistent with Experiments 1.1 and 1.2, a repeated measures ANOVA with within-subject factors contrast (5 levels) and suppression method (BM or CFS) showed the expected main effect of contrast [$F(2.54,45.7) = 154.8$, $p < 0.001$; Figure 4b] and no main effect of suppression method [$F(1,18) = 0.30$, $p = 0.59$]. Consistent with Experiment 1.2, there was no significant interaction between contrast and suppression method [$F(2.31, 41.6) = 0.78$, $p = 0.48$], suggesting that d' was effectively matched between suppression methods across contrast levels (Figure 4b).

Regarding the main analysis, once again there was not sufficient evidence to reject the null hypothesis that d' is matched between BM and CFS at the PSE [$t(18) = -0.52$, $p = 0.61$, 95% CI = (-0.45, 0.27); Figure 4c]. Nor was their sufficient evidence to reject the null hypothesis that subjects are equally likely to bet on each suppression method at the POE [$t(18) = 0.53$, $p = 0.60$, 95% CI = (47.4%, 54.4%)], again providing no evidence for relative blindsight. TOST equivalence tests again suggested that any POE or PSE shifts in the population have effect sizes no larger than Cohen's d = 0.8 [PSE: $t1(18) = 2.97$, $p1 = 0.004$, $t2(18) = -4.01$, $p2 < 0.001$, 90% CI for PSE shift = (-0.38, 0.21); POE: $t1(18) = 4.02$, $p1 < 0.001$, $t2(18) = -2.96$, $p2 = 0.004$, 90% CI for POE shift = (-2.0%, 3.8%)]. Experiment 1.3 was therefore in line with Experiments 1.1 and 1.2 in providing no evidence for a difference in the relationship between objective performance and subjective awareness between suppression methods.

Interestingly, mean stimulus contrast per subject in the main experiment was significantly lower under BM (9.30 ± 1.45%) than under CFS (17.737 ± 2.60%) [$t(18) = -3.28$, $p = 0.004$]. This decrease in threshold target contrast from FBM to BM is presumably due to the relative lack of

interference with feedforward processing under BM (Breitmeyer, 2015; Harris et al., 2011; Macknik & Livingstone, 1998).

No participants reported noticing a difference between BM and CFS intervals when questioned after the main experiment. Performance on catch trials was again high (orientation judgment accuracy: 97.5 ± 1.10% correct, betting accuracy: 97.2 ± 0.93% correct), suggesting that participants maintained attention throughout the task.

Considering the results from Experiments 1.1 - 1.3 together with those from Peters & Lau (2015), we might predict, by transitive logic, a lack of absolute blindsight under CFS. However, as stimuli were not matched between the current experiments and Peters & Lau (2015), it is still possible that some absolute unconscious perception is permitted under the current FBM and CFS paradigms. We tested this question in Experiment 1.4.

Experiment 1.4: Testing Absolute Blindsight Under Continuous Flash Suppression

Methods

Participants

Twenty-two participants (14 female, ages 18-42, 2 left-handed, 11 left-eye dominant, 5 experienced), gave written informed consent to participate.All participants had normal or corrected-to-normal eyesight and normal stereo vision, and all were either paid $10 USD or given course credit for their participation.

Apparatus and Stimuli

The same apparatus, stimuli, and procedure as in the CFS condition in Experiment 1.3 were used in Experiment 1.4 with the following exceptions. CFS was used to mask target stimuli in all stimulus intervals throughout Experiment 1.4. Critically, as previously described (Peters & Lau, 2015), on each trial, unbeknownst to participants, one interval contained a left- or right-tilted target grating [target present (TP) interval] while the other contained no target grating [target absent (TA) interval; Figure 5a,b]. The target absent interval contained either a vertically oriented grating or no grating at all (Figure 5b). Further, the spatial frequency of all gratings increased from 0.025 to 0.043 cycles per pixel or 1 cycle per degree visual angle (dva), and the gaussian hull was reduced from 100 to 25.5 (Figure 5c). Masks were also temporally contiguous throughout each stimulus interval (Figure 5b), and the size of mask stimuli was increased to fill the entire area within the fusion contour presented to the dominant eye (Figure 5c).

Procedure

The experiment consisted of two ~1 hour sessions across two days. On each day participants completed a variable number of practice trials based on their performance (mean ± s.d. Number of trials on Day 1 = 69.0 ± 30.5, and Day 2 = 41.8 ± 23.7). During the practice trials there were valid targets in both stimulus intervals. This was to both to give subjects additional practice in discriminating valid targets and to strengthen participants' belief that there were also valid targets in both intervals during the main task. All practice trials included trial-by-trial feedback about the accuracy of the orientation discrimination judgment in each interval and the accuracy of the confidence judgment as in the previous experiments.

After the practice trials participants performed 80 trials of an adaptive staircasing procedure to estimate the grating contrast level that would lead to 75% correct orientation discrimination performance. As in the practice trials, there were valid targets in both intervals throughout the staircasing procedure, and confidence judgments were removed to save time. The staircasing procedure used two interleaved 40-trial staircases, and the average of the two resulting threshold contrast estimates was used to estimate each participant's 75% correct contrast threshold. This threshold was then multiplied by five proportions, as in Experiment 1.1-1.3 to target a set of five contrast levels that would lead to orientation discrimination performance scores across the range of 55% to 90% correct.

Following the staircasing procedure on Day 1 participants performed 270 trials of the main task (Figure 5a). Thirty of these were catch trials in which both intervals contained valid targets, and one or more of these had a full contrast of 1 [15 trials where both intervals contained full contrast, and 15 trials where one interval contained one of the five near-threshold contrast values estimated from the staircasing procedure (3 trials per contrast level)]. These catch trials served three functions: 1) to help participants maintain perceptual templates of the target stimuli throughout the main task, 2) to help maintain participants' belief that all trials in the main task contained two valid targets, and 3) to ensure that participants were maintaining attention throughout the task (full contrast stimuli were easy to discriminate and should therefore elicit ceiling performance if participants are paying attention). For non-catch trials on Day 1 we used a randomized full factorial combination of 5 target contrast levels, 2 target absent interval types, 2 interval orders, and 2 target orientations with 5 trials per unique combination. The contrast of vertical gratings was randomly selected to be one of the 5 near-threshold contrast values used

for target stimuli, except for in catch trials, when it could be either one of the 5 near-threshold contrast values or 1.

Following Day 1, orientation discrimination scores were computed across trials for each of the five near-threshold grating contrast levels. The first author then made eyeball estimates of five new contrast levels to be used on Day 2 that were intended to lead to orientation discrimination scores that would fill in any gaps within the targeted 55% to 90% correct range in the Day 1 orientation discrimination performance data. These five new contrast levels were used exclusively on Day 2.

On Day 2, after the initial practice trials, participants performed 440 trials of the main task, 40 of which were catch trials [20 with full contrast targets in both intervals, and 20 with near-threshold contrast values (4 trials per near-threshold contrast level)]. For non-catch trials we used the same randomized full factorial approach as on Day 1, but with 10 trials per unique combination of conditions. At the end of Day 2, participants were asked if they noticed anything consistent difference between the two stimulus intervals throughout the main task. This was intended to gauge whether or not participants noticed that one interval always contained an invalid target.

Data Analyses

As previously described (Peters & Lau, 2015), we plotted the proportion of trials in which participants indicated higher confidence, or bet, on the target present interval as a function of orientation discrimination percent correct scores in the target present interval (Figure 5c,d). If participants can perform the orientation discrimination tasks unconsciously, then this “absolute blindsight curve” should show a range of orientation discrimination proportion correct scores

above 0.5 correct (chance) for which the proportion of trials in which the target present interval was bet on is roughly 0.5. The interpretation would be that over this range, despite being able to reliable perform the discrimination task above chance level, participants are unable to subjectively distinguish the target interval from one in which there is either no gating information period (target absent blank) or one in which the grating information should be uninformative to the orientation discrimination decision (target absent vertical; Figure 5b). This would provide evidence for absolute blindsight on the orientation discrimination task that is free of potential contamination from criterion bias (Peters & Lau, 2015).

The inclusion of the two target absent conditions (blank vs vertical; Figure 5b) was designed to test whether or not participants use a detection heuristic to make confidence judgments when performing this task (Peters & Lau, 2015). This would entail betting on the interval that had greater grating visibility regardless of the introspective sense of the correctness of the associated orientation discrimination judgment. If this were the case, then we should predict a higher proportion of bets on the target interval in the target absent blank condition versus the target absent vertical condition. To test this question we ran a repeated measures ANOVA on percent correct scores with within-subjects factors stimulus contrast (10 levels), target absent type (blank or vertical), and response type (Type 1 or Type 2).

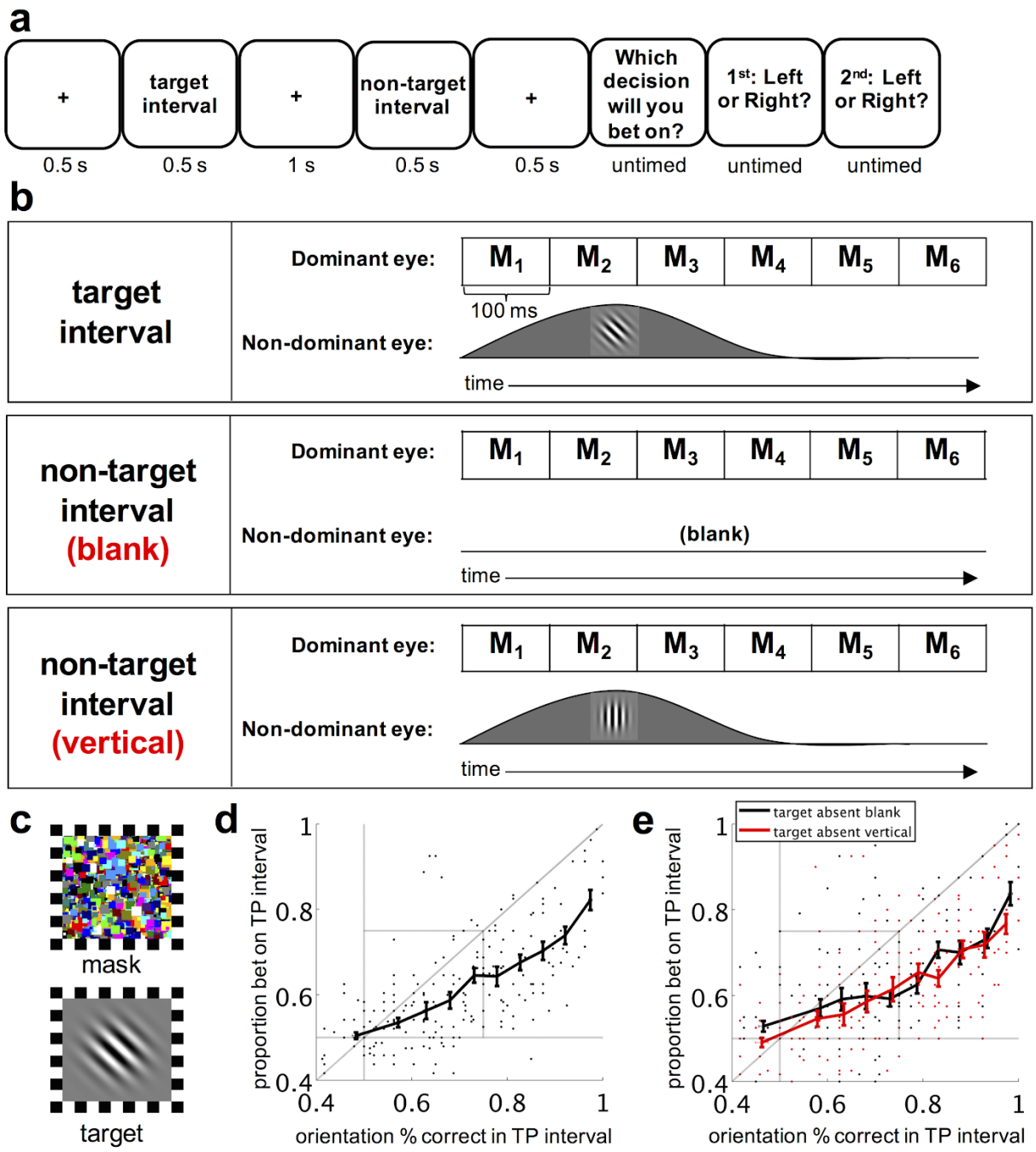

Figure 5. Experiment 1.4 procedure, stimuli, and results. **a)** Procedure. The two stimulus intervals were preceded, separated, and followed by fixation crosses with durations of 0.5, 1, and 0.5 s, respectively. Unbeknownst to participants, in the main task one interval always contained a valid left- or right-tilted target grating [target present (TP) interval] while the other contained either a vertical grating or no grating at all [target absent (TA) interval]. The order of the target present and target absent intervals was randomized across trials. Following the last fixation cross participants indicated in which interval (first or second) they felt more confident in their ability to discriminate the orientation of a grating. **b)** Stimulus intervals. All stimulus interval types (target present, target absent blank, and target absent vertical) included six contiguous

mondrian mask textures presented to the dominant eye for ~100 ms each. In target present and target absent vertical intervals a tilted or vertical grating, respectively, was presented to the non-dominant eye with the same temporal contrast ramping parameters as in the CFS intervals in Experiments 1.1 and 1.3. In target absent blank intervals no grating was presented to the non-dominant eye. **c)** Examples of target and mask stimuli. Note the difference from target and mask stimuli used in Experiments 1.1 - 1.3 (Figure 1a). **d)** "Absolute blindsight" curve. The proportion of trials in which participants indicated higher confidence in, or bet on, the target present interval are plotted as a function of orientation discrimination proportion correct scores in the target present interval. If participants can perform the orientation discrimination task unconsciously then we should see a flat portion of the absolute blindsight curve at y = 0.5 that extends up to an x-value indicating orientation discrimination performance that is significantly above chance (not observed). Mean ± s.e.m. proportion of target interval bets in each of ten evenly spaced bins from 0.5 to 1 target present interval orientation discrimination proportion correct scores are shown. Bin mean x-values correspond to the mean orientation discrimination proportion correct in each bin. Orientation proportion correct scores below 0.5 were lumped into the lowest bin (0.50-0.55) as they are assumed to reflect chance performance. **e)** Absolute blindsight curves separated by target absent interval type (black: blank, red: vertical). Data is plotted the same as in panel d.

Results & Interim Discussion

Absolute blindsight curves show that when orientation discrimination accuracy was greater than chance, the proportion of trials in which participants bet on the target present interval was also above chance. This was the case across all trials (Figure 5d) and across the individual target absent blank and target absent vertical conditions (Figure 5e). This indicates, in line with both Peters & Lau (2015) and Experiments 1.1-1.3, that there was no absolute blindsight for orientation discrimination under CFS.

The repeated measures ANOVA with within subjects factors stimulus contrast (10 levels), target absent interval type (blank or vertical), and response type (Type 1 or Type 2) showed a lack of a significant interaction between response type and target-absent interval type $F(1,21) = 3.06$, $p = 0.10$. Importantly, this suggests that participants were not using a detection heuristic when making confidence judgments. At the debriefing stage at the end of Day 2, while some participants reported noticing that on some trials target gratings appeared to be oriented

vertically, no participants reported noticing that one of the two stimulus intervals always contained an invalid target.

General Discussion

In three experiments we looked for a difference in the relationship between objective performance and subjective awareness, in line with reports of relative blindsight (Lau & Passingham, 2006), between pairs of visual suppression methods. In each case we found no evidence for any such difference, suggesting that the relationship between objective and subjective thresholds for forced-choice orientation discrimination is equivalent under FBM, CFS, IS, and BM. Taking a specific definition of subjective awareness (Giles et al., 2016), which is operationally defined as what is tracked by subjective reports while sensitivity is controlled for, we interpret the results (i.e., the POE analyses) to mean that the different suppression methods impact subjective awareness similarly.

In Experiments 1.1-1.3 we used a modified version of the 2IFC paradigm from Peters & Lau (2015) in which each of two suppression methods, one per 2IFC interval, was used to mask a left- or right-tilted target grating. Subjective awareness was indexed by forcing participants to bet on the interval in which they had higher confidence in their ability to discriminate the orientation of the target grating. This paradigm has several advantages that build on previous studies comparing different visual suppression techniques. For instance, some studies have compared suppression techniques between experiments (Almeida, Mahon, & Caramazza, 2010; Almeida et al., 2008; Almeida, Pajtas, Mahon, Nakayama, & Caramazza, 2013; Faivre, Berthet, & Kouider, 2012), making them vulnerable to potentially confounding idiosyncratic differences between experimental conditions. Further, the forced-choice nature of the subjective judgment

reduces concern about subjective criterion biases that may have been present in previous comparative suppression studies (Izatt et al., 2014; Peremen & Lamy, 2014). To further reduce subjective biases, we took inspiration from earlier studies that compared monocular and binocular suppression conditions within single experiments (Izatt et al., 2014; Jiang, Costello, & He, 2007; Stein et al., 2011) and designed stimuli such that, beyond simple differences in difficulty, the two intervals on a given trial appeared subjectively similar. This has the benefit of minimizing conscious decisional biases (e.g., participants having a conscious preference for backward masked stimuli over CFS-masked stimuli) that would otherwise reduce the chances of finding the hypothesized difference in the relative positioning of objective and subjective discrimination thresholds between suppression methods.

While Experiments 1.1-1.3 showed a lack of evidence for relative blindsight for orientation discrimination judgments under CFS versus monocular pattern masking conditions, in Experiment 1.4 we further found no evidence for absolute blindsight under CFS. We therefore interpret these findings together to suggest, in line with Peters & Lau (2015), that objective and subjective thresholds do not dissociate under any of the currently examined suppression techniques. That is to say, we consider the current results to be further evidence against the idea that normal observers have any capacity for unconscious orientation discrimination. This idea is in line with others who have argued that objective thresholds should, a priori, be considered equivalent to subjective thresholds in forced-choice perceptual tasks (Ian Phillips, 2017; Snodgrass & Shevrin, 2006). These findings also suggest that controlling for criterion bias may be a critical experimental difference between studies that report evidence for unconscious forced-choice discrimination sensitivity in normal observers (2011; Lamy, Salti, & Bar-haim, 2008; 2015) and those that report evidence against it (Peters & Lau, 2015).

The results of Experiment 1.4 also provide evidence against the idea that participants ignored instructions to rate confidence specifically in their performance on the orientation discrimination task, and instead rated confidence based on the detectability of target stimuli. Previous studies using orientation discrimination tasks have shown that in slightly different psychophysical contexts participants do in fact rate confidence based on stimulus detectability (Koizumi et al., 2015; Maniscalco et al., 2016). Therefore, the demonstration that participants are not using such a heuristic here is an important demonstration of the efficacy of the present 2IFC confidence paradigm in the investigation of visual awareness.

An important limitation is that it remains an open question whether a different visual suppression technique can selectively impair subjective awareness while leaving objective discrimination performance relatively intact. Future studies should compare visual suppression techniques that are more distant from each other in terms of how much unconscious priming they allow, e.g., FBM and visual crowding (Breitmeyer, 2015), or that have been functionally characterized to act at different points in the visual processing stream, e.g., visual crowding and object substitution (Chakravarthi & Cavanagh, 2009) or metacontrast masking and interocular suppression (Breitmeyer, Koç, Öğmen, & Ziegler, 2008). They can also focus on suppression methods that rely on attentional manipulations (e.g., attentional blink, inattentional blindness), which may allow for higher levels of unconscious processing (Kouider & Dehaene, 2007) that include unconscious forced-choice discrimination. The current paradigm provides a useful means for comparing such suppression techniques, while maintaining a rigorous control for criterion bias. However, a challenge in designing these studies will be in creating stimuli that make the techniques under comparison appear superficially indistinguishable.

It should also be emphasized that we extend the current interpretation of a lack of unconscious perception only to direct perceptual tasks such as forced-choice detection and discrimination tasks (Green & Swets, 1966; MacMillan & Creelman, 2004), and not to other established indirect perceptual effects like subliminal priming (Hannula, Simons, & Cohen, 2005; Kouider & Dehaene, 2007; though see Phillips (2017) for a discussion on whether priming effects should constitute genuine cases of perception per se). Even if we assume that normal observers do have some capacity for direct unconscious perception, our results suggest that we should not expect hierarchical relationships for subliminal priming among suppression methods (e.g., Kouider & Dehaene, 2007; Faivre, Berthet, & Kouider, 2014; Breitmeyer, 2015) to apply to direct unconscious perception. For example, Almeida et al. found greater subliminal priming effects for tool stimuli (2008, 2010) and emotional faces (2011) under BM than under CFS, while Izatt et al. (2014) found greater subliminal face priming effects under FBM than under IS. These hierarchical relationships among suppression methods for subliminal priming clearly conflict with the null results for differences in direct unconscious processing between suppression methods observed here. However, even some previously suggested hierarchical relationships between suppression methods should be approached with caution, as judgments of prime visibility in these studies were vulnerable to criterion bias (Izatt et al., 2014; Peremen & Lamy, 2014) The 2IFC paradigm described in Peters & Lau (2015) provides a means for future priming studies to ensure invisibility of primes without this potential confound.

In conclusion, we have shown a lack of a difference in the relationship between objective and subjective thresholds for forced-choice orientation discrimination between four commonly used visual suppression techniques. Taken together with previous evidence (Peters & Lau, 2015) and

the lack of an absolute blindsight effect found in our Experiment 1.4, the current results suggest that when criterion bias is sufficiently controlled for, normal observers do not demonstrate direct unconscious perception. Whether this capacity can be demonstrated under a different set of visual suppression conditions is a matter for future studies to investigate. The present results should, however, place helpful constraints on future hypotheses and methodological choices for studying conscious and unconsciousness visual perception.

VIII. The Role of Prefrontal Cortex in Visual Consciousness

Background: The Key to Consciousness May Not Be Under the Streetlight

According to a famous fable (Kaplan, 1964), one night, a drunk man was looking for his lost keys under a streetlight. As it turned out, he had lost them somewhere far away. When asked why he didn't go back to where he had lost the keys to look, he replied, "but the light is here!" Of course, seeing things clearly is easier in some places than it is in others, and in looking for the neural mechanisms for consciousness in the brain, there may be similar temptations. Specifically, neural coding is relatively straightforward and extremely sparse (Olshausen & Field, 2004) in sensory areas. Such coding can be roughly understood as having a 'labeled lines' architecture (Gross, 2002), where the representational content of individual neurons is described in terms of receptive field locations and specific features. This is in contrast to the prefrontal cortex (PFC), where neurons show a high degree of mixed selectivity (Mante et al., 2013), such that identifying perceptual content has proved to be more challenging. As such, despite ample evidence that PFC activity underlies subjective judgments in perceptual tasks (S. Dehaene & Naccache, 2001; Lau & Rosenthal, 2011), the causal status of PFC activity for consciousness is debated (Boly et al., 2017; Odegaard et al., 2017).

One point of debate concerns the observation that the link between PFC activity and consciousness is weakened when subjects do not have to attend to and report about the relevant stimuli. Specifically, under these conditions, PFC signals reflecting the difference between conscious and unconscious perception are typically reduced, especially for conventional neuroimaging measurements (Lau & Rosenthal, 2011). As such, it may look as if these signals were primarily driven by report and attention. However, PFC activity isn't entirely abolished when attention and explicit reports are omitted; using more sensitive invasive multiunit neuronal recordings, it has been shown that unreported and unattended stimulus features can be read out from PFC, nearly as well as for reported and attended features (Mante et al., 2013). Similar decoding approaches can be applied to neuroimaging data (Cortese et al., 2016). Yet, if one focuses on traditional univariate analyses for neuroimaging data, indeed it might seem like the bright streetlights are not there in PFC.

A second point of contention is that, if PFC is truly critical for conscious experiences, one may expect lesion to this region to affect some specific aspects of subjective perception. Indeed, a group study of patients with mostly unilateral PFC lesions showed a 50% decrease in their ability to correctly introspect perceptual (but not memory) content (Fleming, Ryu, Golfinos, & Blackmon, 2014). As in careful psychophysics studies, such effects were observed using near-threshold, i.e., degraded, visual stimuli. However, some have argued that these near-threshold situations are "virtually irrelevant" from the perspective of everyday conscious perception (Haun, Tononi, Koch, & Tsuchiya, 2018). It is not clear to what extent such arguments are meant to write off the meaningfulness of psychophysics for conscious perception in general. But the point may be that again, to some, the streetlights are not there in PFC.

The sensitivity of near-threshold methods may be needed, however, because unilateral PFC lesions in humans often do not always lead to the complete abolishment of functions, including ‘textbook’ PFC functions such as working memory (Curtis & D’Esposito, 2004). On the other hand, complete bilateral lesions are rare and often misidentified (Odegaard et al., 2017). For complex systems like the brain (and maybe PFC in particular), the traditional logic of using unilateral lesion methods to demonstrate absolute necessity for functions may therefore not be as straightforward as was once thought (Jonas & Kording, 2017); relatively clear cases like Broca’s area seem to be exceptions rather than the norm.

An analogy may help to illustrate this point. Suppose one builds a computational neural network using current artificial intelligence methods, with the goal of generating sentences to describe some pictures (Figure 6). Lesioning different parts of this network may lead to different levels of impairment; some lesions may afford higher degrees of ‘graceful degradation’ or fault tolerance (Achard & Bullmore, 2007) than others. It should be clear that using such information to identify the functions of different subparts of the network may therefore be misleading.

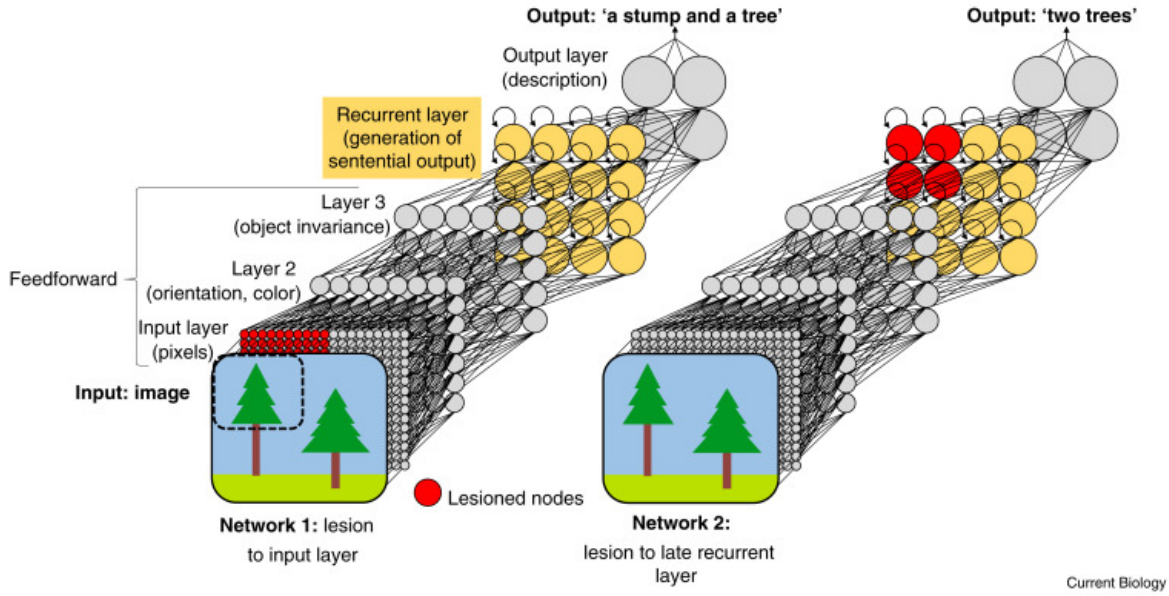

Figure 6. 'Graceful degradation' in a recurrently connected layer of a neural network. Two neural networks are designed to describe images in words. In each network, the first three layers constitute a feedforward architecture that crudely reflects the structure of a mammalian visual system. The fourth layer is a recurrent neural network, which somewhat mimics the highly recurrently connected nature of frontal and parietal cortices. Network 1 (left) contains a 'lesion' to the upper left quadrant of the input layer (red; extent of lesion as it corresponds to the input image is shown by the dotted lines). Because the lesion occurs at the input level, a quarter of the information in the image is irrevocably lost. The network thus makes the error of identifying the tree on the left side of the image as a stump. Network 2 (right) contains a lesion to the upper left quadrant of the recurrent layer. Image information in the feedforward network is preserved, and because the non-lesioned nodes in the recurrent layer are so highly interconnected, processing at this level may show limited impairment, which may be overcome with additional training. This well-known phenomenon of 'graceful degradation' may thus give the false impression that higher layers are not causally relevant.

This is of course not to say that the network in Figure 6 would be a precise model of the brain, but we can gain important intuitions by thinking of PFC as playing similar roles as the nodes at the higher levels. Even neglecting feedback from high to lower layers, which is known to be important in conscious mammalian brains, one can see that nodes at all levels contribute causally to overall function despite the varying effects of lesions. Ignoring the network structure and focusing on the lesion alone may therefore misleadingly suggest that the recurrent layer is causally irrelevant.

Likewise, this can help us understand why traditional methods of PFC electrical stimulation may not drastically modify conscious experiences [though such rare cases have been reported, e.g., (Blanke, Landis, & Seeck, 2000; Quraishi, Benjamin, Spencer, Blumenfeld, & Alkawadri, 2017)]. Stimulating nodes from some deeper or late stage layer may have a relatively small and non-specific impact on the output content due to the relatively complex connections, but this does not mean these nodes are 'content free'. Continuing the same analogy that PFC coding may be more like what happens in higher layers, this may also explain why PFC may misleadingly look like relatively unlit territory as we search for the keys to consciousness.

Admittedly, an analogy is not proof. Thankfully, a recent study (Pal & Mashour, 2018) shed some much needed empirical light on the role of this brain region in consciousness by using pharmacological stimulation that likely impacts PFC activity more broadly than focal lesions or electrical stimulations. They exposed rats to continuous sevoflurane anesthesia and tested whether perfusion of a cholinergic or noradrenergic agonist to either PFC or one of two parietal targets (posterior parietal cortex or medial parietal association cortex) could reverse anesthesia. While in all cases drug perfusion produced electroencephalographic signs of arousal and increased respiration rate, only cholinergic stimulation of PFC produced clear behavioral signs of wakefulness. Their results provide strong evidence for the causal involvement of PFC in consciousness. Hopefully, even to those who are skeptical of near-threshold perceptual effects, the importance of restoring consciousness from anesthesia cannot be denied lightly.

Of course, one caveat of this interpretation is that it elides the important difference between states of consciousness, i.e., wakefulness versus being 'knocked out', and the specific content

of subjective experiences, e.g., the redness of a tomato versus the greenness of its vine. But perhaps, because of some partially shared mechanisms, the two notions of consciousness are intimately linked; after all, supposedly one rarely experiences vivid perceptual contents when adequately anesthetized.

One network-based view for such potentially shared mechanisms is the mesocircuit model of consciousness (Schiff, 2010), which accounts for the results of several studies in which patients with disorders of consciousness showed signs of recovery following various types of mechanical or pharmacological stimulation (Figure 7).

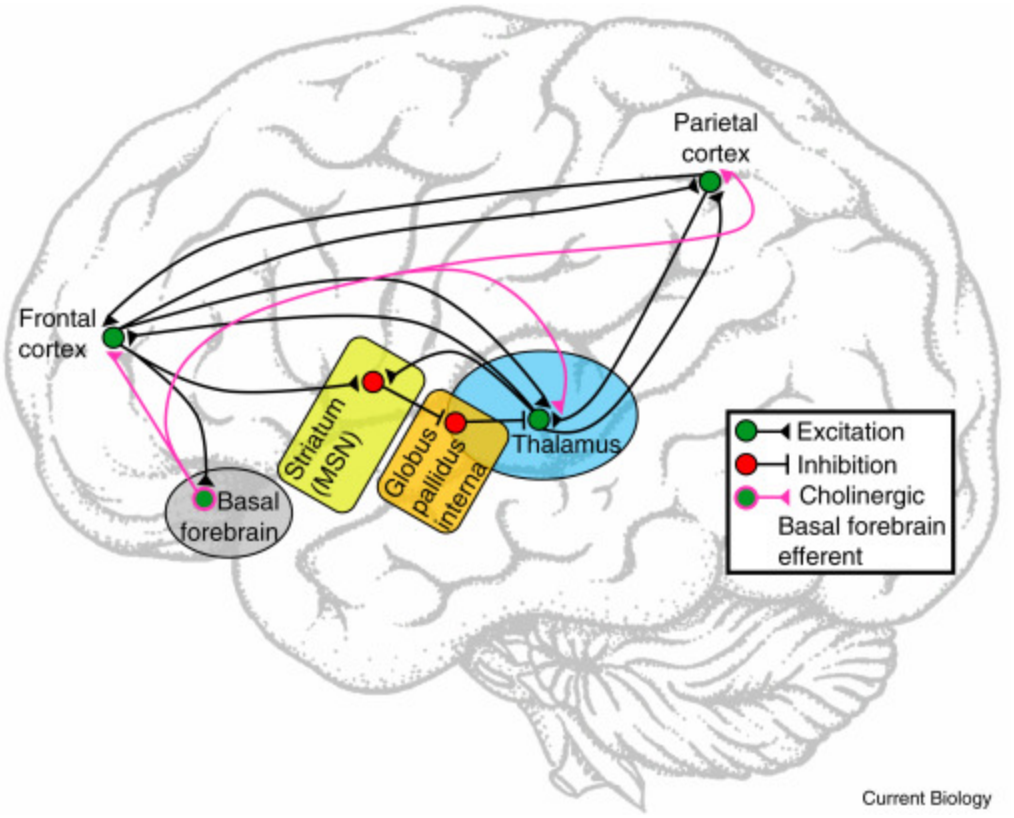

Figure 7. A modified schematic of the 'mesocircuit' that is proposed to underlie recovery of consciousness in several stimulation studies [see (Schiff, 2010)]. Frontal cortex is a critical node, receiving excitatory inputs from the thalamus and parietal cortex while sending excitatory projections to the medium spiny neurons (MSN) of the striatum, consequently disinhibiting the thalamus via inhibitory projections from the globus pallidus interna. The classic mesocircuit is

modified here to show excitatory projections from frontal cortex to the basal forebrain [(Gaykema, Van Weeghel, Hersh, & Luiten, 1991), though see also (Zaborszky, Gaykema, Swanson, & Cullinan, 1997), which highlights potentially complex inhibitory pathways too], a global cholinergic output system. Importantly, cholinergic efferents from the basal forebrain (pink) target each of the major excitatory nodes of the classic mesocircuit (thalamus, frontal cortex, and parietal cortex). Thus, through its direct connections to the basal forebrain, frontal cortex may be uniquely positioned to initiate large scale excitation of the mesocircuit.

In the context of this model, the details regarding the local acetylcholine (Ach) levels measured in each condition by Pal et al. (2018) become intriguing. Specifically, cholinergic stimulation of PFC via the agonist carbachol was unique in causing roughly a 500% increase in local Ach levels. By comparison, in all other conditions, group level increases in local acetylcholine ranged from about 20% to 75%. A possible explanation for this large difference is based on the idea that the basal forebrain, which contains the brain's major cholinergic output system, receives afferent inputs from PFC, but not parietal cortex (Gaykema et al., 1991) (Figure 7). Given this known connectivity profile, perhaps within their experimental setup, only stimulation of PFC could generate large-scale cholinergic activation (via the pink cholinergic efferents shown in Figure 7) of the three major excitatory nodes of the classic mesocircuit (frontal cortex, parietal cortex, and thalamus). This could consequently lead to strong positive excitatory feedback within the mesocircuit, which would explain the wakefulness behavior that was observed in response to stimulation of PFC, but not parietal cortex.

If this interpretation is right, it may further imply that PFC signals to the basal forebrain are more strongly activated by cholinergic as opposed to noradrenergic stimulation. Speculatively, this model would also suggest that cholinergic stimulation of PFC, implemented via reverse dialysis of carbachol as in Pal et al. (2018), might have also resulted in larger increases in parietal Ach levels than direct cholinergic stimulation of parietal cortex itself. If this were true, parietal

cholinergic activity may not be causally irrelevant; it may just be easier to trigger such activity via PFC than via direct parietal stimulation.

Of course, as keys are not always found automatically as we shine light on the ground, the experiments of Pal et al. (2018) understandably do not on their own tell us the full answer to the age old problem of consciousness. But if the above analysis is correct, then we should not write off the causal contribution of parietal cortex just yet, just as other researchers should not write off the role of PFC in consciousness based on the sheer lack of ease of observation. While Pal et al.'s results and experimental setup do not allow us to fully address all of the above hypotheses and questions, they certainly serve to motivate further studies. In sobriety, let us recognize the need for more resources and effort, to take us into these new and exciting areas that were once considered to be 'in the dark'.

Experiment 2: Multivoxel patterns for perceptual confidence are associated with false color detection

Abstract

While it has been proposed that metacognition and conscious perception are related, the exact mechanistic relationship between the two is unclear. To address this question, we combined decoded neurofeedback (DecNef) in functional magnetic resonance imaging (fMRI) with concurrent psychophysics. Participants were rewarded for activating multivoxel patterns for color discrimination confidence while they detected color in mostly achromatic stimuli. We found that occurences of voxel patterns for high color discrimination confidence were associated with

false alarms in the concurrent color detection task, suggesting a link between discrimination confidence and consciousness.

Introduction

Some current theories of consciousness posit a link between consciousness and metacognition (Stanislas Dehaene, Lau, & Kouider, 2017; Lau & Rosenthal, 2011). Intuitively, one cannot consciously see something without having some sense of certainty or uncertainty regarding what is being seen (Dienes, 2007; Fleming & Lau, 2014; Rosenthal, 2018); but see (Block, 2007). While behavioral evidence supports the idea that metacognitive judgments are a meaningful proxy for conscious experiences (Dienes & Seth, 2010; Norman & Price, 2015; Persaud, McLeod, & Cowey, 2007; Rausch & Zehetleitner, 2016; Szczepanowski, Traczyk, Wierzchoń, & Cleeremans, 2013), the extent of this support has been questioned (Norman & Price, 2015; Overgaard, Timmermans, Sandberg, & Cleeremans, 2010; Rosenthal, 2018; Sandberg, Timmermans, Overgaard, & Cleeremans, 2010). It has also been suggested that a common mechanism may underlie biases in conscious perception (e.g. conservative detection) and metacognitive misjudgments (e.g. under-confidence in discrimination); in disorders like blindsight, both seem to be problematic (Ko & Lau, 2012). And yet, these claims have so far not been directly tested.

We capitalized on the findings of a previous study in which we showed that perceptual confidence could be decoded from multivoxel fMRI patterns in lateral prefrontal and parietal cortex (Cortese et al., 2016). Pairing these patterns with reward modulated participants' reported confidence in a subsequent dot motion discrimination task. Our question here concerns whether these changes in reported confidence reflect changes in conscious experience too.

To answer this question, we rewarded participants for simultaneously activating decoded voxel patterns for both perceptual confidence in frontoparietal areas (high vs low confidence) and color perception in early visual areas (red vs green stimulus color), while they viewed a stimulus that was achromatic on the majority (> 97%) of trials. During this closed-loop fMRI procedure, we asked participants at regular intervals to report whether they saw any color in the stimulus. We found that when they falsely detected non-existent color, there was an association with occurrences of multivoxel patterns for high color discrimination confidence, supporting the link between a metacognitive process and conscious perception.

Methods

Experiment Overview

The experiment had four main stages across a total of seven days (Figure 8a): the multivoxel pattern analysis (MVPA) sessions (Days 1-2), pre-DecNef psychophysics (Day 3), DecNef (Days 4-6), and post-DecNef psychophysics (Day 7) (see Figure A1 for further details). During the MVPA sessions on Days 1 and 2 participants (N=17) performed a color lightness task with both red and green stimuli (Figure 8b) and a red/green color discrimination task with confidence judgments (Figure 8c) inside an fMRI scanner, and the resulting blood oxygen-level dependent (BOLD) signal patterns were used to train binary decoders for red versus green color and high versus low confidence, respectively. During the DecNef stage participants performed a real-time neurofeedback task in which they were rewarded for activating decoded multivariate BOLD signal patterns corresponding to redness in visual cortex and high confidence in frontoparietal cortex. To examine whether activation of decoded color and confidence patterns had any correspondence with real-time color perception, participants performed a concurrent color

detection task during DecNef. Finally, to examine whether this neurofeedback manipulation had any effect on red/green color discrimination (Amano, Shibata, Kawato, Sasaki, & Watanabe, 2016), participants performed the same red/green color discrimination task as in the MVPA sessions outside of the scanner during the pre- and post-DecNef psychophysics stages. Days 2 and 3 always occurred on separate calendar weeks, and were thus always separated by at least two days. Days 3-7 were always consecutive.

Participants

Seventeen subjects (2 female, mean ± SD age: 26.0 ± 7.5 years, 2 left-handed) with normal or corrected-to-normal vision participated in the decoder construction stage on Days 1 and 2. Two participants were excluded from analyses following the decoder construction stage for not having accuracies greater than 55% for both the color decoder and at least 2 of the 4 confidence decoders (Cortese et al., 2016). Thus, 15 subjects (1 female, mean ± SD age: 25.4 ± 7.1 years, 2 left-handed) are included in the analyses for the behavioral and DecNef tasks on Days 3-7. The study was conducted at the Advanced Telecommunications Research Institute International (ATR) and was approved by the Institutional Review Board of ATR. All subjects gave written informed consent.

Red/Green Color Discrimination Task

The red/green color discrimination task (Figure 8c) was performed both outside and inside of the scanner on Days 1 and 2, and outside of the scanner only on Days 3 and 7 (Figure A1). At the start of each trial a white fixation circle (diameter ~0.43°) was presented for 1 s on a gray background (rgb[64 64 64]). A circular vertical grating (diameter ~13.5°) and a black annulus (diameter ~0.85°), both centered around the white fixation circle, then appeared for 0.5 s (Figure

8c). The black vertical bars within the grating had a width of ~0.64°, with the area between them subtending the same visual angle.

The majority of pixels in the areas between the black bars had grayscale RGB triplet values (i.e., all RGB channel values were equal) that varied randomly on each frame (frame duration = 16.67 ms) with a mean channel value of 120 and a standard deviation of 51.2. The area between black bars in the grating was thus dynamic. Color strength was adjusted by setting the color of a variable proportion of pixels in the areas between the black bars to either a red or green RGB triplet. Subject-specific RGB triplets were computed from a flicker fusion task (Simonson & Brozek, 1952) performed at the beginning of the Day 1 to ensure psychophysical equiluminance (Supplementary Material). These triplets were fixed for all red/green color discrimination and color detection tasks used throughout the rest of the experiment. The locations of colored pixels varied randomly between frames, but the proportion of colored pixels was constant throughout a given trial.

Offset of the grating was followed by a 1.5-s decision period in which only the fixation circle remained on the screen. Participants were then asked to report the color of the grating (red or green) and to indicate their confidence in their decision on a scale from 1 to 4 per the following instructions: 1 corresponded to a guess, 2 corresponded to having low but non-zero confidence, 3 corresponded to having moderately high confidence without being certain, and 4 corresponded to feeling certain in their decision. Participants had two seconds to make each response. The on-screen locations of each response option (‘Red’ and ‘Green’ for the color judgment and ‘1’,’2’,’3’, and ‘4’ for the confidence judgment) were randomized on each trial. For all iterations of this task that occurred outside of the fMRI scanner on Days 1 and 2 trial-by-trial

feedback (1 s) was given in the form of a green (rgb[0 255 0]) “+1” for correct discrimination responses or a red (rgb[255 0 0]) “-1” for incorrect discrimination responses. The ITIs for this task where 1 s and 5 s when performed outside and inside of the fMRI scanner, respectively.

On both Day 1 and Day 2, prior to the decoder construction session participants performed 80 trials of an adaptive version (QUEST, (Andrew B. Watson & Pelli, 1983)) of the red/green color discrimination task. The adaptive procedure used two interleaved 40-trial staircases to estimate the stimulus strength (in proportion of colored pixels) that would lead to 75% correct accuracy on the task. These procedures were broken down into two 40 trial blocks. The mean of the two 75% correct threshold estimates on Day 1 was used as starting stimulus strength for the red/green discrimination task in the subsequent Day 1 decoder construction session in the scanner. The mean of the two 75% correct threshold estimates on Day 2 was used to determine the stimulus strengths that would be used for the pre- and post-DecNef psychophysics tasks on Days 3 and 7 (see below).

Color Lightness Task

The color lightness task (Figure 8b) was performed both inside and outside of the scanner on Day 1, and inside the scanner on Day 2 (Figure A1). On each trial, a fixation circle with the same parameters as that in the red/green color discrimination task appeared for 1 s. A colored grating stimulus (either red or green) then flashed for 0.5 s durations at 1 Hz and its color lightness either increased or decreased linearly over a period of 6 s (6 presentations in total).

The grating stimulus had the same parameters as that in the red/green color discrimination task except for the following. In the area between the black vertical bars, all pixels were colored

(either all red or all green). On each trial, a set of 6 different equally spaced values for the dominant RGB channel was determined. This set had a variable range across trials but a constant mean equal to the dominant channel value in the corresponding RGB triplet determined by the flicker fusion task. On each frame of a given 0.5-s grating presentation, the dominant RGB channel value of a given colored pixel was drawn from a normal distribution with a mean of the corresponding set value and a SD of 51.2. The value of the non-dominant RGB channels for a given colored pixel was determined by taking the difference between the dominant and non-dominant rgb channel values from the relevant RGB triplet determined by the flicker fusion task, and subtracting it from the dominant RGB channel value for that pixel. Thus, for each subject the difference between dominant and non-dominant RGB channel values per color was constant across all colored pixels for all grating presentations in this task.

After grating presentation there was a 2-s decision period in which only the fixation circle remained on the screen. Participants then had 2 s to indicate whether the flashing grating stimulus increased or decreased in lightness over time. Because the color tended to look more saturated when lightness decreased, the response options for decreases and increases in lightness were “more color” and “less color”, respectively (Figure 8b).

Participants performed 16 trials of this task outside of the scanner on Day 1. On these lightness trials, but not those performed inside the scanner, they received the same trial-by-trial feedback for correct and incorrect judgments as they did in the color discrimination task. The first four trials were designed to familiarize subjects with the task, and thus each stimulus in these trials used a large range of lightness values (154 RGB units). Trials 5 to 16 employed a 1-up 1-down staircasing procedure with variably weighted step sizes (Kingdom & Prins, 2010) of 7.68 and

2.56 RGB units, respectively. The starting range of lightness values on trial 5 was 15 RGB units. The mean of all of the lightness range values from all staircased trials in which a reversal (i.e., a correct response following an incorrect response or vice versa) occurred was set as the midpoint of the uniform distribution of potential range values used in the first color decoder run of the subsequent Day 1 decoder construction session (see Supplementary Material).

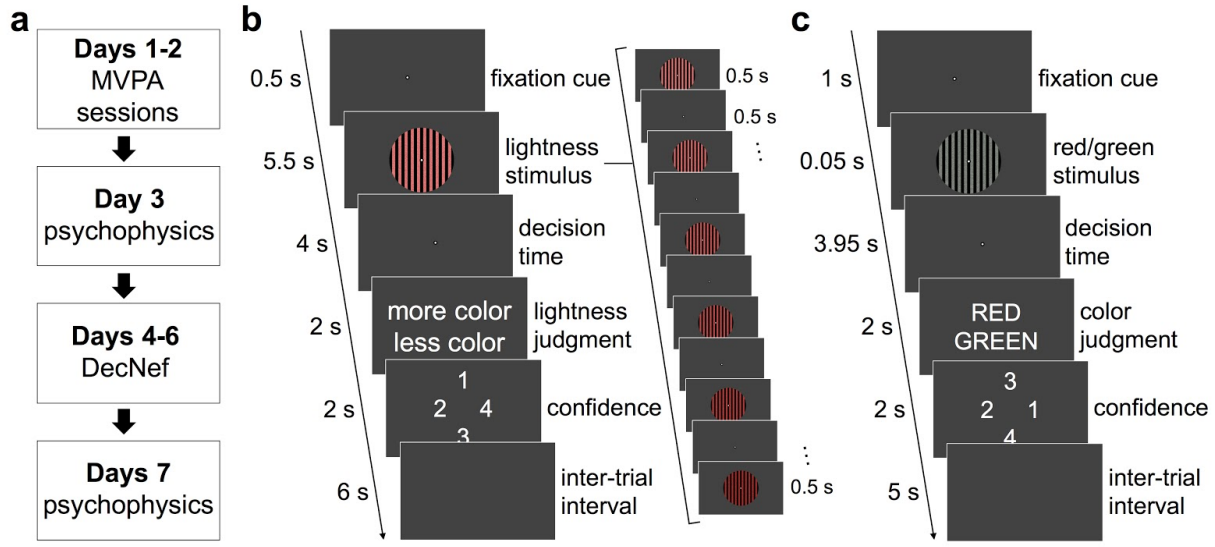

Figure 8. Experiment 2 overview and decoding tasks. **a)** Experiment flowchart. On Days 1 and 2, participants performed color lightness and color discrimination tasks in the fMRI scanner to decode multivoxel patterns for color and confidence, respectively. On Days 3 and 7 participants performed a red/green color discrimination task outside of the fMRI scanner. On Days 4-6 participants performed a DecNef task in which they were rewarded for the simultaneous activation of multivoxel patterns for red color in visual cortex and high confidence in frontal and parietal cortex. **b)** Color decoder task. Participants viewed 6 circular colored (either all red or all green on a given trial; see Methods) vertical gratings presented for 500ms each and flashed at a frequency of 1 Hz (6 s total). Grating color lightness either increased or decreased (shown) with successive presentations. After a 2 s decision period, subjects indicated whether color lightness increased or decreased by selecting the “less color” or “more color” options, respectively. **c)** Confidence decoder task. After a 1 s fixation period, a colored grating (red or green) was presented for 500 ms. After a 1.5 s post-stimulus interval, participants indicated whether they perceived the grating to be red or green and rated confidence on the color discrimination task from 1 (guessed) to 4 (certain). The same task was used in the psychophysics sessions outside of the fMRI scanner on Days 3 and 7. Text in panels b and c is enlarged compared to its actual size during the experiment for clarity.

Color and confidence MVPA

Color and confidence decoders were trained on multivoxel BOLD signal patterns acquired while participants performed the color lightness and red/green color discrimination tasks, respectively. The color decoder was trained on voxel activities in a region of interest (ROI) spanning visual areas V1, V2, V3, and V4 (denoted hereafter as V1-4). Separate confidence decoders were trained on voxel activities in each of four frontoparietal ROIs: inferior parietal lobule (IPL), inferior frontal sulcus (IFS), middle frontal gyrus MFG, and middle frontal sulcus (MFS) (Figure 9). Each task was performed in separate 16-trial runs. Color (lightness task) and confidence (red/green color discrimination task) runs alternated consecutively for each participant, with the order pseudorandomized across participants. Participants performed as many of each run as possible across two 90 minute scanning sessions on Days 1 and 2 (mean ± SD across subjects: 9.5 ± 0.9 color runs and 9.9 ± 0.8 confidence runs).

Iterative sparse logistic regression (Yamashita, Sato, Yoshioka, Tong, & Kamitani, 2008) was used to select and weight the most informative voxels for distinguishing red vs green color in the visual ROI and high vs low confidence in the four frontoparietal ROIs as previously described (Amano et al., 2016; Cortese et al., 2016). Decoding accuracy was validated using an iterative leave-one-run-out procedure. For each cross validation run, the SLR algorithm selected and weighted a subset of voxels in the relevant ROI. These voxels were then removed, and the algorithm was applied again, selecting and weighting a new, unique subset of voxels. This process was repeated iteratively, 10 times for each cross validation run. Decoding accuracies were then averaged across cross validation runs for each iSLR iteration, and the number of

iterations that led to the highest decoding accuracy was selected as the optimal number to be subsequently used during DecNef.

Following the cross validation procedure a separate training run was performed on the entire dataset using the optimal number of iterations. The resulting decoder was used for the subsequent DecNef sessions on Days 4-6. The output of the color decoder reflected the probability of the participant viewing a red stimulus, while the output of the confidence decoder reflected the probability of the participant being in a state of high perceptual confidence. See Supplementary Material for further details on color and confidence decoder optimization.

DecNef sessions

All participants in the MVPA session who had accuracies of 55% or higher for color decoding and for at least two of the frontoparietal ROIs for confidence decoding (N=15) were included in the DecNef sessions on Days 4-6. Each DecNef run (mean ± s.e.m. = 9.6 ± 0.4 runs per day) started with an initial 29 second fixation period, during which a white fixation cross (diameter ~0.84°) was presented at the center of the screen. This was followed by 16 trials in which participants were rewarded for activating the patterns identified in the MVPA session as corresponding to red in V1-V4 and high confidence in the four frontoparietal ROIs (Figure 10a). On each trial, after a 1 s cue, participants viewed a vertical grating with the same dimensions as the gratings shown during the MVPA session for 6 s, during which time they were instructed to try to use their minds to activate a pattern of brain activity in order to make the size of a subsequent feedback stimulus (a black disc) as large as possible. The feedback disc appeared for 2 s after a 6 s rest period (Figure 10a).

For online decoding, the BOLD signal was head motion corrected in real time using Turbo-Brian Voyager software (Brain Innovation, Netherlands). The BOLD signal corresponding to the interval from the start of the run until the last measured TR in the prevailing trial was then extracted from the voxels that were selected in each ROI during the MVPA session. Linear detrending and z-score normalization was then performed on these extracted voxel activities. The resulting detrended, z-score normalized signal in each ROI was then averaged across the 6 second rest period of the prevailing trial, which should correspond to neural activity during the 6 second induction period when adjusting for an estimated 6-s hemodynamic delay, and was inputted into the corresponding color or confidence decoder. The resulting decoding likelihoods (LLs) determined the size of the feedback disc according to the following formula:

$$0.667 * [(LL_{red} / 2) + (LL_{high\ confidence} / 2)] + 0.333 * (LL_{red} * LL_{high\ confidence}).$$

The size of the feedback disc also corresponded to a monetary reward earned on each trial (max = 18.75 yen or approximately $0.15 US dollars per trial). Successive trials were separated by a 5 s ITI.

At the end of each DecNef run participants separately reported whether they perceived any red or any green in the induction grating stimulus on any of the 16 trials in that run, and indicated how confident they were in this judgment on the same 1 to 4 scale that was used during decoder construction trials (Figure 10b). The order in which the red and green perception questions were asked at the end of each DecNef run was randomized across runs. Importantly, on 97.4 ± 0.2% of trials, the induction stimulus was achromatic, while on the remaining trials (4 per day, the induction stimulus was either slightly red (2 trials) or slightly green (2 trials).

Specifically, on each day of neurofeedback one run contained one red trial, a different run contained one green trial, and a third run contained both one red and one green trial. Run order was randomized between subjects, but the three runs containing color trials were constrained to always occur within the first 8 runs on a given day to avoid a given subject missing a run with a color trial due to time constraints.

Given this setup, each run can be categorized as one of four classic types according to signal detection theory: hits (reported seeing color during a run in which at least one trial contained a colored induction stimulus), misses (reported seeing no color during a run in which at least one trial contained a colored induction stimulus), false alarms (reported seeing color during a run in which no trials contained a colored induction stimulus), and correct rejections (reported seeing no color during a run in which no trials contained a colored induction stimulus). The color manipulation was designed to induce a nonzero baseline false alarm rate for perceiving color, e.g., reporting the perception of red in a given DecNef run when no red was present in any induction stimuli during that run.

At the end of Day 7, participants were asked two debriefing questions. First, they were asked whether they thought they received real or sham neurofeedback. Second, they were asked to guess, assuming they had been receiving real neurofeedback, whether they were rewarded for activating a pattern of brain activity corresponding to red perception or green perception. For additional details on the decoded neurofeedback procedure, see Supplementary Material.

Pre-/post-DecNef psychophysics

On Days 3 and 7 participants (N=15) performed the same red/green color discrimination task from Days 1 and 2, with the following differences in stimulus parameters. Three stimulus levels (proportion of colored pixels), fixed across color to preserve equiluminance, were used to target percent correct scores of 65%, 75%, and 85%. As in the red/green color discrimination tasks performed outside of the scanner on Days 1 and 2, the ITI was 1 s.

Participants first performed 10 practice trials with trial-by-trial feedback (as described in the Red/Green Color Discrimination Task section above). They then performed 6 blocks of 51 trials each with self-paced breaks between blocks and no trial-by-trial feedback. Of the 306 total trials, 276 had stimulus strengths near perceptual threshold, with 46 trials at each of the three near-threshold stimulus strengths for each color. Of the remaining 30 trials, 15 had a high percentage of colored pixels (45%), which was intended to help maintain perceptual templates for color, and 15 had zero colored pixels. All trial types were randomly interleaved across blocks.

The three stimulus strengths were determined for each participant by multiplying their mean Quest-estimated threshold stimulus strength (in proportion of colored pixels) from Day 2 by three proportions (mean ± SD proportions across subjects = 0.57 ± 0.17, 1.06 ± 0.10, 1.56 ± 0.16 for low, medium, and high stimulus strengths, respectively). These proportions were adjusted on a subject-by-subject basis according to the Quest procedure's tendency to over- or underestimate threshold stimulus strength when considering all of the across-subject data that had been collected at the time. The resulting mean ± s.e.m. performance scores across Days 3 and 7 for low, medium, and high stimulus strengths were 65.0 ± 2.2%, 77.0% ± 2.7%, and 84.0% ± 2.4% correct (d' = 1.00 ± 0.16, 1.83 ± 0.23, and 2.37 ± 0.22, respectively).

Individual participant data from each day were fit with cumulative normal psychometric functions with free parameters α (threshold) and β (slope), and fixed parameters γ (lapse rate) = 0 and δ (guess rate) = 0 also using the Palamedes toolbox (Kingdom & Prins, 2010; Prins & Kingdom, 2018). Values on the abscissa were equated across subjects to equal ±1, ±2, and ±3 to reflect the low, medium, and high stimulus strength conditions for each color, respectively, with positive and negative values corresponding to red and green stimuli, respectively. The mean and s.e.m. of the individual psychometric fits are shown by the dark lines and lighter bounded regions, respectively (Figure 11c). The point of subjective equality, which corresponds to the stimulus strength at which participants are equally likely to choose red or green on the color discrimination task (i.e., 50% on the ordinate), was estimated as the threshold parameter, α, from the fitting procedure. Given that stimulus strength values were equated across subjects for psychometric curve fitting, the resulting mean ± SD point of subjective equality (PSE) values ($PSE_{pre\text{-}DecNef}$ = 1.66 ± 2.00, $PSE_{post\text{-}DecNef}$ = 0.34 ± 0.72) can be thought of as the proportion of the lowest stimulus strength necessary for the stimulus to be equally likely to have a subjective appearance of redness or greenness, with positive values reflecting red stimulus strength and negative values reflecting green stimulus strength.

Apparatus

Stimuli for tasks performed outside of the fMRI scanner were presented on an IBM P275 CRT monitor with a 1280 x 960 resolution and a 60 Hz refresh rate. All visual stimuli were generated with custom Matlab R2014a (Natuck, MA) scripts using PsychToolbox 3.0.12. Stimuli for tasks performed inside of the fMRI scanner were presented on an LCD projector that also had a 1280 x 960 resolution and a 60 Hz refresh rate. Repeated measures ANOVAs were performed using

SPSS v22 and were adjusted for violations of the assumption of sphericity with the Greenhouse-Geisser correction when necessary.

MRI Parameters

MRI images were acquired using 3T MRI scanners (Siemens, Verio [N=15] or Siemens, Trio [N=2]) at the ATR Brain Activation Imaging Center. Both scanners used head coils. Functional images for MVPA and DecNef sessions were acquired using gradient EPI sequences with 33 contiguous slices (repetition time (TR) = 2 s, echo time (TE) = 26 ms, flip angle = 80 deg, voxel size = 3 x 3 x 3.5 mm^3, 0 mm slice gap) oriented parallel to the AC-PC plane, covering the entire brain. T1-weighted MR images (MP-RAGE; 256 slices, TR = 2s, TE = 26 ms, flip angle = 80 deg, voxel size = 1 x 1 x 1 mm^3, 0 mm slice gap) were also acquired during the first MVPA session. These images were used for automatic brain parcellation in Freesurfer (Fischl et al., 2002).

fMRI preprocessing

fMRI images from decoder construction sessions were preprocessed as previously described (Cortese et al., 2016). T1-weighted structural images were processed with an automatic parcellation procedure based on volumetric segmentation and cortical reconstruction using the FreeSurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/). The IPL, IFS, MFG, MFS ROIs (Figure 9) used in subsequent analyses were defined using this procedure. Visual ROIs were defined using a probabilistic atlas (L. Wang, Mruczek, Arcaro, & Kastner, 2015). Average inflated cortical surfaces shown in Figure 9 were generated using Freesurfer and displayed using PySurfer (https:// pysurfer.github.io/). Average ROIs in Figure 9 were generated in Freesurfer for display purposes; voxels were included in each average ROI if they were present

in the individual ROIs of at least half of the 17 decoder construction participants (Figure 9a). Gray matter masks were generated using the mrVista software package for Matlab (http://vistalab.stanford.edu/software/), which uses functions from the SPM suite (http://www.fil.ion.ucl.ac.uk/spm), to ensure that only gray matter voxels were used for subsequent analyses. Three-dimensional rigid-body motion correction was applied in mrVista to align functional scans to the T1-weighted structural image for each participant. Day 2 Localizer scans were slice-time corrected and averaged across stimulus groups, and a coherence analysis was applied to identify voxels in visual cortex that responded maximally to the localizer stimulus (Wandell & Winawer, 2011). No temporal or spatial smoothing was applied. For all color and confidence decoder construction scans, we removed voxels with exceptional values, extracted BOLD signal time courses from each remaining voxel in each ROI, applied linear detrending, and z-score normalized the BOLD signal per run to account for potential baseline differences between runs.

Results

Decoding color and confidence

Based on 10-fold cross validation, color decoding accuracy (mean ± s.e.m.) in visual cortex was 71.5 ± 1.7%, while the mean confidence decoding accuracy across frontoparietal ROIs was 64.9 ± 1.4% (IPL: 65.9 ± 1.5%, IFS: 66.2 ± 1.6%, MFG: 64.9 ± 1.4%, MFS: 62.6 ± 1.2%; Figure 9). Decoding accuracy in each ROI was significantly greater than chance (50% correct) [V1-4: $t(16) = 12.5$, $p < 0.001$, 95% CI = (0.68,0.75); IPL: $t(16) = 10.3$, $p < 0.001$, 95% CI = (0.63,0.69); IFS: $t(16) = 10.2$, $p < 0.001$, 95% CI = (0.63,0.70); MFG: $t(16) = 10.6$, $p < 0.001$, 95% CI = (0.62,0.68); MFS: $t(16) = 10.3$, $p < 0.001$, 95% CI = (0.60,0.65); Bonferroni corrected ($\alpha_{corrected} = 0.01$), two-tailed one-sample t-tests]. The mean ± s.e.m. numbers of selected voxels in each

ROI were the following: V1-V4 = 140.4 ± 23.0, IPL = 107.4 ± 13.1, IFS = 71.5 ± 12.1, MFG = 80.4 ± 14.6, MFS = 79.1 ± 12.6). Two participants were excluded from subsequent DecNef analyses for failing to have accuracies greater than 55% for color decoding and for at least two of the four frontoparietal confidence decoders.

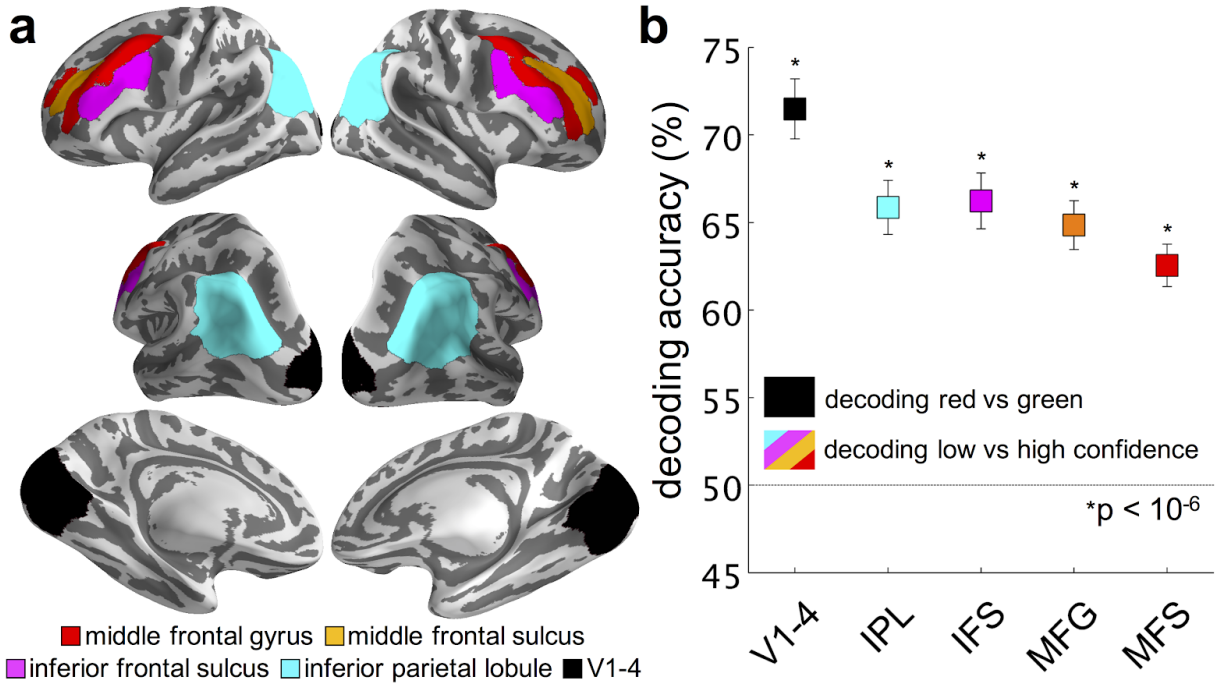

Figure 9. Decoding ROIs and accuracies. **a)** Average regions of interest (ROIs) used for decoder construction. Voxels were included in each of the displayed average ROIs if they were present for at least half of the 17 decoder construction participants. ROIs are displayed on an average (N=17) inflated cortical surface using the Freesurfer and PySurfer software packages. **b)** Decoding accuracies (reported as mean ± s.e.m., N=17). All decoders were trained using sparse logistic regression(Yamashita et al., 2008) and tested with 10-fold cross-validation. Color decoding (red vs green) accuracy in V1-V4: 71.5 ± 1.7%. Confidence decoding (high vs low) accuracy in IPL: 65.9 ± 1.5%, IFS: 66.2 ± 1.6%, MFG: 64.9 ± 1.4%, MFS: 62.6 ± 1.2%. All decoding accuracies were significantly higher than chance (50% correct) as measured by Bonferroni corrected ($\alpha_{corrected}$ = 0.01), two-tailed one-sample t-tests. V1-4: combined visual areas V1, V2, V3, & V4; IPL, inferior parietal lobule; IFS, inferior frontal sulcus; MFG, middle frontal gyrus; MFS, middle frontal sulcus.

The color manipulation in the induction stimulus succeeded in establishing a non zero baseline false alarm rate (FAR) for red perception in 9 of 15 DecNef participants (FAR_{red} = 14.2% ± 2.2%)

and for green perception in 14 of 15 DecNef participants (FAR_{green} = 23.0% ± 5.9%). One participant was excluded from analyses of false alarm and correct rejection runs because they did not make any false alarms. One additional participant was excluded from these analyses due to a failure to make correct rejections for both red and green responses on any single run. Thus, in 13 DecNef participants, we could analyze whether there was any connection between activation of multivoxel patterns for color or confidence and false color perception by comparing color and confidence induction likelihoods between false alarm and correct rejection runs.

In subjects who made red false alarms, there was no significant difference in color induction likelihoods between red false alarm runs and correct rejection runs [$t(8) = -0.34$, $p = 0.75$, 95% CI = (-0.24, 0.18), two-tailed paired-samples t-test; Figure 11a]. In subjects who made green false alarms there was no significant difference in color induction likelihoods between green false alarm runs and correct rejection runs [$t(12) = -0.34$, $p = 0.74$, 95% CI = (-0.08, 0.06), two-tailed paired-samples t-test; Figure 11a]. However, collapsing across color, high confidence induction likelihoods were significantly higher during false alarm runs than they were during correct rejection runs [$t(12) = 2.75$, $p = 0.02$, 95% CI = (0.01,0.06), two-tailed paired-samples t-test; Figure 11b].

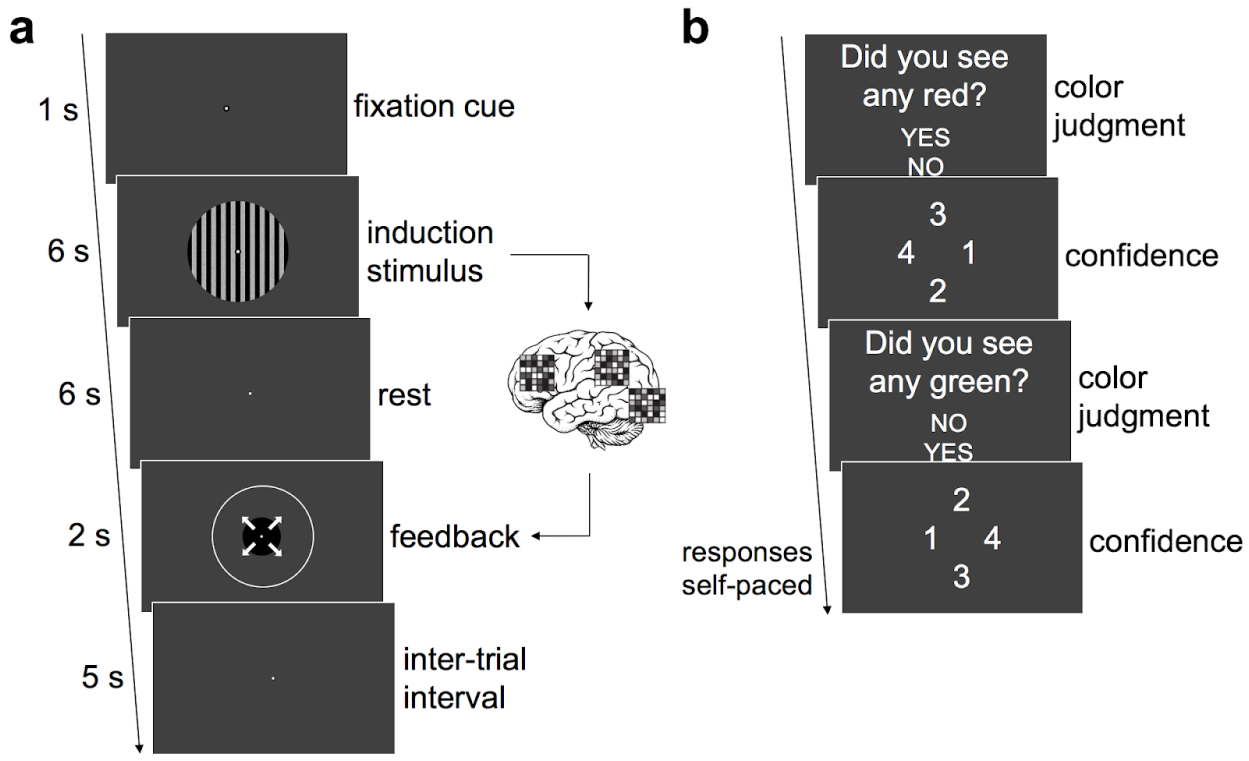

Figure 10. DecNef task. **a)** Trial structure. Participants were told that, after a 1 s cue, while an induction stimulus (vertical grating) was present, they should try to activate a pattern of brain activity so as to maximize the size of a subsequently presented black feedback disc. A 6 s rest period separated the induction and feedback stimuli to account for hemodynamic delay and real-time processing of fMRI images. BOLD signal in the visual (V1-4) and frontoparietal (IPL, IFS, MFG, MFS) ROIs from the induction period was processed by the previously-trained color and confidence decoders, respectively. The magnitude of the resulting red and high confidence likelihoods determined the size of the feedback disc such that participants were maximally rewarded for simultaneously activating a red pattern in visual cortex and a high confidence pattern in frontoparietal cortex (see Methods). **b)** End of run questions. At the end of each run participants were asked whether they perceived red or green in the induction grating during any of the 16 trials in that run. They were also asked to rate confidence in each of these judgments on a scale from 1 (low) to 4 (high). Text in panel b is enlarged compared to its actual size during the experiment for clarity. V1-4: combined visual areas V1, V2, V3, & V4; IPL, inferior parietal lobule; IFS, inferior frontal sulcus; MFG, middle frontal gyrus; MFS, middle frontal sulcus.

If the learned co-activation of the decoded red and high confidence patterns can influence conscious visual perception in real time, then we might predict the false alarm rate (FAR) for red perception during neurofeedback to increase across the three days of neurofeedback. However, a repeated measures ANOVA with within-subjects factor time (3 days) showed that the red FAR

did not increase over time [$F(1.23,17.21) = 0.55$, $p = 0.50$]. Correspondingly, a repeated measures ANOVA with a within-subjects factor of time suggested that neither red nor high confidence induction likelihoods increased over the three days of DecNef (red induction: $F_{2,28} = 0.63$, $p = 0.54$; high confidence induction: $F(2,28) = 0.79$, $p = 0.46$). Mean induction likelihoods across days were 0.47 ± 0.04 and 0.52 ± 0.01 for red and high confidence induction, respectively. Because participants did not successfully learn to perform the neurofeedback task we are precluded from addressing whether DecNef learning in this context can affect real-time color perception.

All DecNef participants (N=15) also performed a red/green color discrimination task outside of the fMRI scanner on the day before the first DecNef session and on the day after the last DecNef session. The color discrimination task was the same as that used for confidence decoding in the MVPA sessions (Figure 8c) except that three constant stimulus strengths were used for each color (whereas stimulus strength was modified on a per run basis during the MVPA sessions; see Methods). The resulting individual participant data from each day were fit with cumulative normal psychometric functions (Figure 11c). In line with a previous DecNef study of color perception (Amano et al., 2016), participants were significantly more biased towards choosing red after DecNef, as indicated by a significant leftward shift in the point of subjective equality (PSE) [$t(14) = 2.43$, $p = 0.03$, 95% CI = (0.16, 2.48), two-tailed paired-samples t-test; Figure 11c].

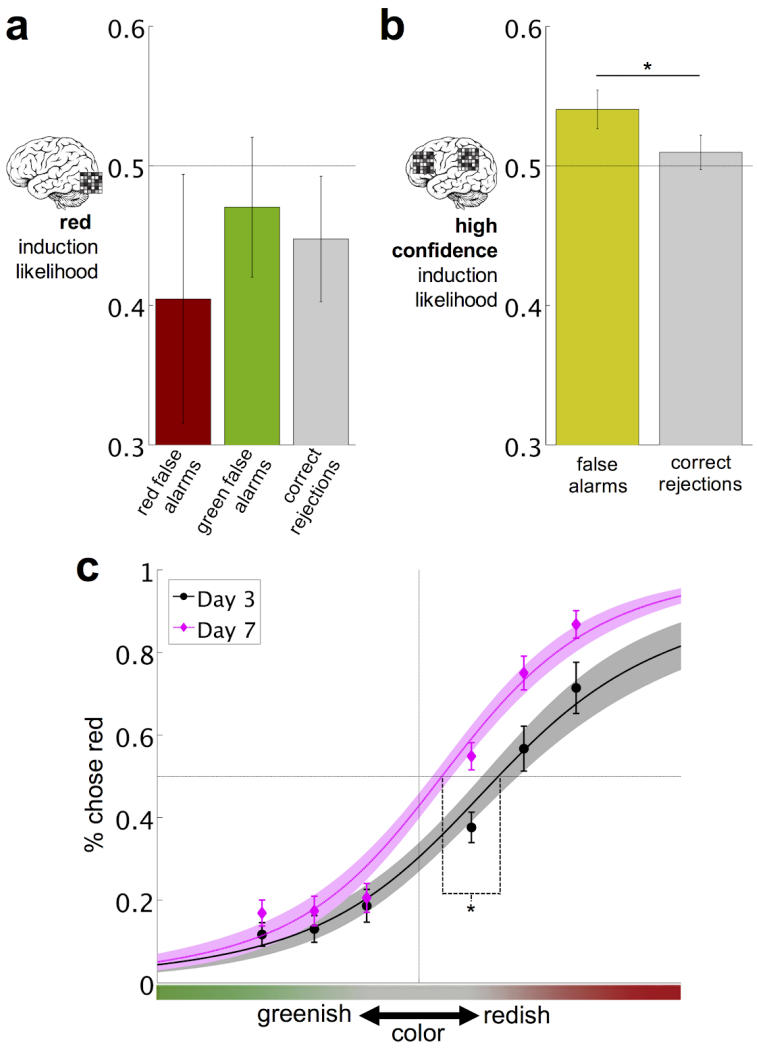

Figure 11. Induction likelihoods during false alarm versuss correct rejection runs and pre-/post-DecNef psychometric functions for color discrimination. **a)** Red induction likelihoods during red false alarm, green false alarm, and correct rejection runs. Induction likelihoods were not significantly different between false alarms and correct rejections for either color [red false alarm versus correct rejection runs: $t(8) = -0.34$, $p = 0.75$, 95% CI = (-0.24, 0.18); green false alarm versus correct rejection runs: $t(12) = -0.34$, $p = 0.74$, 95% CI = (-0.08, 0.06), two-tailed paired-samples t-tests]. **b)** High confidence induction likelihoods during false alarm runs (collapsed across color) and correct rejection runs. High confidence induction was higher under false alarm runs than it was under correct rejection runs [$t(12) = 2.75$, $p = 0.02$, CI = (0.01,0.06)]. **c)** Pre- vs post-DecNef psychometric curves. Participants performed a red/green color task (see Figure 8c) with the method of constant stimuli on the day before the first DecNef session (Day 3, gray) and on the day after the last DecNef session (Day 7, purple). Psychometric curves were fit to individual participant data using cumulative normal distribution functions. Shown is the mean (black/magenta line) ± s.e.m. (gray/light magenta shaded area) of the individual fits. A significant negative shift in group mean point of subjective equality (PSE) was observed from Day 3 to Day 7 [$t(14) = 2.43$, $^*p = 0.03$, two-tailed paired-sample t-test], showing a post-DecNef reduction in an initial bias toward choosing green.

The signal detection theoretic measures d' and criterion (Green DM, 1966; MacMillan & Creelman, 2004), c, were computed for the color discrimination task with Hits corresponding to trials in which participants correctly discriminated red stimuli as red and False Alarms corresponding to trials in which participants incorrectly discriminated green stimuli as red. There was a trend, though non-significant, toward an overall improvement in discrimination d' from Day 3 to Day 7 [d'_{pre} = 1.48 ± 0.18 , d'_{post} = 1.79 ± 0.22, t(14) = -1.91, p = 0.08, 95% CI = (-0.66, 0.04), two-tailed paired-samples t-test] which may be attributable to perceptual learning (Dosher & Lu, 2017). There was also a non-significant trend towards a decrease in criterion [c_{pre} = 0.50 ± 0.12, c_{post} = 0.20 ± 0.08, t(14) = 1.99, p = 0.07, 95% CI = (-0.02, 0.61), two-tailed paired-samples t-test], which is consistent with the observed negative shift in PSE from the psychometric function analyses. There was no significant change in red/green color discrimination confidence from Day 3 to Day 7 [mean $confidence_{pre}$ = 2.13 ± 0.14, mean $confidence_{post}$ = 2.18 ± 0.13, t(14) = -0.54, p = 0.60, 95% CI = (-0.28,0.17), two-tailed paired-samples t-test].

During the debriefing session following the red/green color discrimination task on Day 7, eight out of 15 DecNef participants (53.3%) indicated on the forced-choice question that they thought they had been receiving sham neurofeedback. Further, only five out 15 DecNef participants (33.3%) responded that, assuming they had been receiving real neurofeedback, they were rewarded for activating a pattern of brain activity corresponding to red perception. These results suggest that participants were unaware of the true targets of neurofeedback.

Discussion

Using a DecNef paradigm that targeted activation of multivariate decoded patterns for color in visual cortex and perceptual confidence in frontoparietal cortex, we found that participants were more likely to activate patterns for high confidence during fMRI runs in which they also falsely perceived color. Activation of decoded patterns for color in visual areas, on the other hand, was not associated with false color perception. These results suggest that the decoded pattern for color discrimination confidence is critically related to conscious visual perception.

This result also provides support for confidence judgments in the extant debate about which subjective measure is ideal for measuring conscious awareness (Overgaard et al., 2010; Rosenthal, 2018). Optimization of subjective measures is critical for the study of consciousness; the measure one selects can make the difference in whether or not a priming effect is considered to be truly subliminal (Wierzchoń, Asanowicz, Paulewicz, & Cleeremans, 2012) or whether above-chance orientation discrimination sensitivity is considered a case of Type 1 or Type 2 blindsight (Rausch & Zehetleitner, 2016). In each of these cases, confidence judgments were found to be the most exhaustive and conservative measure of conscious awareness. While the current results do not rule out the possibility that decoded patterns for other subjective measures like visibility judgments might show a similar association with conscious perception in the same DecNef paradigm, they are informative nonetheless in providing at least a partial neural basis for the well-supported behavioral link between confidence judgments and consciousness.

These results also shed light on the current debate about whether or not prefrontal cortex is part of the core neural basis of consciousness (Boly et al., 2017; Odegaard et al., 2017). Given that

three out of the four confidence ROIs were in prefrontal cortex, the observed association between confidence induction and conscious perception suggests that prefrontal cortex is critically involved in consciousness. Indeed, comparing high confidence induction under false alarm vs correct rejection runs in each of the individual frontoparietal ROIs as well as in a collective prefrontal ROI (spanning IFS, MFS, and MFG) suggests that the main effect of high confidence induction under false alarms is driven by activity in prefrontal cortex (Figure A3). Also supporting the idea that PFC is uniquely involved in the generation of conscious percepts is the result that the decoded patterns for color in visual cortex were not associated with false color perception.

To further investigate whether the information in the ROI-specific decoders was shared with other ROIs during neurofeedback, we performed an information leak analysis as previously described (Cortese et al., 2016; Shibata, Watanabe, Sasaki, & Kawato, 2011); Figure A4) using data from the 13 DecNef subjects whose induction data were considered in the false alarm versus correct rejection analyses shown in Figure 11a and 11b. Briefly, this analysis quantifies the extent to which multivariate BOLD patterns in one ROI can predict the output of a decoder trained in another ROI. The pattern of results shown in S4 suggests minimal information leak between frontal and visual ROIs, with more intermediate information leak between parietal and frontal and parietal and visual ROIs. Given that the main result of high confidence induction likelihoods being higher during false alarm runs was found to be driven largely by frontal activity (Figure A3), the leak analysis further supports the idea that this effect was independent of activity in visual cortex. Furthermore, supporting the notion that the output of the confidence decoder was meaningfully related to perception is the fact that average high confidence

induction likelihoods in confidence ROIs across all three days of neurofeedback were positively correlated with average confidence judgments (Figure A5).

The finding that participants' red/green discrimination psychometric functions shifted in the direction of a higher overall proportion of red responses is consistent with a previous DecNef study targeting color representations in visual cortex (Amano et al., 2016). A potentially important difference, however, is that in the previous study it was reported that participants successfully learned to activate the targeted decoded color patterns more frequently over time through DecNef training, whereas in the current study such learning did not occur. One possible explanation for this difference is that despite the lack of such learning in the current experiment, an association still formed between the induction stimulus, which was matched in its achromatic parameters to the target stimuli in the Day 3 and Day 7 psychophysics tasks, and spontaneous induction of decoded patterns for redness.

An alternative explanation is that the observed shift in psychometric functions in the current study was due to an initial Day 3 bias towards choosing green being minimized over time via non-DecNef-related perceptual learning. Future studies should investigate this question by reducing such initial biases through longer training periods and more extensive stimulus titration based not only on task performance (as was the case here) but also on response bias measures like the signal detection theoretic criterion (Green DM, 1966; MacMillan & Creelman, 2004).

A limitation of the current study is the intermittent nature (i.e., run-by-run) of the psychophysical data collected during DecNef. Trial-by-trial measures of color perception would provide

considerably greater power for evaluating the relationship between confidence and conscious awareness. Run-by-run psychophysical measures are not without precedent (Cheesman & Merikle, 1984), and they were selected here as a means of reducing trial times. This was in turn intended to facilitate participants learning to induce the targeted multivoxel color and confidence patterns. However, given the lack of such a learning effect here, it may be optimal for future studies to prioritize the greater power and signal-to-noise ratio afforded by trial-by-trial perceptual judgments.

It remains an open question, however, what prevented participants from learning to activate the targeted color and confidence patterns in the DecNef task. As such an effect would allow for the investigation of a causal relationship between perceptual confidence and consciousness, this is an important issue for future DecNef studies to investigate. The present DecNef study was the first to target two categorically different perceptual targets (color and confidence); one possibility, therefore, is that the combination of these patterns becomes too complex for participants to learn to generate endogenously in a consistent manner. Another related possibility is that the neurofeedback procedure was simply spread across too many decoders (one for color and four for confidence). The complexity of the neurofeedback procedure, both in terms of categorical targets and number of decoders, may similarly have been responsible for the failure of confidence neurofeedback to modulate confidence judgements as previously reported (Cortese et al., 2016).

Future studies should investigate the limits of what human participants can learn to regulate via DecNef both in terms of the number and distribution of decoders throughout the brain, and in terms of the complexity of the perceptual content being targeted. For example, it is an open

question whether confidence DecNef would benefit from using a single decoder spanning the four frontoparietal ROIs used here, as might be suggested by successful approaches using whole brain decoders (deBettencourt, Cohen, Lee, Norman, & Turk-Browne, 2015). It would also be informative to repeat the current study with confidence DecNef alone, and to limit the neurofeedback to only prefrontal ROIs (Figure A3). If learning occurred in either of these modified contexts, or some combination of them, then it would suggest that the difference in learning indeed stems from the difference in either the number of decoders or the categorical complexity of the targets of neurofeedback.

Another limitation and potential roadblock to induction learning is the suboptimality of the decoding procedure itself. It is possible that simply increasing decoding accuracy in the first two days of the study would have lead to induction learning. One recent study showed that decoding accuracies could be significantly improved by integrating fMRI hyperalignment (Haxby et al., 2011) into the DecNef procedure (Taschereau-Dumouchel et al., 2018). Hyperalignment improves decoding accuracies by taking advantage of shared high-dimensional patterns in representational content between subjects. Another recent study showed that offline simulations can be used to estimate optimal parameters for experiment timing and real-time fMRI preprocessing, which can lead to greater decoding accuracy and neurofeedback performance (Oblak, Sulzer, & Lewis-Peacock, 2018). Similar approaches should be applied going forward in order to optimize the efficacy of DecNef tasks, which should in turn allow us to ask questions about a potential causal relationship between confidence and consciousness.

In conclusion, the present study found an association between the occurrence of decoded patterns for high perceptual confidence in frontoparietal cortex and subjective color perception.

Furthermore, pre- and post-DecNef psychophysics revealed a shift in participants' psychometric functions for red/green color discrimination that was consistent with a previous color-DecNef study (Amano et al., 2016). The results discussed above support both the efficacy of confidence judgments as a subjective measure for consciousness and the notion that conscious perception may rely critically on activity in prefrontal cortex.

IX. Phenomenological Richness, Subjective Inflation, and the Continued Search for Blindsight in Normal Observers

Background: Subjective Inflation, Phenomenology's Get-Rich-Quick Scheme

Introduction

There is a longstanding and currently lively debate about whether or not visual phenomenology overflows cognitive access (see recent themed issue of Philosophical Transactions of the Royal Society B (Fazekas & Overgaard, 2018; Matthews, Schröder, Kaunitz, van Boxtel, & Tsuchiya, 2018; Naccache, 2018; Odegaard, Chang, et al., 2018; Overgaard, 2018; Ian Phillips, 2018; Sergent, 2018; Stazicker, 2018; Usher, Bronfman, Talmor, Jacobson, & Eitam, 2018; Ward, 2018)). The question of phenomenological overflow is often rephrased as asking whether phenomenology is rich or sparse. On the Rich view, a snapshot of visual phenomenology is highly detailed across the visual field, while cognitive access is constrained to a subset of that detail given the limited capacities of attention, memory, and reporting mechanisms (Block, 1995, 2007, 2014; Bronfman, Brezis, Jacobson, & Usher, 2014; Koch & Tsuchiya, 2007; Lamme, 2003, 2010; Sligte, Vandenbroucke, Scholte, & Lamme, 2010; Sperling, 1960; Tsuchiya et al., 2015; Usher et al., 2018). This view is supported by anecdotal reports from subjects in partial-report studies (Landman, Spekreijse, & Lamme, 2003; Sligte et al., 2010; Sperling, 1960)

in which they indicate having seen more of a given stimulus array than their objective task performance would suggest. The Rich view further posits, based on both partial-report and dual-task studies (Braun & Julesz, 1998; Matthews et al., 2018; Sperling & Dosher, 1986), that attention is not necessary for phenomenology.

Proponents of the Sparse view argue, often relying on the results of inattentional blindness (Mack & Rock, 1998; Neisser & Becklen, 1975; Simons, 2000) and change blindness (Simons & Rensink, 2005) studies, that phenomenology itself is constrained by, and thus scales with, cognitive access. For example, when subjects fail to notice a difference between two consecutively presented images at a minimally attended location, a Sparse interpretation is that there was insufficient phenomenological detail at that location for the change to be noticed. On this view, a snapshot of visual phenomenology is highly detailed around the focal point of attention where there is strong cognitive access, but loses detail as attention drops off in the periphery, where there is minimal cognitive access (Cohen & Dennett, 2011; Stanislas Dehaene, Changeux, Naccache, & Sergent, 2006; Kouider & Gardelle, 2010; Naccache, 2018; Sergent, 2018; Sergent et al., 2013; Ward, 2018; Ward, Bear, & Scholl, 2016). Furthermore, on this view, phenomenology and cognitive access tend to scale with attention, though this relationship is not necessarily monotonic. Importantly, proponents of this view argue that, without consensus on an operational definition of attention, we cannot definitively claim that attention is not necessary for phenomenology (Cohen, Cavanagh, Chun, & Nakayama, 2012; Cohen & Dennett, 2011; Kouider & Gardelle, 2010).

It has been suggested that the debate about phenomenological overflow, and by extension, the Rich vs Sparse debate, may be empirically intractable (Cohen & Dennett, 2011; Stanislas

Dehaene et al., 2006; Kouider & Gardelle, 2010; Overgaard & Fazekas, 2016; I. Phillips, 2011; Ian Phillips, 2018; Sergent et al., 2013; Stazicker, 2018; Ward et al., 2016). The major issue usually raised is that while empirical studies of phenomenology fundamentally rely on subjective reports, unaccessed phenomenology is, by definition, unreportable. And if unaccessed phenomenology cannot be detected, its presence neither confirmed nor denied, then the results of any given study can be interpreted as being consistent with either the Rich or the Sparse view (Cohen & Dennett, 2011; Kouider & Gardelle, 2010; Overgaard, 2018; I. Phillips, 2011; Ian Phillips, 2018; Stazicker, 2018; Ward, 2018).

A classic experiment that combines the retrocuing paradigm made famous by George Sperling (Sperling, 1960) with a change blindness paradigm (Landman et al., 2003) provides an example where such alternative interpretations are available. Participants were shown two consecutive arrays of eight rectangles, where each rectangle could be oriented either vertically or horizontally. The subjects' task was to detect whether one of the rectangles switched orientations between the first and second arrays. Because subjects were instructed to initially fixate at a central point that is roughly equidistant from the locations of each of the rectangles in the to-be-flashed arrays, attention was thought to be diffuse and minimal with respect to each individual object in the first array. The authors found that when a retrocue indicating the location of the object whose orientation could potentially change was presented after offset of the first array, but prior to onset of the second array, performance on the change detection task was better than it was when no retrocue was presented.

On a Rich interpretation of this result, the orientations of the minimally-attended rectangles in the first array are present in phenomenology, and the postcue allows participants to access and

remember the orientation of the cued rectangle before that information can be overwritten by the second array. In this case, a failure in change detection amounts to a failure of memory. On a Sparse interpretation, orientation information for each of the minimally-attended rectangles is represented unconsciously, with the resulting phenomenology being filled-in and/or summarized. Retrocued attention then makes the initially unconscious orientation information for the cued object available for report. On this interpretation, a failure in change detection amounts to both a failure of phenomenological representation and a failure to access the relevant unconscious orientation information before onset of the second array. Because subjects' behavior could be identical under both views, neither interpretation is obviously more tenable than the other.

A similar problem arises when trying to address more directly whether attention is necessary for consciousness. A common way to test this is to try to demonstrate that some task can be performed in the absence or near absence of attention, for example using a dual-task paradigm (Braun & Julesz, 1998; F. F. Li, VanRullen, Koch, & Perona, 2002; Matthews et al., 2018; Sperling & Dosher, 1986; Van Boxtel, Tsuchiya, & Koch, 2010). But again, without a working operational definition of attention, it is unclear how the complete elimination of attention could be unequivocally demonstrated in an experimental setting. Some have argued that this similarly makes the debate about the necessity of attention empirically intractable (Cohen et al., 2012; Cohen & Dennett, 2011; Kouider & Gardelle, 2010).

In light of these concerns, Ned Block has suggested that currently the best approach to the Rich vs Sparse debate is to consider the extant empirical evidence and use inference to the best explanation (Block, 2007; Harman, 1965). Here, we agree with Block's methodological appeal. We highlight an approach that, instead of attempting to investigate visual phenomenology in the

complete or near-complete absence of attention, exploits the fact that attention is graded, and investigates the interaction between attention and phenomenology. Such an approach consistently reveals a phenomenon known as subjective inflation, wherein study participants exhibit liberal detection criteria or are overly confident when evaluating minimally attended or peripheral stimuli (Odegaard, Chang, et al., 2018). We argue, in line with (Odegaard, Chang, et al., 2018), that this can explain the subjective impression of richness across the visual field appealed to by Block (Block, 2007), while accounting for both behavioral (Mack & Rock, 1998; Neisser & Becklen, 1975; Simons, 2000; Simons & Rensink, 2005) and physiological (Azzopardi & Cowey, 1993; Strasburger, Rentschler, & Jüttner, 2011) limitations in minimally attended perception, all without invoking phenomenological overflow.

Evidence for Subjective Inflation

It is well established that attention boosts objective perceptual performance (Carrasco, 2011). It has also been shown that subjective and objective measures of perception can be dissociated in clinical cases of blindsight (Weiskrantz, 1986, 1999), as well as in normal observers (Lau & Passingham, 2006). It follows that if we want to isolate a subjective measure of phenomenology for comparison between conditions of low and high attention (or low and high retinotopic eccentricity), then objective task performance should be treated as a potential confound and matched between those conditions (Lau, 2008).

Rahnev et al. (Rahnev et al., 2011) employed the strategy described above by matching objective performance, as indexed by the signal detection theory (SDT) sensitivity measure, d' (Green DM, 1966), between attended and minimally attended conditions in a visual detection task (Figure 12). They found that the SDT criterion measure, c, was systematically more liberal

in the minimally attended condition (Figure 12b). On a detection task, a lower value of c equates to a higher proportion of trials in which the observer incorrectly reports the presence of the target stimulus, i.e., a higher false alarm rate (Rahnev et al., 2011), Supplementary Figure 3]. This difference in false alarm rate between attended and minimally attended conditions is of particular note given that stimulus strength is matched (at zero) across all false alarm trials.

One could argue that the observed shift in criterion reflects a change in non-perceptual decisional bias (Witt, Taylor, Sugovic, & Wixted, 2015). However, we interpret this shift to reflect a genuine change in subjective perception, as it was robust to trial-by-trial feedback and increases in monetary rewards, manipulations that are thought to minimize non-perceptual response biases (Rahnev et al., 2011; Solovey et al., 2015; Witt et al., 2015). On this interpretation, the observed criterion shifts in (Rahnev et al., 2011) suggest that subjective awareness is systematically inflated relative to objective processing capacity in conditions of minimal attention. It is worth emphasizing that this effect may seem, at least initially, to be counterintuitive: despite performance matching, one might reasonably expect that when attention is reduced observers might be more conservative in their detection judgments. This could be due to prior knowledge that, under normal conditions, unattended vision is relatively poor. However, we observe the opposite: given matched objective capacity, when attention is reduced, participants are more likely to say something is there when it is not.

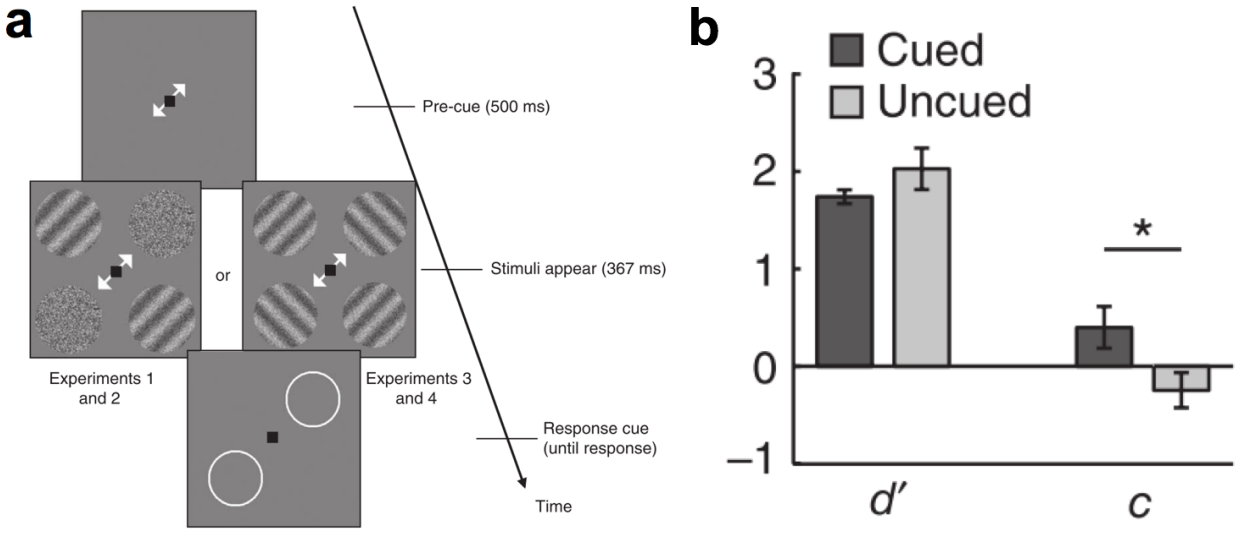

Figure 12. An empirical demonstration of subjective inflation from (Rahnev et al., 2011). **a)** Stimuli and task. Each trial was initiated by a pre-cue that indicated to which diagonal (top-right and bottom-left or top-left and bottom-right) the subject should attend. This was followed by a stimulus array, in which each diagonal (independently) contained either a set of tilted gratings embedded in noise or noise alone. This was followed by a response cue, which prompted the subject to indicate, in a forced-choice manner, whether or not the diagonal in the stimulus array indicated by the response cue had contained a set of gratings. The pre-cue and response cues were congruent on 70% of trials. **b)** Behavioral results show matched detection sensitivity, d', between cued and uncued conditions, but a more liberal criterion measure, c, in the uncued condition relative to the cued condition. Reprinted with modifications and permission from (Rahnev et al., 2011).

This inflation effect has been replicated and extended in several independent studies. One study found a consistent decrease in detection criterion in peripheral vision compared to central vision despite matched detection d' (Solovey et al., 2015). This effect was again robust to trial-by-trial feedback. Similarly, in a study that was not designed to test inflation, more liberal change detection criteria were observed for fragile visual short term memory compared to working memory, despite matched change detection d' between the two conditions (Vandenbroucke et al., 2014). The authors interpreted this as being analogous to the inflation effect in (Rahnev et al., 2011), as there was presumably higher attention in the working memory condition. Another study, which used a relatively naturalistic setting of simulated driving, found more liberal criteria for detecting the color of a simulated pedestrian's shirt when attention was minimized (M. K. Li,

Lau, & Odegaard, 2018). In a control experiment in which the identity of the detection target (the color to be detected out of a set of 11 colors) was not revealed until after presentation of the target stimulus, the inflation effect was no longer present. This suggests that the observed criterion shift in (M. K. Li et al., 2018) was not due to simple confirmation bias, which, much like the trial-by-trial feedback results in (Rahnev et al., 2011) and (Solovey et al., 2015), indicates that the observed inflation effect was perceptual as opposed to decisional.

Inflation was also recently observed in both peripheral summary statistics and crowding tasks (Odegaard, Chang, et al., 2018). In the latter task, inflation was indexed by an increase in confidence ratings on incorrect trials for crowded compared to uncrowded stimuli. Surprisingly, this effect was observed despite discrimination d' being lower in the crowded condition. Also recently, an inflation-like effect, similar to the flashed face distortion effect (Tangen, Murphy, & Thompson, 2011), was found in which repeatedly flashed peripheral color stimuli were rated as more saturated than physically-saturation-matched, non-flashing central color stimuli (Sivananda, Peters, Lau, & Odegaard, 2017). Further, an inflation-like effect has been found in which subjects failed to notice drastic changes in peripheral text during reading (McConkie & Rayner, 1975).

Inflation and the Richness Debate

To see how inflation fits into the Rich vs Sparse debate, it is helpful to consider that phenomenology tends to feel rich. For example, in the case of the famous Sperling retrocueing experiments (Sperling, 1960), participants indicated, anecdotally (Ward, 2018), that they felt as though they saw all of the characters in a briefly flashed 3x4 array in detail, despite not being able to report all of that detail after the fact. Proponents of the Rich view traditionally take these

reports at face value and assume that this detail was indeed phenomenologically represented outside of focal attention, albeit in a fragile and fleeting manner. It is worth emphasizing, however, that in many partial-report studies these reports of global richness are not systematically collected per experimental procedure, and instead come from retrospective subjective impressions (Landman et al., 2003; Sligte et al., 2010; Sperling, 1960; Ward, 2018). And it is possible that such retrospective reports result from a demand effect (Orne, 1969). However, for the sake of argument, we take these reports at face value here.

The Sparse view, on the other hand, suggests that what is phenomenologically represented outside of focal attention is gist-like, a summary representation of low level features (Cohen, Dennett, & Kanwisher, 2016; Kouider & Gardelle, 2010; Ward, 2018). There is evidence for this view from studies in which nonsensical or dramatically homogenized alphanumeric characters in peripheral vision go unnoticed (de Gardelle, Sackur, & Kouider, 2009; McConkie & Rayner, 1975), and studies in which participants cannot subjectively distinguish between physically distinct images that are matched for low-level image statistics (Freeman & Simoncelli, 2011; Wallis, Funke, Alexander, Gatys, & Wichmann, 2018). Yet, a proponent of the Rich view might argue that, despite these studies, non-focally attended phenomenology does not feel summarized or homogeneous, so there is still something left to be explained. In other words, if the Sparse view is correct, why is introspection so mistaken?

We suggest, in line with (Odegaard, Chang, et al., 2018), that this mistaken feeling of richness is the result of subjective inflation. If partially-summarized, minimally-attended representations are subjectively inflated above what would be expected based on objective processing capacity, as in the examples above, then inflation can explain the apparent richness of the resulting

phenomenology despite well-established physiological limitations in minimally-attended and peripheral visual processing (Azzopardi & Cowey, 1993; Carrasco, 2011; Strasburger et al., 2011).

The exact mechanism of inflation here may be unclear (but see (Odegaard, Chang, et al., 2018; Rahnev et al., 2011; Solovey et al., 2015)). Still, the very occurrence of inflation provides an explanation based on an operationally defined comparison between attended and minimally attended phenomenology. And that explanation is neutral as regards the more difficult and possibly empirically intractable issues surrounding the overflow argument. Therefore, a combination of summary statistics and subjective inflation arguably provides the best explanation of the subjective phenomena that advocates of the Rich view appeal to. Since this view incorporates both the relevant behavioral data and anecdotal reports of richness, which may well be illusory (Kouider & Gardelle, 2010), we suggest that it constitutes a position intermediate between the Sparse and Rich views.

This inflation account is consistent with behavioral results observed in several additional studies of peripheral or minimally attended vision. For example, inflation appears to be similar to perceptual filling-in (Komatsu, 2006; Odegaard, Chang, et al., 2018). A recent study found that filled-in percepts at the blindspot are rated as more reliable than perceptually equivalent, but externally veridical percepts (Ehinger, Häusser, Ossandon, & König, 2017). This shows a similar pattern to inflation in which a percept that is based on a less veridical representation of the external world is actually granted a subjective boost. This leads to the question of whether filling-in at the blind-spot and inflation share a common mechanism. If so, then we might also expect the mechanism that underlies known cases of peripheral inflation (M. K. Li et al., 2018;

Sivananda et al., 2017; Solovey et al., 2015) to underlie instances of peripheral filling-in like in the “uniformity illusion” (Otten, Pinto, Paffen, Seth, & Kanai, 2016) or in the case of peripheral color in natural scenes (Balas & Sinha, 2007). Presumably, such questions are empirically tractable. For example, future studies could combine an approach similar to (Ehinger et al., 2017) with the performance-matching strategies described above, e.g., (Rahnev et al., 2011; Solovey et al., 2015), to see if subjective judgements about filled-in percepts are, operationally speaking, inflated relative to judgments about veridical percepts.

Additionally, as the peripheral filling-in examples above are presumably driven by an expectation based on foveal perception, inflation similarly appears to be influenced by expectations based on prior knowledge. For example, the control experiment in (F. F. Li et al., 2002), described above, suggests that the content of inflation depended, at least in part, on an expectation about a specific color. This is consistent with the studies mentioned above in which participants failed to notice manipulations to alphanumeric characters during partial-report (de Gardelle et al., 2009) and reading (McConkie & Rayner, 1975) tasks. It is also consistent with the result that stronger filling-in effects were observed when the to-be-filled-in content of a natural scene percept was colored compared to when it was grayscale, presumably based on the prior knowledge that natural scenes contain color (Balas & Sinha, 2007). Similar prior expectation effects based on central visual information have been shown for peripheral motion perception (Zhang, Kwon, & Tadin, 2013) and peripheral object recognition (Wijntjes & Rosenholtz, 2018).

This reliance on expectation supports the important partial awareness hypothesis put forth by Kouider et al. (Kouider & Dupoux, 2004; Kouider & Gardelle, 2010), and suggests that inflation may underlie cases in which expectation influences minimally attended phenomenology (Balas

& Sinha, 2007; de Gardelle et al., 2009; M. K. Li et al., 2018; McConkie & Rayner, 1975). One benefit is that expectation is an easily manipulated experimental variable; thus, the influence of expectation on minimally attended phenomenology should also be amenable to empirical investigation going forward. For example, one prediction is that inflation should be stronger when expectations are higher. Here, we take expectations to be functionally different from attention in that they depend on the predictability of the stimulus. Experimentally, attention and expectation could be manipulated separately, e.g., by telling a subject to attend to location A, where the statistical likelihood of a detection stimulus appearing is known to be low relative to location B (Kok, Rahnev, Jehee, Lau, & De Lange, 2012). If inflation depends critically on expectation, then we should expect inflation effects to be stronger when the subject attends to location A compared to when they attend to location B.

We therefore propose that the content of inflated phenomenology can be affected by at least 3 factors: 1) summary computations of minimally-attended and/or peripheral low-level visual features (Cohen et al., 2016; Freeman & Simoncelli, 2011; Wallis et al., 2018; Ward et al., 2016), 2) expectations based on attended and/or foveal representations (Balas & Sinha, 2007; Otten et al., 2016; Suárez-Pinilla, Seth, & Roseboom, 2018; Toscani, Gegenfurtner, & Valsecchi, 2017; Wijntjes & Rosenholtz, 2018; Zhang et al., 2013), and 3) expectations based on prior knowledge (Balas & Sinha, 2007; de Gardelle et al., 2009; McConkie & Rayner, 1975; Wijntjes & Rosenholtz, 2018). In the case of prior knowledge, it is a topic for future investigation whether this knowledge needs to be explicit, or if it can be implicit (e.g., as in implicit statistical knowledge found in visual search (Zinchenko, Conci, Müller, & Geyer, 2018)).

Summary

We have argued that subjective inflation may explain the apparent introspective content that the Rich view assigns to minimally attended and peripheral phenomenology. It can explain why phenomenology may appear to overflow cognitive access, without countenancing actual overflow. By denying overflow but explaining why it seems to occur, inflation provides an intermediate position between the Rich and Sparse views that borrows and extends ideas from the partial awareness (Kouider & Dupoux, 2004; Kouider & Gardelle, 2010) and summary statistics hypotheses (Cohen et al., 2016).

On this position, minimally attended phenomenology is built on summary representations (Cohen et al., 2016), the subjective reliability of which is inflated above what would be predicted based on objective performance. This operationally definable effect is likely based on prior expectations (Balas & Sinha, 2007; de Gardelle et al., 2009; Kouider & Dupoux, 2004; Kouider & Gardelle, 2010; M. K. Li et al., 2018; McConkie & Rayner, 1975). So phenomenological richness need not be seen as mapping onto high representational capacity, as the traditional Rich view claims to be the case in early visual cortex (Block, 1995, 2007; Lamme, 2003, 2010). Rather, inflation supports strong physiological evidence that peripheral representational capacity in early vision is limited (Azzopardi & Cowey, 1993; Strasburger et al., 2011), and suggests that our exaggerated sense of rich peripheral phenomenology is mediated by some later stage prediction-based mechanism (Suárez-Pinilla et al., 2018). The neural basis of this mechanism is a topic for future investigation, but at this point, it seems to be at odds with the multi-level overflow account offered by proponents of the Rich view (Block, 1995). One prediction along these lines is that when comparing false alarm trials to correct rejection trials for minimally attended stimuli in a visual detection task, the decodability of illusory, inflated perception should

be high in later stages of the visual processing hierarchy, but not in primary visual cortex (Suárez-Pinilla et al., 2018). Furthermore, if expectation-based inflation effects are mediated by top-down feedback to visual areas (Miconi & VanRullen, 2016), then this should be reflected in the relative timecourses of decodability between frontoparietal areas and visual areas. This prediction could be tested with a time-sensitive imaging method like magnetoencephalography (MEG).

Finally, one interesting consideration is that whether the inflation hypothesis really favors one side of the richness debate over the other may ultimately depend on the extent to which proponents of the Rich view categorize illusory percepts as instances of “rich” phenomenology. Indeed, Block (Block, 2011) and others (Bronfman et al., 2014), p. 1395] have conceded this point, e.g., acknowledging that “a minor illusion effect,” (Block, 2011), p.5], as in (de Gardelle et al., 2009), can still be consistent with the Rich view. Whether the majority of the results discussed here would constitute such minor illusions is a topic for future discussion; though it is worth pointing out that the more the Rich view sees such illusion effects as constituting “rich” phenomenology, the more the dividing lines between the Sparse and Rich views will blur. Most importantly, we hope that the issues raised here in connection with inflation will stimulate new ideas about how to approach the puzzle of phenomenology, ultimately giving the field more useful data for resolving, or perhaps dissolving, the richness debate.

Impaired Introspective Access in Dot Motion Discrimination

Abstract

Previous studies have shown that subjective judgments about peripheral or minimally attended visual perception can be inflated above what would be expected based on objective performance (Knotts, Odegaard, Lau, & Rosenthal, 2018). This effect has been observed for grating detection and discrimination, color detection, orientation summary perception, and crowding, but it is not known if it generalizes to dot motion discrimination. Surprisingly, when we compared central and peripheral dot motion discrimination directly in a two-interval forced choice task with relative confidence judgments, we observed the opposite of the predicted inflation effect: subjective confidence ratings were higher for central stimuli when objective performance was matched between center and periphery. We followed this with a series of behavioral experiments that revealed another surprising effect: metacognition is impaired for dot motion descrimination in both central and peripheral vision, more so than when an analogous task is used with orientation discrimination. Whether these results can be interpreted as evidence for blindsight in normal observers versus an inflation-like effect that may be inherent to the motion system is discussed.

Introduction

In Experiments1.1-1.4 we found no evidence for either relative or absolute blindsight in normal observers performing an orientation discrimination task under a variety of monocular and binocular suppression techniques. This supports a previous finding for a lack of absolute blindsight in normal observers for orientation discrimination under forward and backward masking (Peters & Lau, 2015). Importantly, all of these experiments used centrally presented stimuli. Based on the subjective inflation effects described in the previous section, we

hypothesized that relative blindsight could be found by using the comparative 2IFC approach from Experiments 1.1-1.3 to compare central and peripheral stimuli. Specifically, per inflation, we hypothesized that when objective performance on a given perceptual task is matched between central and peripheral conditions, participants will indicate higher confidence in peripheral stimuli.

In addition to introducing an eccentricity manipulation in our continued search for relative blindsight, we switched to using a dot motion discrimination task. There are a few reasons why dot motion may be more conducive to inflation-based relative blindsight. First, in the two-streams hypothesis of vision, motion perception has been associated with the “less conscious” dorsal stream (Goodale, 2011; Milner & Goodale, 2008; Tapia & Breitmeyer, 2011). Second, motion discrimination for random dot motion kinematograms inherently involves a summary computation. This can be intuited from imagining a natural analog to a random dot motion kinetogram, like the movement of a school of fish. One does not need to possess simultaneous awareness of the movement of every individual component to perceive the global trajectory of the group (but see Bronfman et al., 2014; Block, 2014; Haun et al., 2018 for a different perspective). As we listed summary computations as one of the putative major contributors to subjective inflation effects in the previous section, random dot motion perception appears to be an ideal candidate task for generating an inflation-like effect.

In the current series of experiments we compared subjective awareness in central and peripheral dot motion discrimination first using a simultaneous central/peripheral dot motion task. To anticipate, in the simultaneous central/peripheral task we found the opposite of the hypothesized inflation effect: participants were more confident in central motion judgments. This

led to three additional experiments, testing both relative and absolute blindsight for central and peripheral dot motion discrimination, in which it was found that metacognition for dot motion discrimination appears to be impaired across the visual field.

Experiment 3.1: Simultaneous center and peripheral dot motion discrimination

Methods

Participants

Four participants (3 female, ages 18-33, all right handed), including the first author, gave written informed consent to participate. One subject was removed from analyses due to failure to perform the left/right motion discrimination task above chance. All participants had normal or corrected-to-normal eyesight and normal stereo vision, and all were either paid $13.25 USD per hour plus performance-contingent bonus money (see Procedure section below) for their participation. This and all subsequent experiments in this chapter were conducted in accordance with the Declaration of Helsinki and were approved by the UCLA Institutional Review Board.

Apparatus & Stimuli

All stimuli were generated with custom Matlab R2014a (Natuck, MA) scripts using PsychToolbox 3.0.12 on a gamma-corrected Dell E773c CRT monitor with a resolution of 1024 x 768 pixels and a refresh rate of 75 Hz. Participants used a chin rest and viewed stimuli at a fixed distance of 42 cm in a dimly lit room. Eye movements were tracked continuously with an SMI REDn Scientific remote eye tracker (SensoMotoric Instruments GmbH, Teltow, Germany) at a sample rate of 60 Hz.

All dot motion kinematograms contained 196 dots and subtended a circular region with a diameter of 3.5° for an average density of 16 dots per degree squared. Individual dots were square with a width of 2 pixels or ~0.08° and had a speed of 1°/s. The background was a uniform gray with an rgb triplet value of [109, 109, 109] and a luminance of ~16.1 cd/m^2. Dots were white with a luminance of ~97.6 cd/m^2, resulting in a Michelson contrast of ~0.72. Peripheral dot motion stimuli were centered at points 19.5° northeast, northwest, southeast, or southwest from central fixation (dashed circles in Figure 13a). Central and peripheral stimuli were presented simultaneously for 1200 ms, and all dots had a full lifetime over this interval.

The starting locations of all dots were selected randomly for each motion stimulus. Each stimulus contained a subset of coherently moving dots whose directions of motion were either leftward or rightward with no vertical component. The motion directions of the remaining "incoherent" dots were determined as follows. Assuming an even number of incoherent dots, the motion directions (in degrees) of half of these dots ($D_1...D_{N/2}$) were selected randomly from the 360 whole numbers from 1 to 360. The directions of the remaining incoherent dots ($D_{N/2+1}...D_N$) were determined by adding 180° to the randomly selected directions from the first half. If there was an odd number of incoherent dots for a given stimulus, then the value of N/2 in the computations just described was rounded down, and the final remaining incoherent dot was randomly selected to move either upward (90°) or downward (270°) with no horizontal motion component. Dots were plotted within a square region with a width equivalent to the diameter of the circular motion stimulus (3.5°), but only dots falling within the largest circle within the square were displayed. Dots exiting the square plotting boundary were replotted at a random location on the opposite side.

Eye tracking

At the start of each session participants performed a 10-point eye tracking calibration procedure using a default SMI REDn Scientific Matlab script (SensoMotoric Instruments GmbH, Teltow, Germany). Participants then performed an additional 1-back calibration task during which they fixated on a centrally presented alphanumeric character surrounded by 4 concentric rings of additional characters with 2^{n+1} characters in every nth concentric ring. All characters were presented in size 11 arial font and changed at a rate of 1.4 Hz. Each nth concentric ring had a diameter of n x 1.3°. Participants were instructed to fixate the centrally presented character and press a button every time the same character was presented twice consecutively, referred to hereafter as a target pair.

Participants first performed a practice run of the 1-back task with 11 character changes and two target pairs. If they did not detect both target pairs they had to repeat the practice. They then performed a main run with 39 letter changes and eight target pairs. If they did not detect at least seven of the eight target pairs they had to perform another run. Participants were instructed to keep their gaze fixed firmly on the central character throughout the task for the purpose of calibrating the eye tracker. Gaze position was recorded continuously throughout the main 1-back task, and the mean position across all eyetracking samples was set as the functional central fixation point for all subsequent real-time eye tracking analyses. Additionally, a functional "gaze contingency radius" was set by doubling the standard deviation of gaze positions and adding 1.75°. If gaze position deviated from the functional central fixation point by a value greater than this gaze contingency radius during specific gaze-contingent portions of subsequent tasks, the prevailing trial was aborted and a warning message was displayed. If the standard deviation of eye movements on the main 1-back task exceeded 9.75°, participants

were informed that the variance of their gaze position was too high, and the 8-target version of the task was repeated. These calibration and gaze-enforcement procedures were implemented in the same manner in all of Experiments 3.1-3.4.

Procedure

After completing the eye tracking calibration procedure participants performed 10 practice trials of the main task (Figure 1a). Participants first had to focus on a centrally presented fixation cross for a full 1000 ms to initiate the rest of the trial sequence. If, on any trial throughout the experiment, it took longer than 10 s to initiate the trial sequence via fixation, the 1-back task was repeated and a new functional central fixation point and gaze contingency radius were determined. Upon successful initiation of a trial the display cross remained on the screen for a variable interval between 100 and 250 ms, after which the central and peripheral dot motion kinematograms were displayed for 1200 ms. If gaze deviated beyond the gaze contingency radius described above during the variable fixation or stimulus display period the trial was aborted and a warning message was displayed. The location of the peripheral dot motion stimulus varied randomly across trials. After offset of the dot motion stimuli participants first made a Type 2 judgment, indicating via button press whether they were more confident in their ability to discriminate the direction of motion in the central or peripheral stimulus. They then made Type 1 left versus right motion discrimination judgments for each dot motion stimulus. The Type 2 judgment preceded the Type 1 judgments to prevent participants from using their Type 1 reaction times to inform their Type 2 judgments.

For all parts of the experiment the order of the Type 1 questions (central first versus peripheral first) was varied between blocks, with block order pseudorandomized across subjects. During

the practice session, five trials were performed per Type 1 question order to familiarize subjects with the blocking procedure. Trial-by-trial feedback on performance of both Type 1 (correct vs incorrect) and Type 2 (indicated higher confidence in a correct vs incorrect Type 1 decision) was also provided during the practice session.

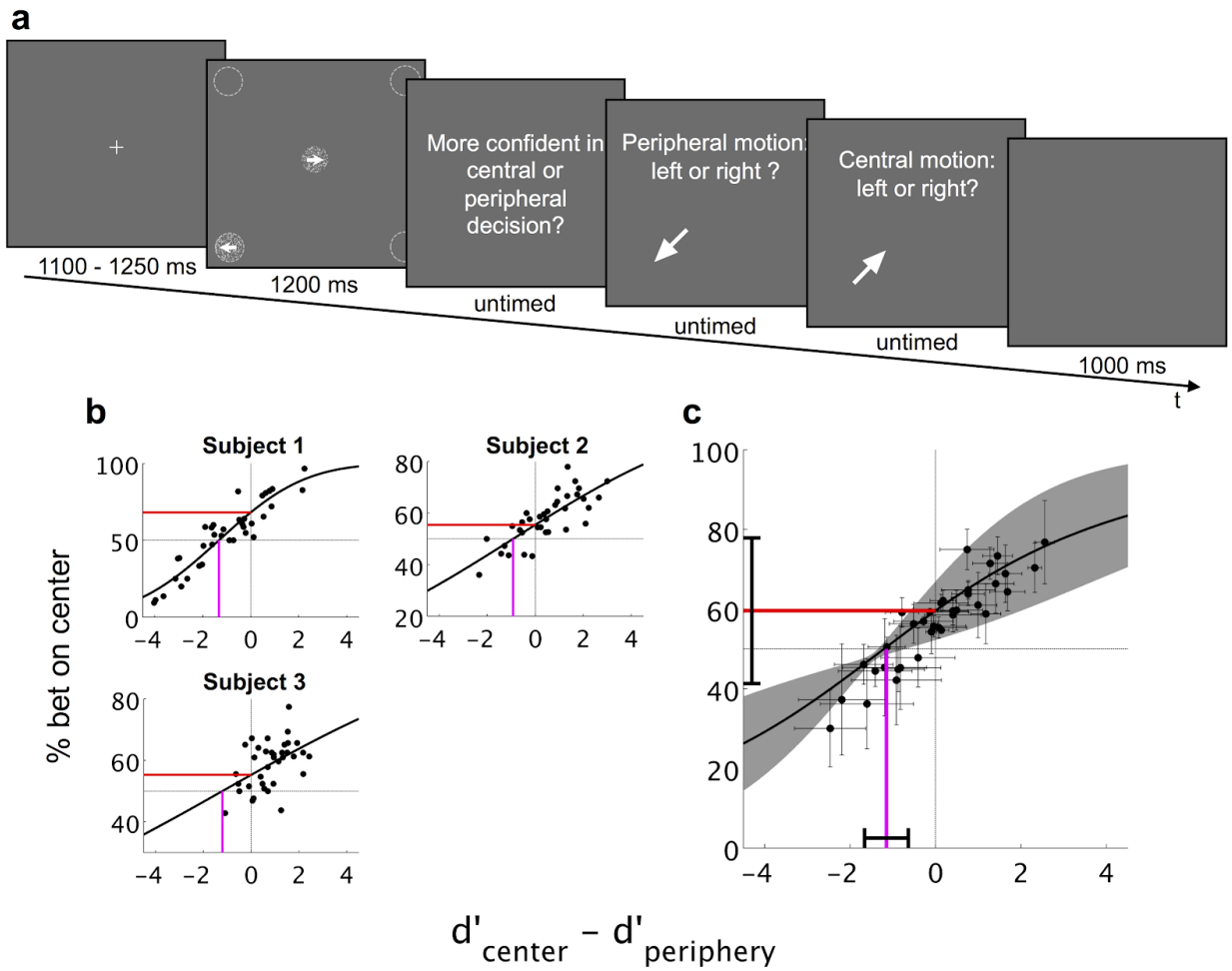

Figure 13. Procedure and results for simultaneous central and peripheral dot motion discrimination task (Experiment 3.1). **a.** On each trial, after a brief gaze-contingent fixation period, participants viewed one central and one peripheral random dot motion kinematogram each with a nonzero net motion coherence either to the left or to the right. The peripheral stimulus could appear at one of four locations, indicated by the dashed circles. Participants then reported the location, central or peripheral, at which they were more confident in their ability to accurately discriminate the net direction of dot motion. They then judged the net direction of motion at each location, with the order of question (central first or peripheral first) varied between blocks. This was followed by a 1000 ms intertrial interval. Font sizes are enlarged here relative to their actual size during the experiment for clarity. **b.** Individual psychometric curves.

For each participant the percentage of trials in which higher confidence was indicated for the central motion judgment is plotted as a function of the difference in motion discrimination d' between central and peripheral locations. Cumulative normal functions with mean and slope as free parameters were fit to the individual participant data. Also shown are the point of subjective equality (PSE; magenta) and point of objective equality (POE, red). **c.** Group level (N=4) data. Black circles represent group mean d' difference scores ± 1 s.e.m., while the black line shows the mean of the individual cumulative normal fits ± 1 SD (gray). The 95% confidence intervals for estimated PSE and POE group means are shown by the black bars near the x- and y-axes, respectively.

Following the practice participants performed 160 trials of an adaptive staircasing procedure (QUEST, Watson & Pelli, 1983) to estimate the percentage of dot motion coherence that would lead to 75% correct Type 1 performance both at center and in the periphery. The trial structure was identical to the practice trials and main task except that confidence judgments were not collected. Four randomly interleaved 40-trial staircases were used per eccentricity condition, and 75% coherence thresholds were estimated by taking the mean of the four coherence threshold estimates for each eccentricity condition. We then multiplied these estimated coherence thresholds by a set of six proportions to generate six coherence values for each eccentricity condition that would provide a range of Type 1 percent correct scores from roughly 60% to 90% (or approximately d' = 0.51 to d' = 2.56) correct on the subsequent main task. Mean ± s.e.m. proportions used across subjects were as follows: $\text{Proportions}_{\text{center}}$ = 0.17 ± 0.05, 0.34 ± 0.10, 0.51 ± 0.16, 0.71 ± 0.20, 0.95 ± 0.23, 1.27 ± 0.21 ; $\text{Proportions}_{\text{periphery}}$ = 0.18 ± 0.04, 0.30 ± 0.05, 0.42 ± 0.07, 0.59 ± 0.11, 0.77 ± 0.17, 1.03 ± 0.22. If a trial was aborted during the staircasing procedure due to improper gaze position, the trial was repeated with motion directions and the location of the peripheral stimulus newly randomized to avoid stimulus predictability.

Following the adaptive staircasing procedure participants performed 2,304 trials of the main task (Figure 13a) using a full factorial combination of 6 coherence levels per eccentricity condition, 4

peripheral stimulus locations, 2 motion directions per eccentricity condition, and 2 Type 1 question orders, with 2 trials per unique combination of conditions. Trials were divided into 48 48-trial blocks. If 5 gaze errors were made within a single block, subjects immediately repeated the 1-back eyetracking calibration task to generate a new functional central fixation point and gaze contingency radius. If 10 gaze errors were made within a single block, participants immediately repeated both the initial SMI REDn Scientific 10-point calibration task and the 1-back calibration task. After any recalibration procedures participants returned the trial of the main task on which they had left off.

Given the large trial number, the main task was completed over several sessions on different days (mean ± s.d. days per subject: 7 ± 1). At the end of each session a total performance score was computed by adding the total proportion of correct Type 1 judgment to the proportion of Type 2 judgments in which participants indicated higher confidence in a correct Type 1 judgment. Starting from the second session of the main task, if a participant's total performance score was higher than that in the previous session, they earned a $10 bonus.

Data Analyses

For our main analysis we adopt the approach of fitting Type 2 psychometric curves as described previously (Knotts, Lau, & Peters, 2018). Briefly, for each participant we found the difference in motion discrimination d' between central and peripheral stimuli for all 36 combinations of central and peripheral coherence levels. We then plotted the percentage of trials in which higher confidence for each combination of central and peripheral coherence levels against the corresponding difference in motion discrimination d' (Figure 13b,c). As in Experiments 1.1 - 1.3, cumulative normal functions were then fit to each participant's data with free parameters α

(threshold) and β (slope), and fixed parameters γ (lapse rate) = 0 and δ (guess rate) = 0 using the Palamedes toolbox (Kingdom & Prins, 2010; Prins & Kingdom, 2018). Also as in Experiments 1.1 - 1.3, from these curves we estimated the point of subjective equality (PSE; magenta lines in Figure 13b,c) and the point objective equality (POE; red lines in Figure 13b,c) and performed two-tailed one sample t-tests to see whether these these were significantly different from 0 or 50%, respectively (see Data Analysis section under Experiment 1.1).

According to the inflation hypothesis we predicted a significant positive shift of the PSE away from 0 such that when participants are equally confident in their motion judgments at center and periphery, they will be better at discriminating motion at center. Similarly, we predicted a significant negative shift of the POE below 50% such that when participants are equally good at discriminating motion direction at center and periphery, they will be more likely to indicate high confidence in their peripheral motion judgments.

Analyses for each of Experiments 3.1-3.4 were conducted in Matlab R2014a (Natuck, MA), with the exception of repeated measures ANOVAs, which were conducted in SPSS v22 (IBM, Armonk, NY, USA). All repeated measures ANOVAs were adjusted for violations of the assumption of sphericity with the Greenhouse-Geisser correction when necessary.

Results & Interim Discussion

The mean ± s.d. percentage of eyetracking errors across all subjects on the main task was 5.18 ± 5.12%, indicating that participants were efficient in keeping their gaze fixated throughout the task. One subject was removed from the main analyses due to failure to perform the central Type 1 task above chance, despite motion coherence values for central stimuli being

surprisingly high (up to 90%). Type 1 performance scores for each of the remaining 3 subjects fell within the desired range (mean ± 1 s.e.m. d'_{center}: 0.47 ± 0.12, 0.74 ± 0.28, 1.33 ± 0.38, 1.75 ± 0.42, 2.02 ± 0.35, 2.90 ± 0.26, $d'_{periphery}$: 0.34 ± 0.10, 0.57 ± 0.11, 1.26 ± 0.29, 1.62 ± 0.34, 2.16 ± 0.59, 2.94 ± 0.75). A repeated measures ANOVA with dependent variable motion discrimination d' and within subjects factors motion coherence (6 levels) and eccentricity (central versus peripheral) showed an expected main effect of motion coherence [$F(1.13,2.25) = 26.194$, $p = 0.028$], i.e. that increased motion coherence led to higher performance. The ANOVA also showed no main effect of eccentricity [$F(1,2) = 0.01$, $p = 0.93$] and no interaction between coherence and eccentricity [$F(1.26,2.52) = 0.161$, $p = 0.77$]. These data suggest that d' was effectively matched between central and peripheral stimuli.

Previous studies on feature-based attention have shown objective performance advantages in discrimination tasks (e.g., orientation and motion) for covertly attended stimuli when simultaneously discriminated, overtly attended stimuli at a different retinal location contain congruent feature information (Lu & Itti, 2005; Sàenz, Buraĉas, & Boynton, 2003; Sally, Vidnyánsky, & Papathomas, 2009; White & Carrasco, 2011). To test for any such effect of the overt attention difference between central and peripheral stimuli in our task we ran an additional ANOVA on discrimination d' with the within-subjects factors motion coherence (6 levels), eccentricity (2 levels: central or peripheral), and motion direction congruence (2 levels: congruent or incongruent). The ANOVA showed no significant interaction between eccentricity and congruence [$F(1.20,2.41) = 0.27$, $p = 0.69$] and no main effect of congruence [$F(1,2) = 7.63$, $p = 0.11$] as may have been predicted by the feature-based attention studies above. However, this is likely due to the fact that the adaptive staircasing procedure was specifically designed to match performance between central and peripheral motion discrimination judgments (see

Experiment 3.1 Methods). To determine whether central-peripheral stimulus congruence had any impact on subjective awareness, we also ran a repeated measures ANOVA on the percentage of trials in which higher confidence was indicated in the central stimulus with the same within subjects factors, and again found no evidence for either an interaction between stimulus congruence and eccentricity [$F(1,2) = 2.03$, $p = 0.29$] or a main effect of stimulus congruence [$F(1,2) = 0.02$, $p = 0.91$].

The mean ± s.e.m. PSE value was -1.15 ± 0.12 d' difference units, which a two-tailed paired-samples t-test confirmed was significantly lower than 0 [$t(2) = -9.66$, CI(-1.66,-0.64), $p = 0.01$]. This suggests that when participants were equally confident in central and peripheral stimuli, they were better at discriminating motion for peripheral stimuli by an average of 1.15 d' units. Consistent with this result, the mean ± s.e.m. POE value was 60.0 ± 4.3%. While this did not reach statistical significance [$t(2) = 2.26$, CI(41.3%,77.9%), $p = 0.15$], the fact that every participant had a POE of at least 55% ($subject_1$: 68.1%, $subject_2$: 55.4%, $subject_3$: 55.3%) means that when motion discrimination d' was matched, all participants were, on average, more likely to indicate higher confidence in the central dot motion stimulus.

This result was surprising in that it is the opposite of what we predicted based on previous studies showing subjective inflation in the periphery. Nonetheless, the PSE being significantly lower than zero is indicative of relative blindsight (Lau & Passingham, 2006), and points to a tantalizing question about peripheral motion perception: can participants discriminate motion in the periphery unconsciously? We examine this question in Experiment 3.2.

Experiment 3.2: Peripheral Two-Interval Forced Choice Dot Motion Discrimination With Null Interval

Methods

Participants

Four participants (all female, ages 18-21, all right handed, 3 who also participated in Experiment 3.1), gave written informed consent to participate. All participants had normal or corrected-to-normal eyesight and normal stereo vision, and all were either paid $13.25 USD per hour plus performance-contingent bonus money for their participation.

Apparatus, Stimuli, & Procedure

The apparatus, stimuli, and procedure used in Experiment 3.2 were the same as those in Experiment 3.1 with the following exceptions. While the task in Experiment 3.1 allowed us to examine relative blindsight between central and peripheral motion judgments, in Experiment 3.2 we adapted a two-interval forced choice approach (Peters & Lau, 2015) to examine absolute blindsight for dot motion discrimination in the periphery (Figure 14a). Participants were again required to fixate a central cross for 1 s to initiate a trial sequence, after which the cross would remain on the screen for a variable interval from 100 to 150 ms. This was followed by a 400 ms presentation of a peripheral dot motion stimulus at one of the same four peripheral locations from Experiment 3.1 (dashed circles in Figure 14a). After a 1 s inter-interval interval (III) another gaze-contingent cross appeared and was followed by another dot motion stimulus at one of the same four peripheral locations. Critically, motion coherence in one of the two intervals (hereafter denoted as the target-absent interval) was always zero, meaning one interval always contained no net horizontal motion. We refer to the interval containing nonzero net horizontal motion as the target-present interval. Following presentation of both dot motion stimuli, participants made

the Type 2 judgment of indicating the interval in which they were more confident in their ability to discriminate the net left/right motion of the dot motion stimulus. They then made the Type 1 motion discrimination judgments for the dot motion stimuli in intervals 1 and 2, always in that order.

Participants again performed 10 practice trials and an 80-trial adaptive procedure with four interleaved staircases (QUEST, Watson & Pelli, 1983) to estimate the motion coherence that would lead to 75% correct motion coherence discrimination performance. Importantly, in both the practice and adaptive staircasing procedures, both intervals contained nonzero motion coherence. This was intended to keep participants under the impression that both stimulus intervals contained nonzero motion coherence throughout the main task. This also allowed for a more efficient estimation of motion thresholds during the staircasing procedure. The resulting thresholds were multiplied by ten proportions (mean ± s.e.m. proportions: 0.13 ± 0.00, 0.23 ± 0.00, 0.31 ± 0.007, 0.38 ± 0.01, 0.45 ± 0.01, 0.53 ± 0.02, 0.61 ± 0.02, 0.70 ± 0.03, 0.80 ± 0.05, 0.12 ± 0.01) in order to target Type 1 performance levels in the range from roughly 53% to 85% correct. Subjects received trial-by-trial feedback regarding the accuracy of their motion judgments in both intervals during both the practice and adaptive staircasing parts of the experiment.

We used a randomized full factorial combination of target coherence (10 levels), interval 1 peripheral location (4 locations), interval 2 peripheral location (4 locations), motion direction (2 directions), target interval (2 intervals), with 8 trials per unique combination of conditions for a total of 5,120 trials per subject. Trials were divided into 128 40-trial blocks. As in experiment 3.1,

participants performed the man task over the course of several days (mean ± s.d. days per subject: 13.8 ± 1.7).

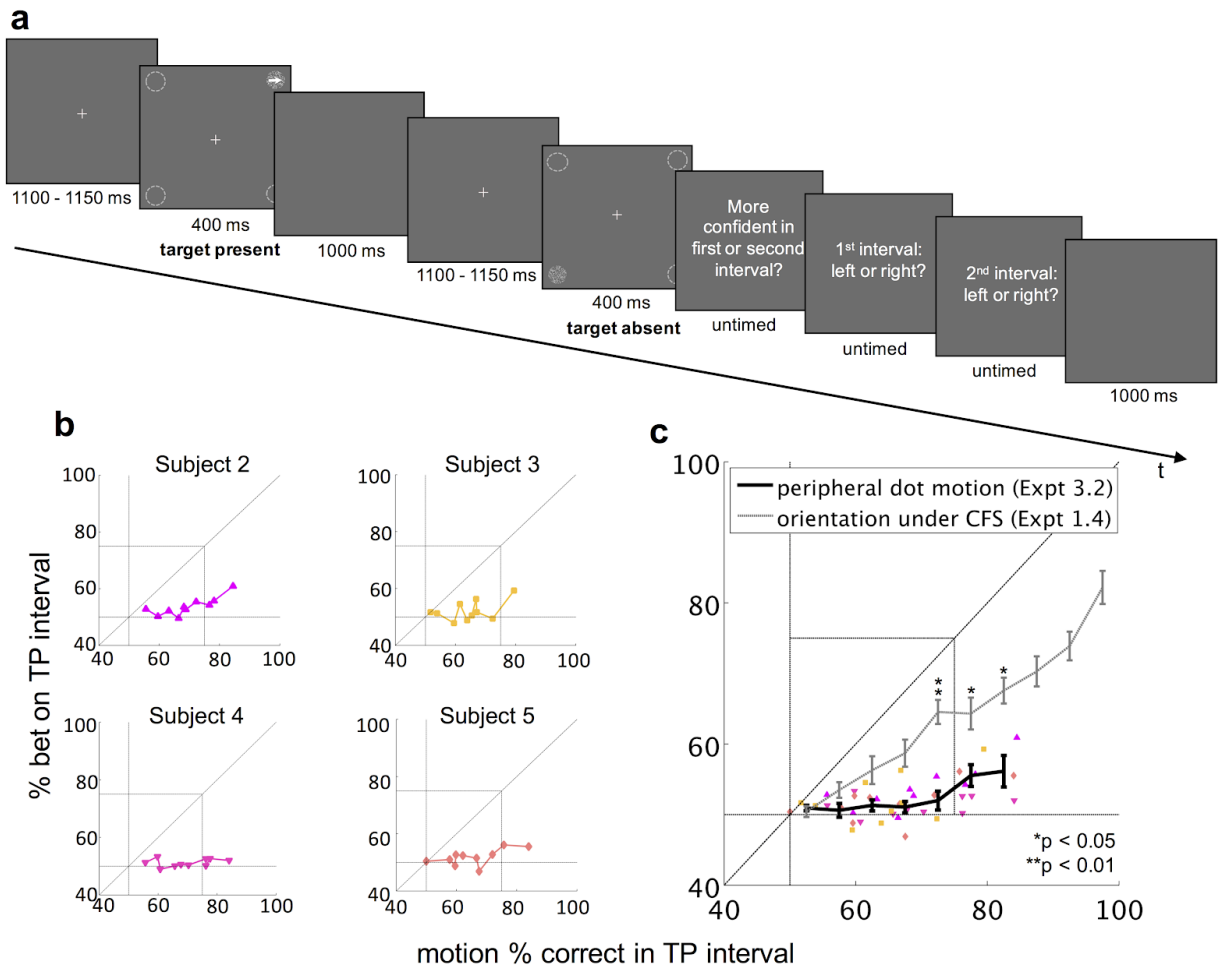

Figure 14. Procedure and results for peripheral two-interval forced choice with null interval task (Experiment 3.2). **a.** On each trial participants saw a brief fixation cross followed by a random dot motion kinematogram presented at one of four peripheral locations (dashed circles). After a 1000 ms interstimulus interval they then saw another fixation cross followed by another dot motion kinematogram presented at one of four peripheral locations. Unbeknownst to participants, one interval always contained a net leftward or rightward motion coherence of zero. Participants then indicated the stimulus interval (first or second) in which they were more confident in their ability to discriminate the net direction (leftward versus rightward) of motion. They then judged the net direction of motion in the first and second intervals, respectively. Font sizes are enlarged here relative to their actual size during the experiment for clarity. **b.** Individual Type 2 versus Type 1 responses. The percentage of trials in which the participant indicated higher confidence in (or "bet on") the target present interval (Type 2 judgment) is plotted as a function of motion discrimination percent correct scores for the target present interval (Type 1 judgment). **c.** Group level Type 2 versus Type 1 data compared with that from Experiment 1.4.

The mean ± s.e.m. percentage of trials in which participants bet on the target present interval in each of seven evenly spaced bins from 50% to 85% Type 1 accuracy in Experiment 3.2 (solid black line) is shown overlaid on the raw data (see panel **b** for mapping to individual subjects). The binned means ± s.e.m. of the same measures from Experiment 1.4 are shown in gray. Binned data are shown centered within each bin. Significant results from Wilcoxon rank-sum tests comparing bin means between experiments are indicated with asterisks (*p<0.05, **p<0.01; note that these are not corrected for multiple comparisons).

Data Analyses

Removing net horizontal motion from one of the two stimulus intervals in every trial allowed us to investigate unconscious peripheral dot motion discrimination in the following way as described above in the Methods for Experiment 1.4. Briefly, we plot the percentage of trials in which the participant indicated higher confidence in the target-present interval as a function of motion discrimination performance for the target-present interval (Figure 14b,c). If a line plot of these data shows a flat region such that x becomes substantially larger than 50% correct Type 1 accuracy while the percentage of trials in which the participant indicated higher confidence in the target-present interval stays at 50%, then it would suggest that despite being able to perform the motion discrimination task above chance, the participant was not able to subjectively distinguish a stimulus with a discriminable signal from one with no discriminable signal (Peters & Lau, 2015). Such behavior could therefore be interpreted as evidence for unconscious motion discrimination.

To formally analyze the flatness of these Type 2 “absolute blindsight” curves we directly compared the data from Experiment 3.2 to that of Experiment 1.4, in which the same 2IFC procedure with a null interval was used. We compared the data in two ways. First, we ran a repeated measures ANOVA on % correct scores with the within subjects factors stimulus strength (10 levels) and response type (Type 1 or Type 2) and the between subjects factor task (peripheral dot motion discrimination or orientation discrimination under CFS). If the Type 2

curve in Experiment 3.2 is indeed flatter than that in Experiment 1.4, then we should expect an interaction between response type and task.

Second, we averaged the data from each experiment in 10 equally spaced bins of Type 1 performance (orientation or motion discrimination) from 50% to 100% correct (Figure 15c). Due to the lower trial number per participant in Experiment 1.4, some Type 1 percent correct scores were lower than 50 % correct. These were lumped into the lowest bin from 50 to 55% correct on the assumption that sub-50% correct scores effectively represent chance performance. To ensure equal weighting of the data from each subject in a given bin, we first computed the mean of each individual subject's data within that bin, and then computed the average of those means. In the case of a significant interaction between response type and task, post-hoc Wilcoxon rank sum tests can be performed on the individual subject means for each task. Comparing binned means will ensure that Type 1 performance is matched when comparing between tasks.

Results & Interim Discussion

The mean ± s.d. percentage of eyetracking errors across all subjects on the main task was 4.60% ± 5.09%, indicating again that participants were efficient in keeping their gaze fixated throughout the task. For the main analysis of absolute blindsight, visual inspection of individual data suggests that each participant placed confidence judgments on the target present and target absent intervals roughly equally until Type 1 performance was at least (Figure 14b, Subject 2) about 65% correct. Group data was binned by target present Type 1 accuracy in seven equally spaced (5%) bins from 50% to 85% correct (Peters & Lau, 2015). Again, visual inspection suggests that at the group level, participants are near chance in their confidence

judgments until at least the 5th bin of motion discrimination accuracies, the lower bound of which is 65% correct.

The repeated measures ANOVA comparing Experiments 1.4 and 3.2 showed a significant interaction between response type and task [$F(1,24) = 17.44$, $p < 0.001$]. The nature of this interaction is clear when plotting binned means of Type 2 responses as a function of Type 1 accuracy (Figure 15c). When subjects perform the CFS task, they appear to reliably bet on the target present interval as soon as Type 1 performance enters the second bin between 55% and 60% correct (Figure 15c, gray line, second bin). This is in sharp contrast to the peripheral dot motion task, where subjects, on average, do not appear to be able to reliably bet on the target present interval until Type 1 performance is somewhere between 70 and 75% correct. This was confirmed by post-hoc rank sum tests betwetween binned Type 2 responses for each task, which showed that over the range of at least 70% to 85% correct Type 1 accuracy, participants were significantly more likely to bet on the target-present interval when performing the central orientation discrimination task under CFS (Figure 15c).

These results provide some preliminary evidence for absolute blindsight for dot motion discrimination in the periphery. In other words, if the target-present interval is subjectively indistinguishable from the target absent interval up to Type 1 motion discrimination accuracies of 65%, this suggests that participants may be able to perform the task up to this level of accuracy without subjective awareness. However, alternative interpretations are considered in the general discussion. Importantly, if the blindsight-like effect found in Experiment 3.2 is critically related to the dissociation found in Experiment 3.1, then we should not expect to find a similar result if we

repeat Experiment 3.2 with all stimuli at central fixation. We asked this question in Experiment 3.3.

Experiment 3.3: Central Two-Interval Forced Choice Dot Motion Discrimination With Null Interval

Methods

Participants

Three participants (all female, ages 18-21, all right handed, all of whom also participated in Experiments 3.1 & 3.2), gave written informed consent to participate. All participants had normal or corrected-to-normal eyesight and normal stereo vision, and all were either paid $13.25 USD per hour plus performance-contingent bonus money for their participation.

Apparatus, Stimuli, & Procedure

The apparatus, stimuli, and procedure used in Experiment 3.3 were identical to those used in Experiment 3.2 with the following exceptions. Instead of presenting dot motion stimuli in the peripheral locations from Experiments 3.1 and 3.2, all dot motion stimuli were presented at the central fixation point (Figure 15a). Furthermore, the total number of trials was halved; we used a full factorial combination of target coherence (10 levels), motion direction (2 directions), and target interval (2 intervals), with 64 trials per unique combination of conditions for a total of 2,560 trials per subject (mean ± s.d. days per subject: 9.00 ± 2.65). Proportions used to determine experimental coherence strengths from Quest-estimated coherence thresholds were equal across all subjects: 0.12, 0.19, 0.25, 0.32, 0.40, 0.49, 0.59, 0.71, 0.86, 1.05.

Data Analyses

Type 2 absolute blindsight curves were plotted the same as in Experiment 3.2. Given that all participants in Experiment 3.3 also participated in Experiment 3.2, we tested for any potential differences in the relationship between Type 1 and Type 2 performance between the two experiments using a repeated measures ANOVA on % correct scores with within-subjects factors motion coherence (10 levels), response type (Type 1 versus Type 2), and eccentricity (central versus peripheral).

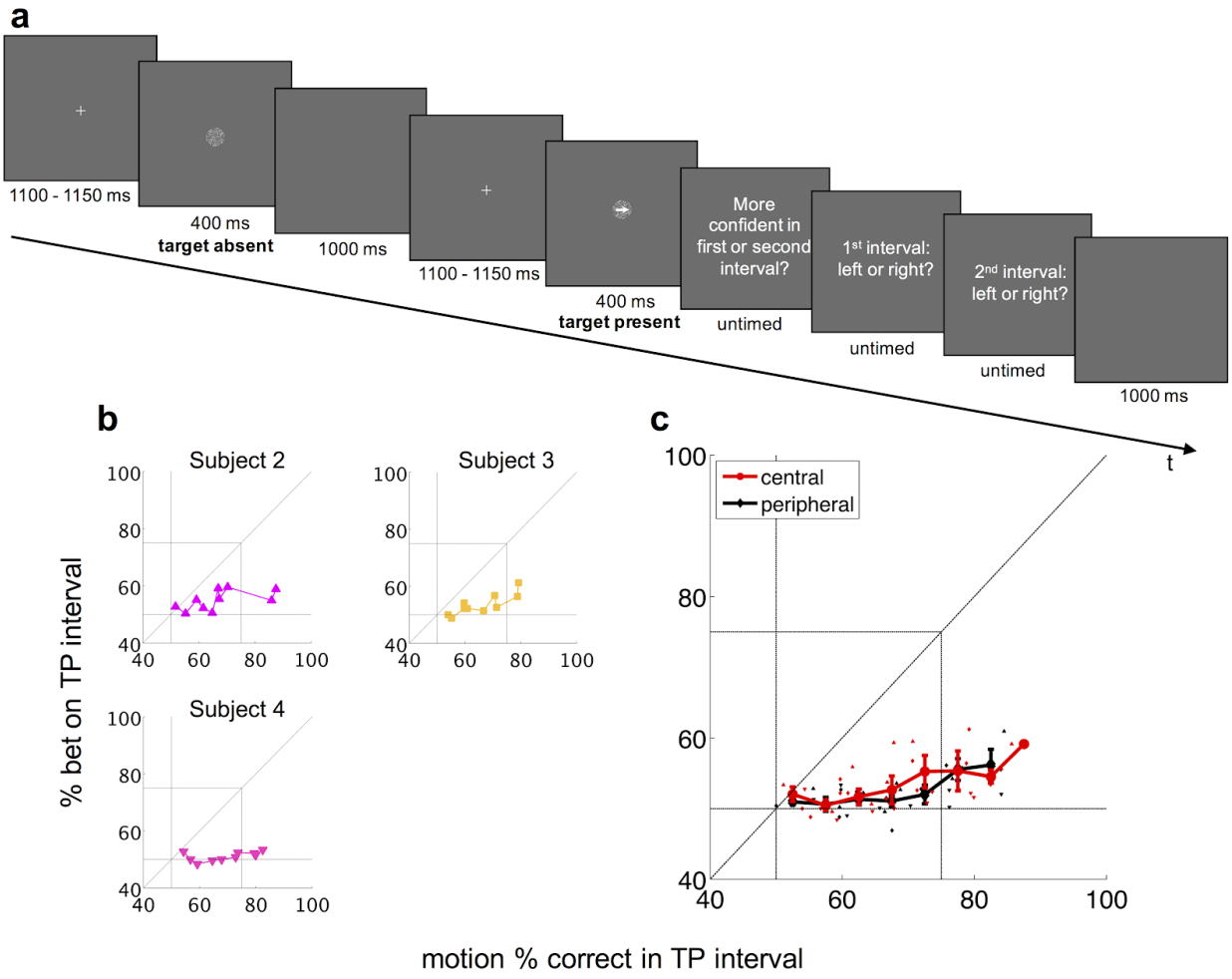

Figure 15. Procedure and results for central two-interval forced choice with null interval task (Experiment 3). **a.** The trial procedure was identical to those in Experiment 3.2 except that dot motion stimuli were always presented centrally. Font sizes are enlarged here relative to their

actual size during the experiment for clarity. **b.** Individual Type 2 versus Type 1 performance plotted the same as in Figure 14. **c.** Group level (N=3) Type 2 versus Type 1 performance data at center (red, Experiment 3.3) compared to periphery (black, Experiment 3.2). Mean ± s.e.m. Type 2 performance scores are overlaid over the raw data.

Results & Interim Discussion

The mean ± s.d. percentage of eyetracking errors across all subjects on the main task was 4.49% ± 5.13%. Individual and group level Type 2 absolute blindsight plots are shown in Figures 15b and 3.3c (red data points and line), respectively. The repeated measures ANOVA showed no significant interaction between motion coherence and eccentricity [f(1.99,3.98) = 2.60, p = 0.19], which confirms that a similar range of Type 1 motion discrimination % correct scores was observed between Experiments 3.2 and 3.3 (see ranges of overlaid raw central and peripheral data in Figure 15c). Critically, there was no interaction between response type and eccentricity [F(1,2) = 1.16, p = 0.39]. This suggests that the extent of absolute blindsight-like behavior for foveal dot motion discrimination is the same as that in the periphery.

The lack of a difference in the relationship between Type 1 and Type 2 judgments between Experiments 3.2 and 3.3 suggests that the absolute blindsight-like behavior found in these experiments and the relative blindsight effect found between central and peripheral judgments in Experiment 3.1 are not underlain by the same mechanism. Perhaps the largest procedural difference between the relative and absolute blindsight paradigms used so far is the simultaneity of the dot motion stimuli. In the relative blindsight task (Experiment 3.1), central and peripheral stimuli are presented simultaneously, whereas in the absolute blindsight tasks (Experiment 3.2 and 3.3), target and non-target stimuli are presented consecutively. In the next experiment, we test whether this simultaneity difference may be critical to the relative blindsight effect observed in Experiment 3.1.

Experiment 3.4: Two-Interval Forced Choice Center Versus Peripheral Dot Motion Discrimination

Methods

Participants

Three participants (all female, ages 18-21, all right handed, two of whom also participated in Experiments 3.1 & 3.2 and one of whom previously participated in only Experiment 3.2), gave written informed consent to participate. All participants had normal or corrected-to-normal eyesight and normal stereo vision, and all were either paid $13.25 USD per hour plus performance-contingent bonus money for their participation.

Apparatus, Stimuli, Procedure, & Data Analysis

The apparatus, stimuli, and procedure used in Experiment 3.4 were the same as that in Experiments 3.2 and 3.3 with the following exceptions. On each trial the dot motion stimulus in one interval was presented centrally, while the dot motion stimulus in the other interval was presented peripherally (Figure 16a). Every dot motion stimulus contained nonzero net leftward or rightward motion coherence. An adaptive staircasing procedure with the same structure as that in Experiment 3.1 was used to target 6 near-threshold motion coherence levels per eccentricity condition (see Experiment 3.1 Methods). We used a full factorial combination of 6 coherence levels per eccentricity condition, 2 motion directions per eccentricity condition, 2 interval orders, and 4 peripheral stimulus locations with 2 trials per unique combination of conditions for a total of 2,304 trials (mean ± s.d. days per subject: 7.7 ± 2.1). Data analysis procedures followed those from Experiment 3.1. Mean ± s.e.m. proportions of Quest-estimated threshold coherence values used for stimuli were as follows: $\text{Proportions}_{\text{center}}$ = 0.13 ± 0.02, 0.25

± 0.01, 0.37 ± 0.01, 0.59 ± 0.01, 0.90 ± 0.00, 1.28 ± 0.08; $Proportions_{periphery}$ = 0.12 ± 0.02, 0.23 ± 0.02, 0.34 ± 0.02, 0.55 ± 0.04, 0.84 ± 0.05, 1.19 ± 0.05.

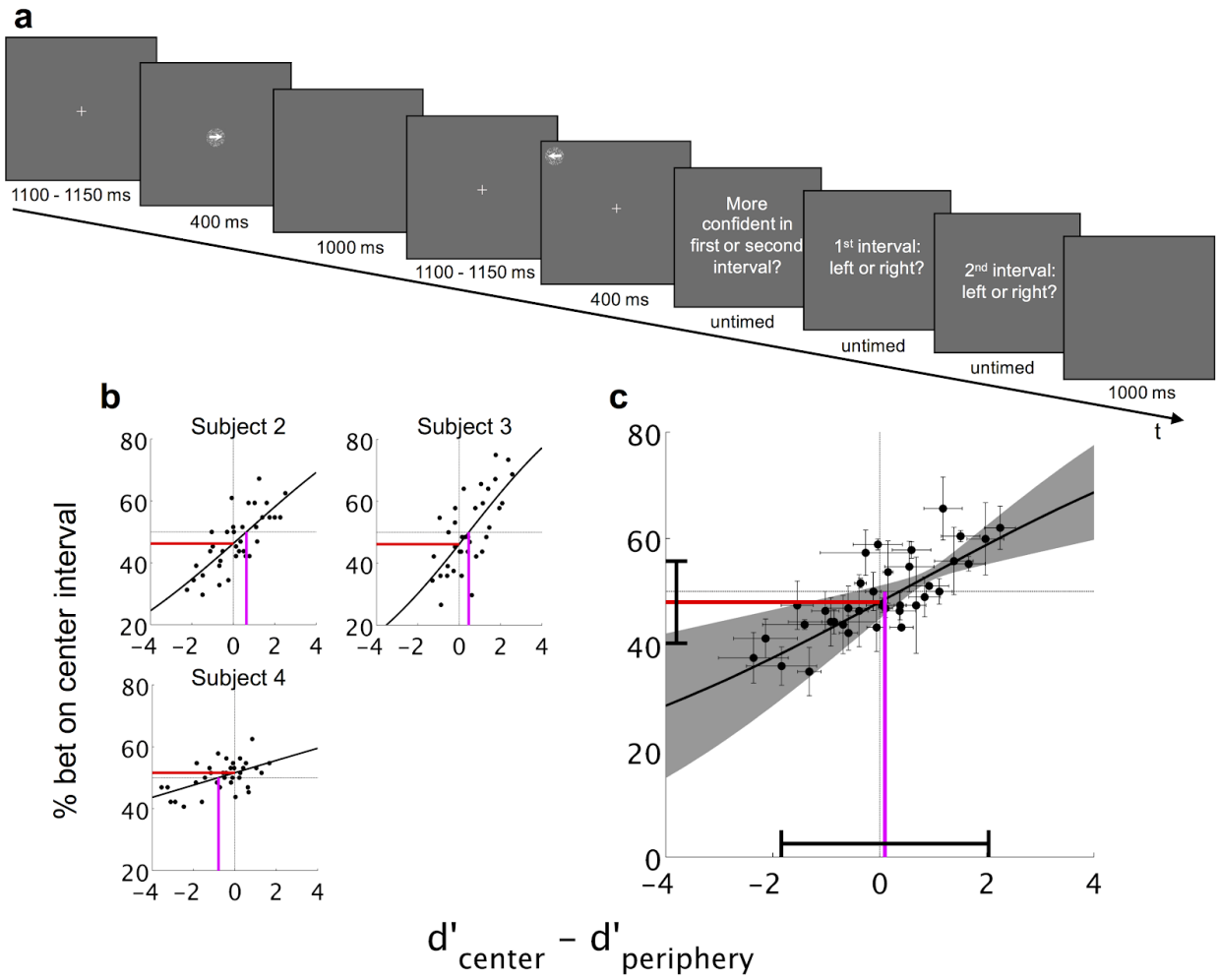

Figure 16. Procedure and results for 2IFC central versus peripheral dot motion discrimination task (Experiment 3.4). **a.** The trial procedure had the same temporal structure as Experiments 3.2 and 3.3. On each trial one dot motion stimulus was presented centrally whereas the other was presented at one of four peripheral locations, with the central vs peripheral location order randomized across trials. Further, all dot motion stimuli contained net leftward or rightward motion greater than zero. Font sizes are enlarged here relative to their actual size during the experiment for clarity. **b.** Individual psychometric curves. Individual Type 2 performance (percentage of trials in which participants indicated higher confidence in the interval containing the centrally presented stimulus) is plotted as a function of the difference between central and peripheral motion discrimination d'. Cumulative normal curves were fit to individual subject data as in Figure 13b. **c.** Group level (N=3) data. Black circles represent group mean d' difference scores ± 1 s.e.m., while the black line shows the mean of the individual cumulative normal fits ± 1 SD (gray). The 95% confidence intervals for estimated PSE and POE group means are shown by the black bars near the x- and y-axes, respectively.

Results & Interim Discussion

The mean ± s.d. percentage of eyetracking errors across all subjects on the main task was 4.45 ± 5.40%. As in Experiment 3.1, type 1 performance scores fell within the targeted range (mean ± 1 s.e.m. d'_{center}: 0.14 ± 0.13, 0.57 ± 0.16, 0.77 ± 0.23, 1.12 ± 0.23, 1.68 ± 0.38, 2.38 ± 0.29; $d'_{periphery}$: 0.27 ± 0.04, 0.50 ± 0.09, 0.85 ± 0.13, 1.34 ± 0.07, 1.95 ± 0.30, 2.78 ± 0.64). The repeated measures ANOVA with dependent variable motion discrimination d' and within subjects factors motion coherence (6 levels) and eccentricity (central versus peripheral) showed the expected main effect of increasing motion discrimination % correct scores with motion coherence [$F(1.13,2.25) = 26.194$, $p = 0.028$]. The ANOVA also showed no main effect of eccentricity [$F(1,2) = 0.01$, $p = 0.93$] and no interaction between coherence and eccentricity [$F(1.26,2.52) = 0.161$, $p = 0.77$]. These data suggest that d' was effectively matched between central and peripheral stimuli.

Type 2 cumulative normal psychometric fits (Figure 16b,c) showed PSE and POE shifts that, contra Experiment 3.1, were neither consistent across participants nor significantly different from the hypothetical null values of 0 and 50%, respectively, at the group level [$t_{PSE}(2) = 0.22$, 95% $CI_{PSE} = (-1.84,2.03)$, $p_{PSE} = 0.85$; $t_{POE}(2) = -1.12$, 95% $CI_{POE} = (40.3\%,55.7\%)$, $p_{POE} = 0.38$]. This therefore suggests that relative blindsight effect observed between center and periphery in Experiment 3.1 is task-specific. This result is also consistent with the lack of a difference in the extent of absolute blindsight-like behavior found at center and periphery in Experiments 3.2 and 3.3.

General Discussion

Across three experiments (3.1-3.3) we found evidence for both relative (Lau & Passingham, 2006) and absolute blindsight (Weiskrantz, 1986) in dot motion discrimination in human participants. In a fourth experiment (3.4) we found evidence that these effects are likely independent, and, in the case of relative blindsight, task-specific.

In Experiment 3.1 we found that when objective motion discrimination performance was matched for central and peripheral stimuli in a simultaneous central/peripheral motion discrimination task, participants were more likely to indicated higher confidence in their central motion discrimination judgments (Figure 13). In Experiment 3.2 we found that even when participants could objectively discriminate the motion direction of a near-threshold peripheral dot motion stimulus at accuracies up to roughly 75% correct, they could not subjectively distinguish such judgments from those made for stimuli with no net coherent motion. This may provide evidence for absolute blindsight (Figure 14). Further, this introspective insensitivity was found to be significantly worse than that observed when the same paradigm was used with orientation discrimination under continuous flash suppression (Figure 14c; Experiment 1.4). In Experiment 3.3 we found that this potential absolute blindsight effect was not limited to the periphery; participants were equally poor at introspecting on their motion judgments when the same task as in Experiment 3.2 was performed with stimuli presented foveally (Figure 15). Finally, in Experiment 3.4, we found that presenting central and peripheral stimuli consecutively in a 2IFC paradigm as opposed to simultaneously (Experiment 3.1), removes the bias towards indicating higher confidence in central judgments when motion discrimination performance is matched in the periphery that was observed in Experiment 3.1 (Figure 16).

While Experiment 3.4 indicates some task constraints on the relative blindsight effect observed in Experiment 3.1, this effect is nonetheless important in providing preliminary evidence for a reliable method for dissociating objective and subjective awareness (Lau, 2008). Surprisingly, the direction of the observed dissociation, that participants showed higher confidence in central vision when objective performance was matched in the periphery, was the opposite of what we hypothesized based on previous reports of subjective inflation in the periphery (M. K. Li et al., 2018; Odegaard, Chang, et al., 2018; Rahnev et al., 2011; Solovey et al., 2015). A potentially important difference is that subjective judgments in Experiment 3.1 involved direct within-trial comparisons between central and peripheral motion judgments, whereas previous inflation effects have been found by comparing differences in the average magnitudes of single stimulus confidence judgments across separate central and peripheral trials. However, the lack of Type 2 bias toward either eccentricity condition in Experiment 3.4 suggests that the Type 2 judgment being a direct comparison cannot alone explain the central bias observed in Experiment 3.1.

Comparing the task structures of Experiments 3.1 and 3.4 suggests that the critical factor may be the simultaneity of the central and peripheral stimuli. In this case, the underlying mechanism may be attentional. For example, previous studies of feature-based attention have found that discrimination performance improves when overtly attended stimuli are congruent with simultaneously presented, covertly attended stimuli with respect to a specific feature (e.g., motion direction or orientation (Lu & Itti, 2005; Sàenz et al., 2003; Sally et al., 2009; White & Carrasco, 2011). However, this effect has only been studied in reference to objective performance (e.g., discrimination d'), and here we unsurprisingly found no such effect, as objective performance was deliberately matched between eccentricity conditions via staircasing. A repeated measures ANOVA further revealed that stimulus congruence also did not impact the

likelihood of indicating higher confidence in central judgments. Together, these results suggest that a feature-based attentional explanation of the subjective bias observed in Experiment 3.1 is unlikely. However, this does not preclude the dual-task nature of Experiment 3.1 from being the critical experimental manipulation underlying the subjective bias effect. On this note, future studies should also examine whether established peripheral inflation effects (Rahnev et al., 2011; Solovey et al., 2015) persist under such dual-task conditions.

Another possible explanation is that the subjective bias observed in Experiment 3.1 is decisional as opposed to perceptual or sensory in nature (Linares, Aguilar-Lleyda, & López-Moliner, 2019; Witt, Sugovic, & Wixted, 2012). It is intuitive that, in the face of uncertainty, participants might default to indicating higher confidence in central judgments simply based prior knowledge and experience that visual acuity is better at the fovea than it is in the periphery. While it is not clear that there is a definitive method for disambiguating sensory and decisional interpretations of perceptual decision making biases, future studies could get a clearer idea of the nature of the central bias from Experiment 3.1 by seeing if it persists in the presence of experimental manipulations like trial-by-trial feedback (Rahnev et al., 2011; Solovey et al., 2015) or button randomization (Linares et al., 2019).

The results of Experiments 3.2 and 3.3 suggest that introspection may be impaired for random dot motion discrimination in a manner that is independent of the relative blindsight-like effect observed in Experiment 3.1. Importantly, the flat portions of the Type 2 psychometric functions in Figures 14b,c and 15b,c provide the first evidence for objective performance in the absence of subjective awareness using a task that is not susceptible to criterion bias (Peters & Lau, 2015). Previous studies using the unbiased 2IFC approach here for orientation discrimination have

shown no such evidence for objective performance without awareness (Peters & Lau, 2015); Figures 5 and 14), suggesting that random dot motion discrimination may be an ideal task for studying subjective awareness. To our knowledge, the only study that has directly examined differences in metacognition between different types of discrimination task when Type 1 performance is matched looked at differences between contrast discrimination and orientation discrimination (Song et al., 2011). Otherwise, direct comparisons between discrimination tasks have typically focused on comparing discrimination accuracy (Halpern, Andrews, & Purves, 1999). Given the current results, future studies should examine more directly, within subjects, whether introspective access to type 1 discrimination sensitivity is really more impaired in random dot motion tasks than other traditional psychophysics tasks like orientation or shape discrimination.

It remains an open question, however, whether the failure to subjectively distinguish between the target present and target absent intervals in Experiments 3.2 and 3.3 really reflects unconscious motion discrimination. An alternative explanation is that participants made subjectively rich false alarms in the target absent intervals. In this case, they might indicate higher confidence in the target absent interval despite being subjectively aware of the externally valid motion in the target present interval. If this interpretation is correct, we might ask why this be the case with dot motion, but not orientation discrimination. One possible answer is that a random dot motion stimulus contains local motion information (e.g., the motion of a single dot) that can be incongruent with the stimulus's global motion information (e.g., in the target absent intervals in Experiments 3.2 and 3.3, a net horizontal motion of zero). Therefore, if one selectively attends local information, it might lead to the false impression of a global motion signal when there is none. An oriented gabor patch, on the other hand, does not typically

contain any such incongruence between local and global orientation information, and therefore does not present this kind of opportunity to make false alarms based on local information.

One way to control for attention to local motion signals in a random dot motion discrimination task is to reduce the lifetime of dots (i.e., the amount of time each dot is on the screen before it is redrawn at a random location; see Appendix B). While a previous study showed that changing dot lifetime does not affect random dot motion detection thresholds (Scase, Braddick, & Raymond, 1996), the influence of dot lifetime on subjective measures of perception has not, to our knowledge, been explored. In each of the current experiments, the dots had full lifetimes, meaning the only time a dot was redrawn was when it would exit the boundary of the stimulus and then reappear on the other side. In theory, if dot lifetime is reduced, then it should be more difficult for subjects to attend to local motion signals. Future studies should investigate how this and other stimulus properties might modulate introspective access on this task. For another example, motion blindsight was found in an actual blindsight patient, GY, when using the same dot motion speed as was used here (1 dva/s), but not at higher speeds (Sahraie et al., 1998). If a similar pattern is found in healthy observers, it may provide additional evidence that the metacognitive deficits found in Experiments 3.2 and 3.3 reflect true unconscious processing.

It should be clarified here that while previous studies suggest that local dot trajectories do not significantly affect global motion perception (Watamaniuk, Sekuler, & Williams, 1989; Williams & Sekuler, 1984), subjects in these studies were not presented with kinematograms containing zero net motion in the directions available for forced choice responding. We believe that this may be an important distinction. It may be that it in such cases when there is no global signal available to sufficiently suppress evidence accumulation from local signals, and participants are

forced to make a decision, that these local signals have the strength to bias either the relevant pooling mechanisms in low level motion processing in area MT (Britten & Heuer, 1999; Britten, Shadlen, Newsome, & Movshon, 1992; Haberman & Whitney, 2012; Newsome & Paré, 1988), or later stage metacognitive systems, presumably in frontal and parietal areas (Stanislas Dehaene et al., 2017; Lau & Rosenthal, 2011). Whether one of these two loci is more strongly implicated in causing the high proportion of Type 2 false alarms observed in Experiments 3.2 and 3.3 is an open question that will be considered in the general discussion of this dissertation.

Regardless of the locus of the metacognitive impairment, we can see that the subjectively rich false alarms interpretation of Experiments 3.2 - 3.4, whether or not the exact mechanism relies on attention to local motion signals, bears a resemblance to the subjective inflation hypothesis. Of course, to the extent that inflation is operationally defined as having higher subjective ratings than would be predicted based on objective performance when either attention is reduced or eccentricity is increased, we cannot find inflation on the absolute blindsight tasks here (Expts 3.2 and 3.3); we only have one valid objective measure in each of these tasks, and there is no manipulation of attention or eccentricity. However, as mentioned above, global motion discrimination in random dot motion kinematograms involves pooling of local motion signals, as has been established in visual area MT in monkeys (Britten & Heuer, 1999; Britten et al., 1992; Haberman & Whitney, 2012; Newsome & Paré, 1988). On the lower level version of the rich false alarm interpretation, for a stimulus with zero net motion, bias in evidence accumulation from local signals may become magnified through pooling to such an extent that the observer subjectively perceives an unambiguous stimulus. This is similar to the presumed contribution of summary statistic computations to subjectively inflated percepts discussed above; in both cases,

the integration and compression of information in the visual system leads to unexpectedly high subjective awareness given the corresponding objective performance.

Finally, it is important to distinguish between theoretically versus operationally defined absolute blindsight when interpreting the current data. Theoretically defined absolute blindsight is true unconscious perception; it is a specific dissociation between objective performance and subjective awareness in which subjective awareness is at zero, and objective performance is not. Operationally defined absolute blindsight, here, entails betting on the target present and target absent intervals roughly equally when objective performance in the target present interval is above chance. This is important because, on the rich false alarms hypothesis, we essentially have operationally defined absolute blindsight without theoretically defined absolute blindsight. In this case, subjective awareness is presumably nonzero when objective performance is above chance; it is only because the Type 2 false alarms in the target absent interval also come with rich sense of subjective awareness that participants bet evenly between the two interval types. Therefore, despite the observed operational absolute blindsight effect, it is not clear whether our results reflect true unconscious motion discrimination, or a strong tendency to hallucinate motion that isn't really there.

One possible way to disentangle these interpretations would be to repeat the same absolute blindsight tasks (Experiment 3.2 and 3.3), but have subjects make absolute confidence judgments (e.g., on a scale from 1 to 4) in each stimulus interval in addition to the relative confidence judgment between the two intervals. If absolute confidence ratings are at floor only within the range of motion discrimination percent correct scores that corresponds to the flat portion of the absolute blindsight curve where subjects are equally likely to bet on target absent

and target present intervals (e.g., from 50% to roughly 70% in Figure 14c), it would provide some evidence that the metacognitive impairment reflects true unconscious perception, or absolute blindsight.

In conclusion, the present results suggest evidence for both relative and absolute blindsight in normal observers performing random dot motion discrimination tasks. The relative blindsight effect, which found that participants were more confident in central than peripheral left/right motion judgments, appears to be independent from previous demonstrations of peripheral inflation (Knotts, Odegaard, et al., 2018). Further, this effect may arise only in the context of simultaneous dual-discrimination tasks. The absolute blindsight effects observed here suggest that random dot motion discrimination at both central and peripheral retinal locations may be uniquely associated with impaired introspective access. But future studies should investigate whether this effect reflects true unconscious perception or a tendency to make rich false alarms. In the latter case, these data support an inflation-like account of dot motion perception across the visual field, which further supports the intermediate theoretical position (discussed in the background section of this chapter) that visual perception is objectively sparse but subjectively rich. In the former case, dot motion perception may represent an ideal task for isolating subjective measures of awareness in the scientific study of consciousness.

X. General Conclusions & Future Directions

The three lines of research described here inform each other in several ways. Perhaps the most striking of these concerns the difference we observed in performance on the null interval 2IFC task (Peters & Lau, 2015) between orientation (Experiment 1.4) and dot motion (Experiments

3.2-3.3) discrimination (Figure 14c); introspective access to objective discrimination judgments appears to be significantly worse for dot motion stimuli.

In Chapter IX we briefly discussed potential candidates for the locus of this introspective impairment in the visual system. Given the task difference, an obvious choice might be area MT (Britten & Heuer, 1999; Britten et al., 1992; Haberman & Whitney, 2012; Newsome & Paré, 1988). This fits with the theory that motion perception is less conscious due to its processing in the dorsal stream (Goodale, 2011; Milner & Goodale, 2008). And, as discussed above, it also fits the hypothesis that pooling mechanisms in MT may be vulnerable to inflation-like effects in the absence of strong external motion signals just as we propose to be the case for peripheral summary computations (Knotts, Odegaard, et al., 2018).

The neurofeedback results from Chapter VIII, however, suggest that the locus of the introspective impairment for dot motion may be a later metacognitive stage in frontal and/or parietal cortex. False color perception was associated specifically with decoded patterns for perceptual confidence and not lower level decoded patterns for color (Figure 11). Further, this association was strongest in PFC (Figure A3). Of course, this was a color task, not a motion task. However, evidence for a dissociation between perceptual confidence and objective perceptual decisions, specifically for dot motion perception, has been reported previously in both humans (Cortese et al., 2016) and monkeys (Odegaard, Grimaldi, et al., 2018). In the former study, perceptual confidence was decoded from the same frontoparietal areas as in the main neurofeedback study discussed in Chapter VIII. Further, this study found that perceptual confidence could be selectively increased via neurofeedback without affecting objective task performance. This very nicely suggests flexible frontoparietal confidence representations as

both a potential substrate for the mechanism of subjective inflation and, consequently, a potential locus for the metacognitive impairment observed in our behavioral dot motion data.

This is not to rule out the idea that it is ultimately an interaction between frontoparietal metacognitive mechanisms and low-level motion representations that underlies the observed metacognitive deficits in Experiments 3.2 and 3.3. At the beginning of Chapter IX we discussed the importance of expectation in mediating inflation-like effects [Knotts et al., kouider, pull a few others?]. This idea is consistent with the frontoparietal interpretation above, as there are several lines of evidence that expectation signals from frontoparietal cortex bias perception (Gau & Noppeney, 2016; Gilbert & Li, 2013; Gold & Shadlen, 2007; Rao, DeAngelis, & Snyder, 2012; Summerfield & Egner, 2009). Of particular importance, it has been found that such prior expectation signals can bias dot motion perception with corresponding modulations of neural activity in area MT in monkeys (Schlack & Albright, 2007), and in even earlier visual cortex in humans (Kok, Brouwer, van Gerven, & de Lange, 2013). On this note, it is critical to point out that in Experiments 3.2 and 3.3, the lack of an external stimulus precluded us from measuring objective performance in the target absent interval. Therefore, unlike the case of the relative blindsight tasks (Experiments 3.1 and 3.4), where objective performance is controlled for, it is more difficult to rule out the biasing of low-level motion representations as a critical contributing factor to the high proportion bets on the target absent interval that we observed.

Future neuroimaging and electrophysiology studies directly comparing dot motion discrimination with other perceptual tasks should examine the question of this task-specific deficit in more detail. In any case, the set of studies conducted for this dissertation provide evidence that dot motion discrimination represents an ideal task for dissociating objective and subjective

perception. The identification of such tasks is critical in that they will ultimately give researchers in the field better experimental access to the subjective component of visual perception.

The data herein further help arbitrate in the debate between first order (Block, 1995, 2007; Lamme, 2003) and higher order theories of consciousness (S. Dehaene & Naccache, 2001; Stanislas Dehaene et al., 2017; Lau & Rosenthal, 2011). On first order theories, phenomenology is rich, and maps onto activations in visual cortex, with prefrontal activity playing the secondary role of granting cognitive access to that phenomenology (Block, 2007). Conversely, on higher order theories, phenomenology is sparse, and maps, at least in part, onto cognitive mechanisms in frontal and parietal regions. The neurofeedback data from Chapter VIII clearly support the higher order view, as prefrontal activity was uniquely predictive of the subjective impression of color in the absence of external color stimulation. Additionally, the operational definition of subjective inflation argued for in Chapter IX supports the higher order view that cognitive access exerts a direct influence on the content of phenomenology, but that higher order mechanisms inflate subjective awareness of that phenomenology when cognitive access is limited. While we suggest that the ultimate position offered by the inflation argument is intermediate between the traditional Rich and Sparse views, the implication of higher order cognitive mechanisms in inflation are clearly in conflict with first order theories of consciousness.

To conclude, we can briefly summarize, in decreasing order of confidence, where this dissertation lands on the three debates in consciousness science it was designed to clarify. First, both the literature review and the association between decoded perceptual confidence and false color detection found in the neurofeedback study in Chapter VIII strongly suggest that prefrontal cortex is indeed critically involved in conscious visual perception. Provided that

sensitive analysis methods are used, the evidence for this connection seems to be clear. Next, on the debate about phenomenological richness, we conclude that an intermediate position based on the operationally defined notion of subjective inflation provides the most parsimonious account of the empirical literature. On this account, phenomenology feels rich despite its sparse representational foundations.

Lastly, where we land on the debate about unconscious perception in normal observers will depend largely on disambiguating the absolute blindsight versus rich false alarms interpretations of the null interval 2IFC data from Experiments 3.2 and 3.3 (discussed at length at the end of Chapter IX). At the present time, the author of this dissertation leans toward the inflation-like, rich false alarms interpretation, and thus, toward the conclusion that we still lack convincing evidence for true unconscious perception in normal observers. The author bases this leaning on both the existing evidence for other inflation and illusion effects in the literature, and on the intuition that this is a less extreme hypothesis overall; i.e., it seems more likely that an observer would mistakenly (from an external perspective) perceive coherent motion in an incoherent dot motion stimulus, than objectively perceive the motion in a coherent dot motion stimulus with a complete lack of subjective awareness. However, this remains a topic for future investigation.

XI. Appendices

Appendix A: Supplementary Information for Experiment 2, Multivoxel patterns for perceptual confidence are associated with false color detection

Supplementary Figures & Tables

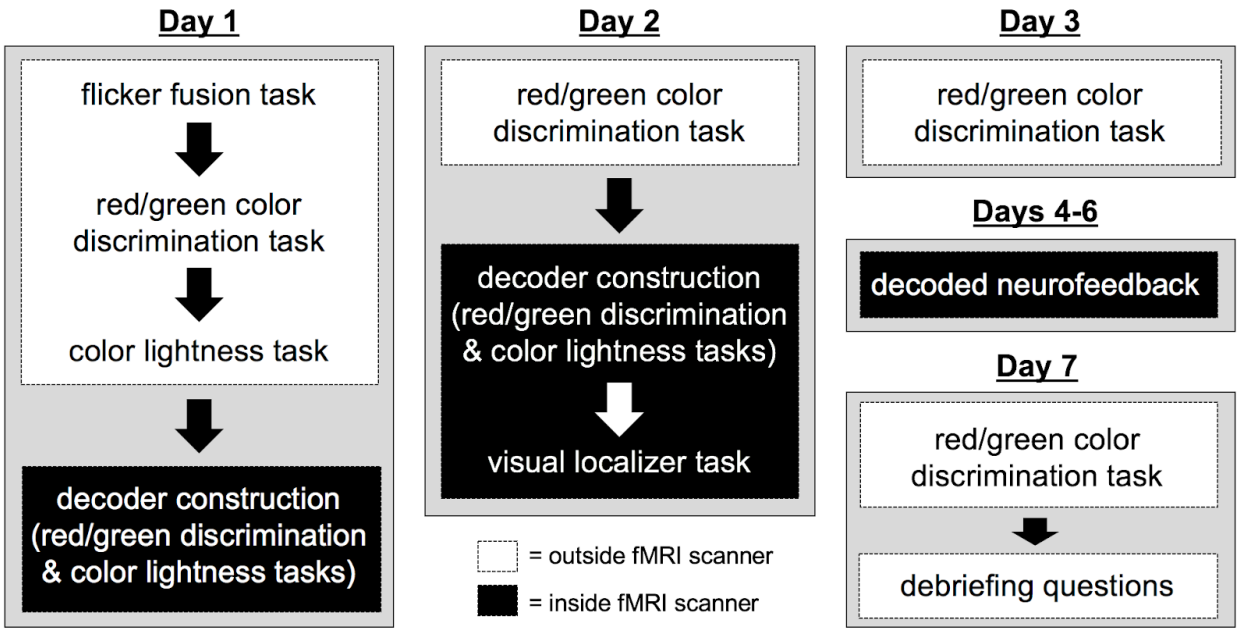

Figure A1. Task-by-task structure of Experiment 2. Tasks that were performed inside and outside of the fMRI scanner are shown in black and white boxes, respectively.

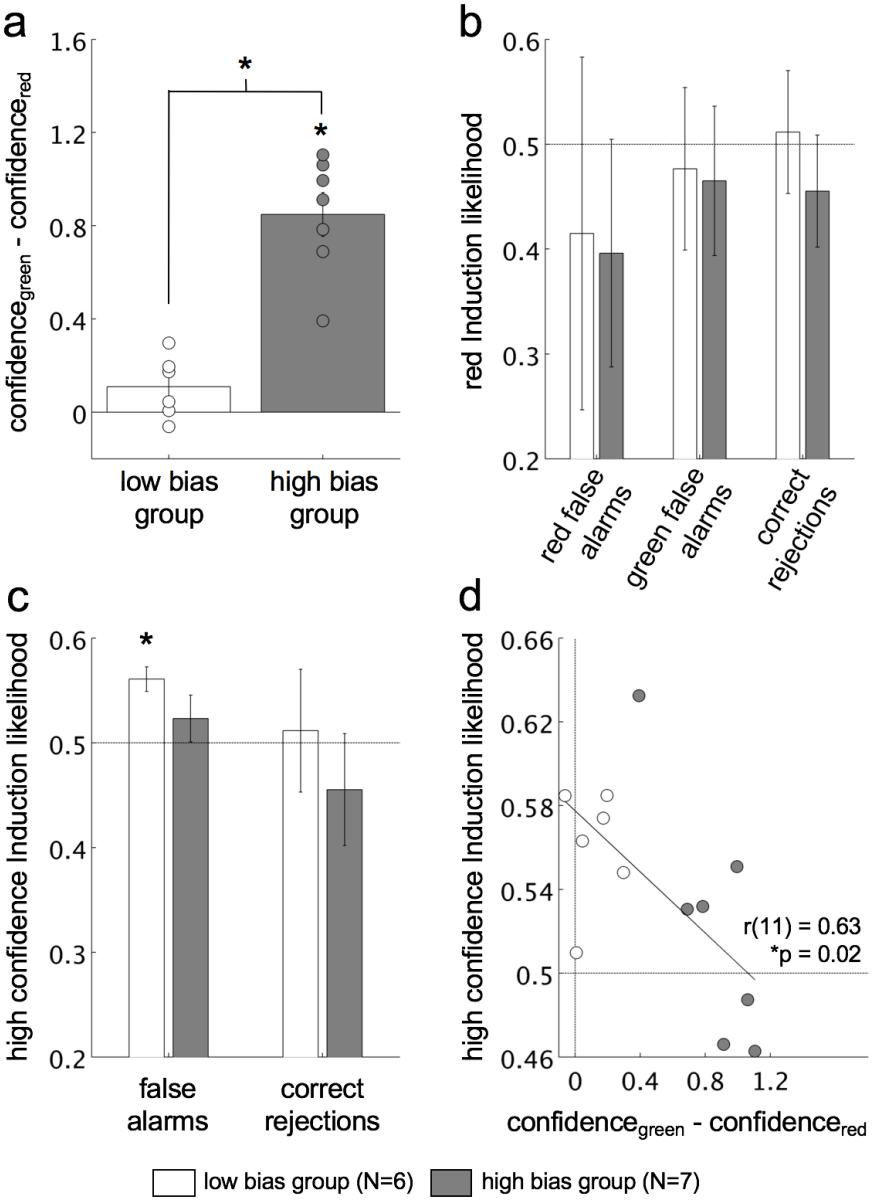

Figure A2. Median split analyses suggest that the association between high confidence induction and false alarms is not mediated by a bias for higher confidence in green decoder construction stimuli. **a)** A median split was conducted on the difference in mean confidence ratings for green stimuli and red stimuli. Because mean confidence ratings for red stimuli were subtracted from those for green stimuli, positive values suggest that, on average, subjects had higher confidence in green stimuli than red stimuli. As expected, the median split resulted in a significant difference in bias between the low bias group (white bar) and high bias group (gray bar) [$t(11) = -6.47$, $p < 0.001$, CI = (-0.99,-0.49), two-tailed, two-sample t-test]. Importantly, a one-tailed, one-sample t-test suggests that bias in the low group is not significantly different from zero at alpha = 0.05 [$t(5) = 1.99$, $p = 0.052$, CI = (-0.001,+inf)], although the low p-value suggests a trend in this direction. **b)** Red induction likelihoods for false alarm and correct rejection DecNef runs after median splitting on color-confidence bias. Median splitting showed no effect of color-confidence bias on red induction likelihoods during either false alarm or correction rejection runs. **c)** High confidence induction likelihoods for false alarm and correct

rejection runs after median splitting on color-confidence bias. Induction likelihoods for participants in the low color-confidence bias group were significantly above chance (0.50) during false alarm runs [$t(5) = 5.20$, $p < 0.01$, CI = (0.53,0.59), two-tailed, one-sample t-test] but not during correct rejection runs. High confidence induction likelihoods were not significantly different from chance for either false alarm or correct rejection runs for participants in the high color-confidence bias group. **d)** Extent of color-confidence bias is inversely correlated with high confidence induction likelihoods during false alarm DecNef runs [$r(11) = 0.63$, $p = 0.02$].

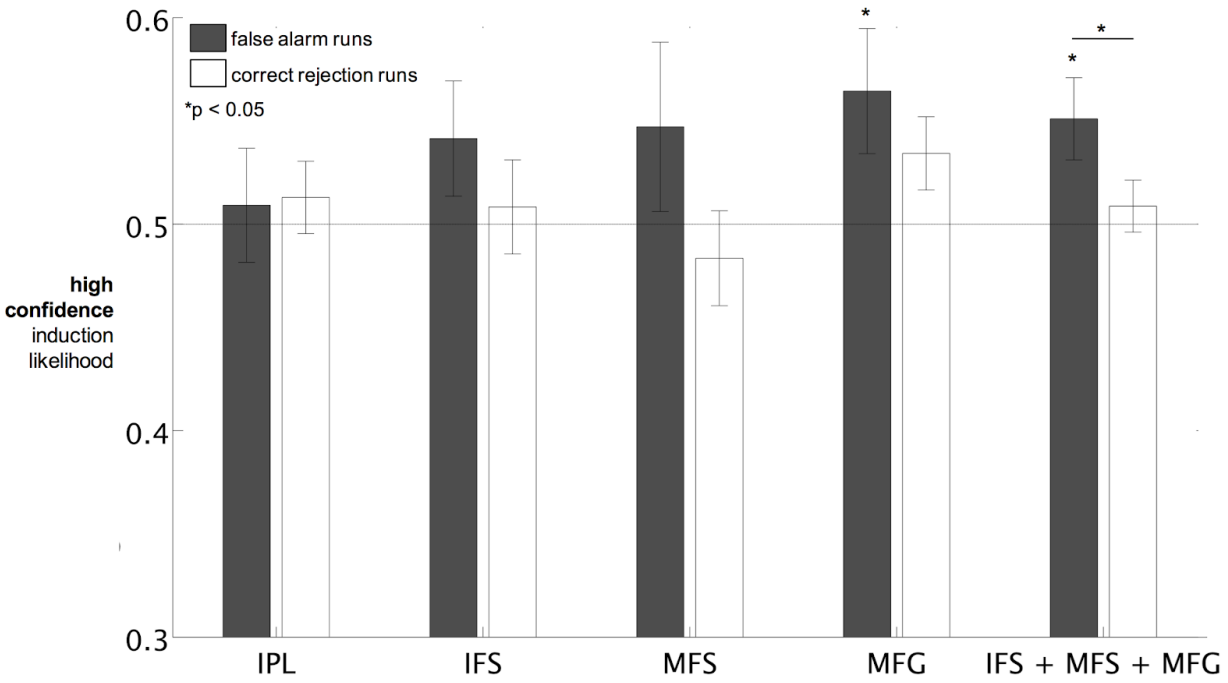

Figure A3. High confidence induction likelihoods during false alarm versus correct rejection runs in individual prefrontal and parietal ROIs and one group prefrontal ROI. While the difference in induction likelihoods between false alarm and correct rejection runs in the IFS + MFS + MFG ROI does not survive Bonferroni correction (family-wise alpha for comparing high confidence likelihoods between false alarm and correct rejection runs in each ROI = 0.01), the pattern of results suggests that the difference in high confidence induction likelihoods between false alarm and correct rejection runs found across all frontoparietal ROIs (Figure 9b) is primarily driven by the prefrontal ROIs. IPL, inferior parietal lobule; IFS, inferior frontal sulcus; MFG, middle frontal gyrus; MFS, middle frontal sulcus. *p < 0.05.

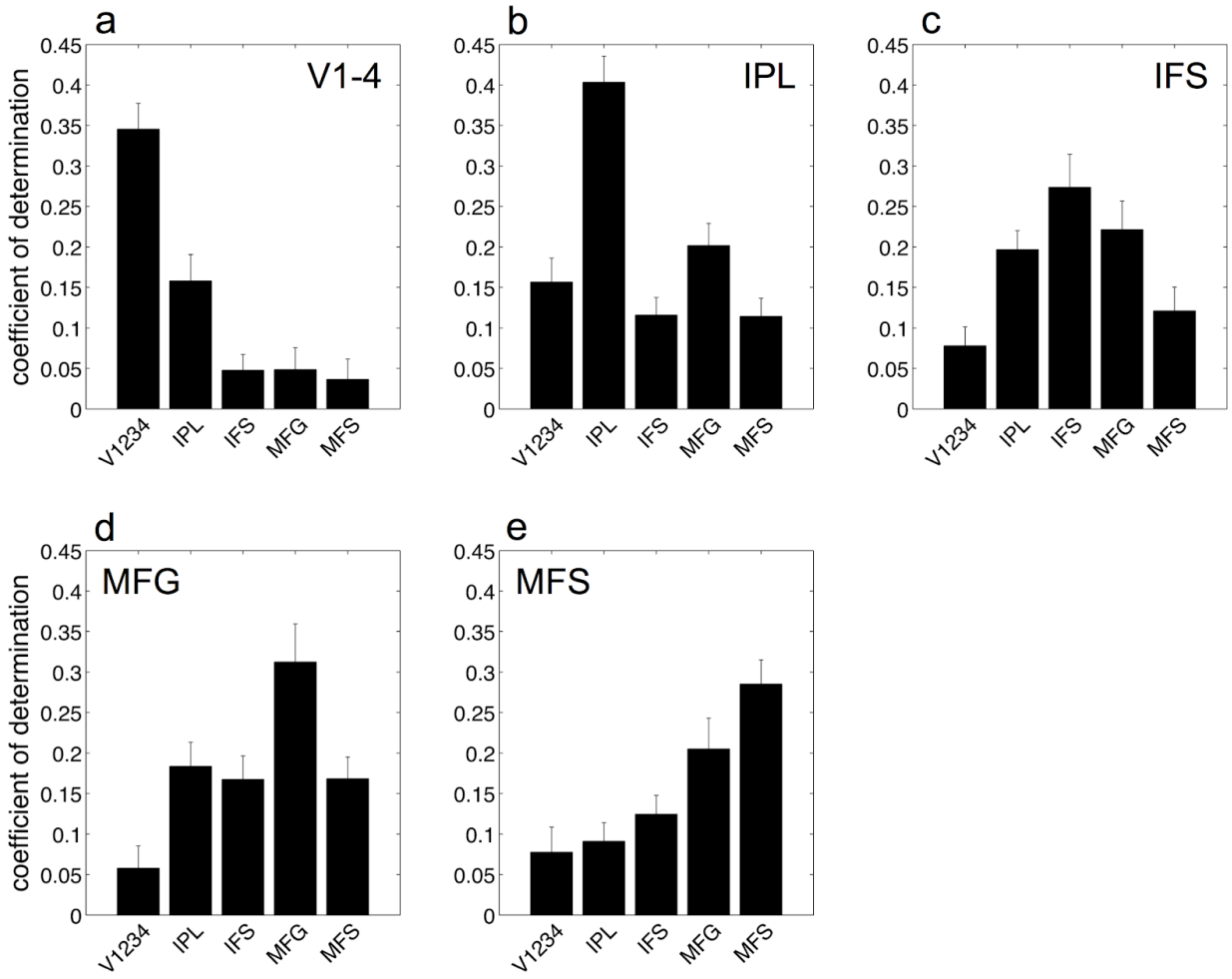

Figure A4. Information leak analysis (N=13). The coefficient of determination (y-axis) is an index of the extent to which voxel activities in a given "predictor" ROI (x-axis) can predict, via sparse linear regression (SLiR), color induction likelihoods in V1-4 **(A)** and confidence induction likelihoods in IPL **(B)**, IFS **(C)**, MFG **(D)**, & MFS **(E)**. The results show minimal "leak" of information outside of target regions, suggesting that induction likelihoods in a given ROI were minimally influenced by the activities of voxels in neighboring ROIs. This relationship is particularly pronounced when looking at "leak" between ROIs in frontal and visual cortices. V1-4: combined visual areas V1, V2, V3, & V4; IPL, inferior parietal lobule; IFS, inferior frontal sulcus; MFG, middle frontal gyrus; MFS, middle frontal sulcus.

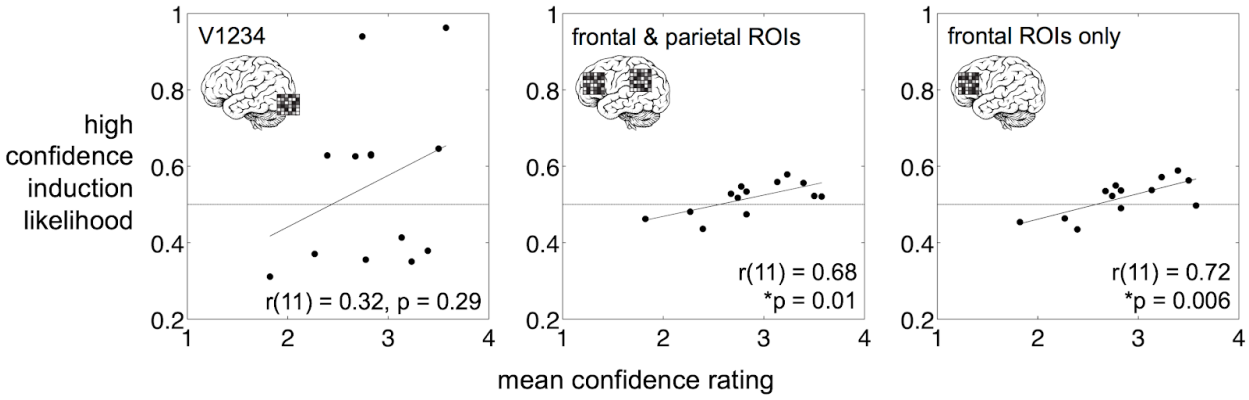

Figure A5. Relationship between mean high confidence induction likelihoods and mean DecNef confidence ratings across runs. High confidence induction likelihoods and confidence ratings were averaged across all DecNef runs for each subject. Confidence ratings were averaged across color for each DecNef run. The confidence decoder in V1234 was trained in the same manner as those in the frontoparietal ROIs (see Methods), but was not used for neurofeedback. Bonferroni corrected Pearson correlations suggest a relationship between DecNef confidence ratings and high confidence induction likelihoods in the collective frontoparietal ROI (middle panel: IPL + IFS + MFS + MFG, r(11) = 0.68, p = 0.01) , and when looking at only the frontal ROIs alone (right panel: IFS + MFS + MFG, r(11) = 0.72, p = 0.006), but not in visual cortex (left panel: V1-4, r(11) = 0.32, p = 0.29). IPL, inferior parietal lobule; IFS, inferior frontal sulcus; MFG, middle frontal gyrus; MFS, middle frontal sulcus.

Participant No.	Localizer	Color Decoder Time Window		Confidence Decoder Time Window	
		Start time relative to target stimulus onset (Δs)	Duration (s)	Start time relative to target stimulus onset (Δs)	Duration (s)
1	N	-2	6	-2	6
2	N	0	8	0	4
3	N	0	8	0	4
4	N	0	8	-2	6
5	Y	0	8	-2	6
6	N	2	6	0	4
7	N	0	8	0	4
8	N	0	8	0	4
9	Y	0	6	0	4
10	N	0	8	0	4
11	N	2	4	0	4
12	N	2	4	0	4
13	N	-2	8	0	4
14	Y	4	4	-2	6
15	N	-2	8	0	4
16	N	2	4	-2	4
17	Y	0	6	0	4

Table A1. Subject-specific temporal windows and V1-4 localizer intersection status that led to maximum decoding accuracy. A Y in the second column indicates that the maximum decoding accuracy was obtained when the V1-4 ROI was intersected with the functional localizer ROI, while an N indicates that the maximum decoding accuracy was obtained when the entire V1-4 ROI was used. Negative and positive numbers in columns 3 and 5 indicate temporal window starting times before and after target stimulus onset, respectively. The decoding parameters shown here were used to train the decoders that were subsequently used for neurofeedback.

Participant No.	Neurofeedback Strategies
1	imagined live rock music, playing sports, doing multiplication problems, a menu, working their old restaurant serving job, being a customer being served by their old self at that restaurant
2	imagined being happy, angry, sad, joyous; relaxed and mind wandered; counted prime numbers; imagined singing with friends; actively tried to think nothing
3	imagined vivid colors: sea and sky, green peppers, carrots; imagined playing soccer, basketball, baseball; imagined music and playing music, Monet paintings, puppies
4	imagined eating food -- it's appearance, texture, and taste; imagined chatting with friends, telling stories, the faces of people they'd seen the previous day, being rich, being an animal, getting hair done at a salon, getting married
5	imagined friends' faces after a car accident that they were in; tried to make themselves feel sad; imagined traveling; imagined painful things like running a marathon when you don't want to; imagined shouting angrily, singing; tried viewing the grating as clearly as possible; played word games
6	imagined what to eat for lunch; imagined a red grating (only 2 consecutive trials - the first made the feedback circle relatively large, the second did not); imagined the induction grating being the top of a barbecue on which they were grilling meats; imagined a green grating; imagined the induction grating being the bars of a jail cell, behind which was a famous baseball player who had been suspended for taking performance enhancing drugs; tried to empathize with that baseball player; wondered, "what is the meaning of this experiment"; imagined being in Las Vegas, viewing scenery through a moving train, the Grand Canyon, barbecue sauce, salt and pepper, what it would be like to run a business
7	imagined foreigners traveling through Japan; integrating equations in their mind; imagined being a train conductor, playing guitar, reading a music score, train station names, the story of a book they recently read, drawing difficult Hanzi; tried thinking about nothing; counted the seconds while the induction stimulus was on the screen; wished that the circle would get bigger
8	mental arithmetic, imagined music, singing, playing tennis and soccer, what that night's dinner would be, events that happened earlier that day; thought about their own research questions
9	tried to relax and think about nothing; imagined what it feels like to be in front of others at a swimming pool; imagined cutting and peeling the vertical bars of the grating, singing, walking on the black bars of the induction grating; tried to distribute attention widely across entire visual field
10	imagined past houses and school experiences, playing tennis, swimming, recent studies in telecommunications, the black disc around the fixation circle being blue, the fixation circle growing in size; focused on noise in the induction grating; imagined throwing a ball or shooting arrows at the fixation point
11	mental arithmetic: addition, multiplication, factoring; imagined running, doing track and field events; mind wandered; thought about how they could not control the size of the feedback circle; imagined the induction grating being green
12	mental arithmetic: addition, subtraction, multiplication, division; imagined writing sentences, scenery near their house, and playing soccer and baseball;
13	imagined friends' faces, feelings of depression, playing soft tennis, the smell of barbecued meat, vegetables, using non-dominant hand to pick up a bean with chopsticks; remembered getting an award and being praised at work, being afraid of high school tennis coaches, going on secret dates in high school
14	imagined singing, a woman dressed as a man singing, the shape of a train, potential names for a future child, a package of red apples, pictures of spouse, memories with spouse, cartoon characters, what they ate for lunch; focused on the induction grating; thought about what to make for dinner
15	imagined biking, the induction grating being a watery sphere, the induction grating being an archery target, the induction grating being green, scenes with blue skies and green lawns, friends' faces

Table A2. Examples of DecNef induction strategies. Participants were asked what strategies they employed during the DecNef task at the end of each neurofeedback session.

Supplementary Methods

Flicker fusion task

On the first day of the experiment a flicker fusion task (Simonson & Brozek, 1952) was used to determine perceptually equiluminant red and green RGB triplets. On each trial of the flicker

fusion task, a flickering circle (30 Hz, diameter ~13.5°) alternated between either red and neutral gray (rgb[128 128 128]) (block 1), green and neutral gray (block 2), or red and green (blocks 3 and 4). The screen background in this and all other tasks both inside and outside of the fMRI scanner was a uniform gray (rgb[64 64 64]). Participants were instructed to use button presses in order to minimize the amount of flicker they perceived in the stimulus as follows. On each trial, one of the two colors textures was used as a reference stimulus while the test stimulus, which was always either red or green, had the corresponding red or green channel of its RGB triplet shifted either up or down when participants pressed either the 'I' or 'K' key, respectively. Participants then pressed the 'Y' key to indicate that they had reached a point of minimal flicker. Non-variable RGB channel values in the test texture (e.g., the green and blue channels in the red test texture) were arbitrarily set to 80. On half of the trials the starting value of the variable channel was set to a random value between 0 and 19, while on the other half it was set to a random value between 236 and 255.

There were three practice trials for the flicker fusion task, after which participants completed 4 blocks of 12 trials each. In both the practice and 12-trial blocks, trials were separated by a 2-s intertrial interval (ITI), during which a uniform gray screen was shown. In the first two blocks the test textures were red and green, respectively, and reference textures were neutral gray. In the third block, the reference texture was red with an RGB triplet that corresponded to the mean of all of the 12 selected minimal flicker inducing red RGBs from block 1, while the test texture was green with the same stimulus parameters as the green test textures in block 2. In the fourth block, the reference texture was green with an RGB triplet that corresponded to the mean of all of the 12 selected minimal-flicker inducing green RGBs from block 2, while the test texture was red with the same stimulus parameters as the red test textures in block 1. For each subject, the

red and green RGB triplets used throughout the rest of the experiment were computed as the mean of all selected minimal flicker inducing RGBs from blocks 1 and 4 and blocks 2 and 3, respectively (mean ± s.e.m.: red = [218.6 80 80] ± [3.11 0 0], green = [80 149.1 80] ± [0 0.98 0]).

MVPA Sessions

For each task performed during the MVPA sessions, in the majority of trials (83.7% ± 0.7% of red/green discrimination trials and 83.6% ± 0.7% of lightness trials), hereafter described as threshold trials, stimulus strength was titrated via a run-by-run thresholding procedure in order to keep performance near 75% correct. This was intended to 1) facilitate a good spread of low to high confidence responses on the red/green discrimination task, and 2) keep participants engaged in the lightness task. The remaining trials either had relatively high stimulus strength (lightness range = 38.4 RGB units in the lightness task, 10.8% ± 0.9% of color trials; 80% of colored pixels for the red/green discrimination task; 10.7% ± 0.7% of confidence trials) or zero stimulus strength (no change in lightness in the lightness task, 5.6% ± 0.1% of color trials; zero colored pixels in the red/green discrimination task, 5.6% ± 0.1% of confidence trials). These high and zero stimulus strength trials were randomly interleaved across runs.

The difficulty of threshold trials in the color decoder task (color lightness) was modulated by changing the range of lightness values across which the colored pixels in the grating stimulus increased or decreased. For a given run, lightness range values were drawn from a uniform distribution, the range of which was arbitrarily set to 3.84 RGB units when the median of the distribution was greater than or equal to 5.12 RGB units, and 150% of the median when the median was less than 5.12 RGB units. The median of this distribution of lightness range values was adjusted per run (with the exception of the first run on Day 1) based on performance in the

preceding run according to the following rules. If the percent correct score on threshold trials in the preceding run was greater than or equal to 95%, between 80% and 95%, between 55% and 70%, or less than or equal to 55%, then the median of the current run's distribution of possible lightness values was scaled by 70%, 80%, 120%, or 130%, respectively.

For the first color decoder run on Day 1, the median of the distribution of possible lightness range values for threshold trials was set to the mean lightness range across all lightness task reversal trials from the Day 1 pre-decoder construction 1-up 1-down lightness task staircasing procedure (see above). The run-by-run thresholding procedure succeeded in maintaining group performance on threshold trials near perceptual threshold (mean ± s.e.m. percent correct = 73.7% ± 1.4% , d' = 1.53 ± 0.08).

The difficulty of the confidence decoder task (red/green discrimination) was modulated by changing the proportion of colored pixels in the grating stimulus. On the first confidence decoder run of Day 1, the proportion of colored pixels for a given trial was drawn from a uniform distribution, the minimum and maximum of which corresponded to mean Quest-estimated stimulus threshold from the Day 1 pre-decoder construction adaptive staircasing procedure multiplied by 1.2 and 1.6, respectively. The multipliers in this case were both greater than 1 to account for the observation from pilot subjects that the Quest procedure on Day 1 tended to underestimate the color stimulus strength that would lead to 75% correct accuracy on the red/green color discrimination task inside the scanner. On subsequent runs, the range of this distribution was scaled according to the same rules as those in the color decoder task (see above). Group performance on threshold trials in this task was also maintained near perceptual

threshold (mean ± s.e.m. percent correct = 74.3% ± 1.57%, d' = 1.67 ± 0.10, confidence ratings = 2.15 ± 0.14).

Because the confidence decoder was trained on confidence responses from a red/green discrimination task, one concern is that the confidence decoder may be confounded with color. Figure A2 shows that indeed confidence responses were higher on average for green stimuli compared to red stimuli in all but one DecNef participant. To investigate whether this bias, referred to hereafter as color-confidence bias, underlies the relationship between high confidence induction and false color perception during DecNef we performed a median split on DecNef study participants according to the difference in their mean confidence judgements for green versus red stimuli (Figure A2a). Replotting DecNef color induction likelihoods (Figure A4a) after median splitting suggested that DecNef color induction was not affected by color-confidence bias (Figure A2b).

To specifically test whether our main finding that high confidence induction likelihoods were higher during false alarm runs was affected by color-confidence bias we ran a mixed model ANOVA on high confidence induction likelihoods during DecNef with the within-subject factor run type (2 levels: false alarm runs, correct rejection runs) and the between-subjects factor color-confidence bias (2 levels: low, high). The ANOVA showed no main effect of either run type [$F(1,11) = 2.62$, $p = 0.13$] or color-confidence bias [$F(1,11) = 1.01$, $p = 0.34$] and no significant interaction [$F(1,11) = 0.07$, $p = 0.80$]. However, post-hoc Bonferroni corrected ($\alpha_{corrected} = 0.0125$) two-tailed one-sample t-tests for each of each combination of bias group and run type showed that high confidence induction likelihoods were significantly above chance (0.50) in the low color-confidence bias group during false alarm runs [$t(5) = 5.20$, $p < 0.01$, CI = (0.53,0.59)],

while high confidence induction likelihoods in each of the other three groups were not significantly different from chance [high bias, false alarm runs: $t(6) = 5.20$, $p = 0.34$, CI = (0.47, 0.58); low bias, correct rejection runs: $t(5) = 0.20$, $p = 0.85$, CI = (0.36, 0.66); high bias, correct rejection runs: $t(6) = -0.84$, $p = 0.43$, CI = (0.32, 0.59); Figure A2c]. Furthermore, there was a negative correlation between the extent of color-confidence bias and high confidence induction likelihoods during false alarm runs ($R^2 = 0.40$, $p = 0.02$; Figure A2d). Taken together, these results suggest that color-confidence bias did not underlie the high high confidence induction likelihoods observed during false alarms DecNef runs. Conversely, the median split analysis suggests that this effect was most strongly driven by the study participants who showed the smallest color-confidence bias.

fMRI localizer scans

In order to determine the subregions of V1, V2, V3 and V4 that retinotopically mapped to the grating stimuli in the color and confidence decoder tasks, during the second decoder construction session (Day 2) participants were presented with a flickering colored checkerboard localizer stimulus that occupied the same subregion of the visual field as those grating stimuli (0.425°-6.75° eccentricity). The localizer stimulus was presented alternately, in 8-s periods, with a second flickering colored checkerboard stimulus whose dimensions corresponded to the black annulus inside of the grating stimuli in the color and confidence decoder tasks (0.215°-0.425° eccentricity). Each stimulus was presented 14 times per run (224 s total), and each participant performed 3 runs. To ensure that participants maintained their gaze at the fixation point throughout each run, they performed a change detection task in which they pressed a button every time the fixation point changed color.

Optimization of color and confidence MVPA

Color and confidence decoders were constructed using sparse logistic regression as previously described (Amano et al., 2016; Cortese et al., 2016; Yamashita et al., 2008); see Methods in main text). To account for variability in hemodynamic delay (Buckner, 1998), for each region of interest (ROI), we trained separate decoders for each of several time windows for each type of decoder construction run (color or confidence). All decoding windows were shifted back in time to account for an assumed average hemodynamic delay of 6 seconds. In what follows, we indicate the time window according to the event to which the time window is supposed to correspond assuming this 6-s shift; e.g., when we indicate that a time window started at the target stimulus onset, that means that the first fMRI image that was analyzed in that time window was the one that was captured 6 seconds after induction cue onset.

For color runs, where the target stimulus was flashed over a period of 6 s, we trained decoders over nine different time windows: four 2-s windows starting at timepoints -2 s, 0 s, +2 s, and +4 s relative to target stimulus onset, three 6-s windows starting -2 s, 0 s, and +2 s relative to target stimulus onset, and two 8-s windows starting at -2 s and 0 s relative to target stimulus onset, where negative and positive numbers correspond to earlier and later time points, respectively. For confidence runs, which had a much briefer target stimulus display time (0.5 s) we trained decoders over three different time windows: two 2-s windows starting at timepoints -2 s and 0 s relative to target stimulus onset, and one 3-s windows starting -2 s relative to target stimulus onset. Further, for color runs, two decoders were trained over each time window, one in which the V1-4 ROI was intersected with the functional localizer ROI, and one in which it was not. This led to a total of 18 different localizer/time window combinations for color decoding.

For each decoder type for each participant, the time window-localizer combination for color decoding and time window for confidence decoding that resulted in the highest cross-validated accuracy was used for training the corresponding DecNef decoder on the entire dataset (summarized in Table A1). In all cases, the BOLD signal was averaged across all time points within a given time window, and the resulting averaged samples were used for the iSLR decoder training procedure.

For confidence decoding, to ensure an equal number of samples in the low and high confidence classes for decoder construction we used the following downsampling approach. Confidence ratings of 1 and 4 were always allocated to the low and high confidence training classes, respectively. The classes to which confidence ratings of 2 and 3 were allocated were determined so as to minimize the difference in the sample number between the two classes. Thus, confidence ratings could be divided into low and high confidence classes in the following three ways: low confidence = ratings of 1, high confidence = ratings of 2-4 (N = 9), low confidence = ratings of 1 and 2, high confidence = ratings of 3 and 4 (N = 6), and low confidence = ratings of 1-3, high confidence = ratings of 4 (N = 2). After assigning confidence ratings to their respective decoder classes, the class with the higher number of training samples was downsampled to equate the total number of samples between classes. The to-be-removed samples were chosen randomly, over four separate iterations. Cross-validated decoding accuracies were calculated for each downsampling iteration as described above, and the training samples that were used in the iteration that resulted in the highest decoding accuracy were subsequently used for training the DecNef decoder.

An additional confidence decoder was trained in the same manner using the visual ROI (V1-V4) that was used for color decoding. Confidence decoding accuracy in this ROI was significantly greater than chance by a one-sample, two-tailed t-test [mean s.e.m. decoding accuracy = 65.5 ± 1.8%; $t(16) = 8.61$, $p < 0.001$, CI = (61.7, 69.3)]. However, induction of high confidence patterns in this ROI did not correlate with actual confidence judgments during neurofeedback as was the case with the frontoparietal ROIs [$r(11) = 0.32$, $p = 0.29$; Figure A5, left].

Decoded Neurofeedback

For a description of the main components of the decoded neurofeedback (DecNef) procedure, see the Methods section in the main text. To ensure that the correct voxels were targeted during DecNef, we computed the correlation between the detrended, z-score normalized signal in each ROI during DecNef (see Methods in main text) and the mean detrended, z-score normalized signal in the corresponding ROIs across all decoder construction trials for each subject. Any trial that resulted in a correlation value less than $r = 0.6$ in any of the five target ROIs was excluded from the current analyses [median (interquartile range) = 1.3% (0.5% - 4.3%) of trials excluded per participant].

For all but two DecNef participants, a small proportion of DecNef trials [median (interquartile range) = 1.1% (0.4% - 1.8%) of total DecNef trials per participant (N=13)] motion-corrected BOLD signal data could not be retrieved in time for feedback stimulus presentation. On these trials a blank gray screen was presented during the feedback interval (Figure 10a). Participants were instructed beforehand that any such trials would be the result of computer malfunctions and should be ignored. These trials were omitted from all data analyses.

On achromatic trials, normally distributed gray RGB triplets for pixels between the black vertical bars in the induction stimulus were generated as described in the Red/Green Color Discrimination Task section under Materials and Methods in the main text. To generate red and green induction stimuli, the same procedure was followed, but a value of 12.8 was either subtracted from the green and blue channels or added to the green channel, respectively, of each dynamic voxel.

At the end of each DecNef session participants were asked what strategies, if any, they used during neurofeedback to try to maximize the size of the feedback circle. Strategies ranged from actively trying to think nothing to imagining the induction stimulus being the top of a barbecue, on which meats were being grilled (Table S2).

Appendix B: Peripheral 2IFC with null interval dot motion lifetime control

We tested whether the metacognitive inefficiencies observed in Experiments 3.2 and 3.3 might be due to attention to local motion signals by having two of the participants from both Experiments 3.2 and 3.3 repeat the task from Experiment 3.2 with identical stimulus parameters except for the following: dot lifetime was reduced from 400 ms to 80 ms and the total number of trials was decreased from 5,120 to 2,560. Reducing dot lifetime should make it harder to discriminate local motion signals. If this effect contributes to the flatness of the observed absolute blindsight curves (Figures 14b,c and 15b,c), then we should see a higher percentage of bets on the target present interval when dot lifetime is reduced.

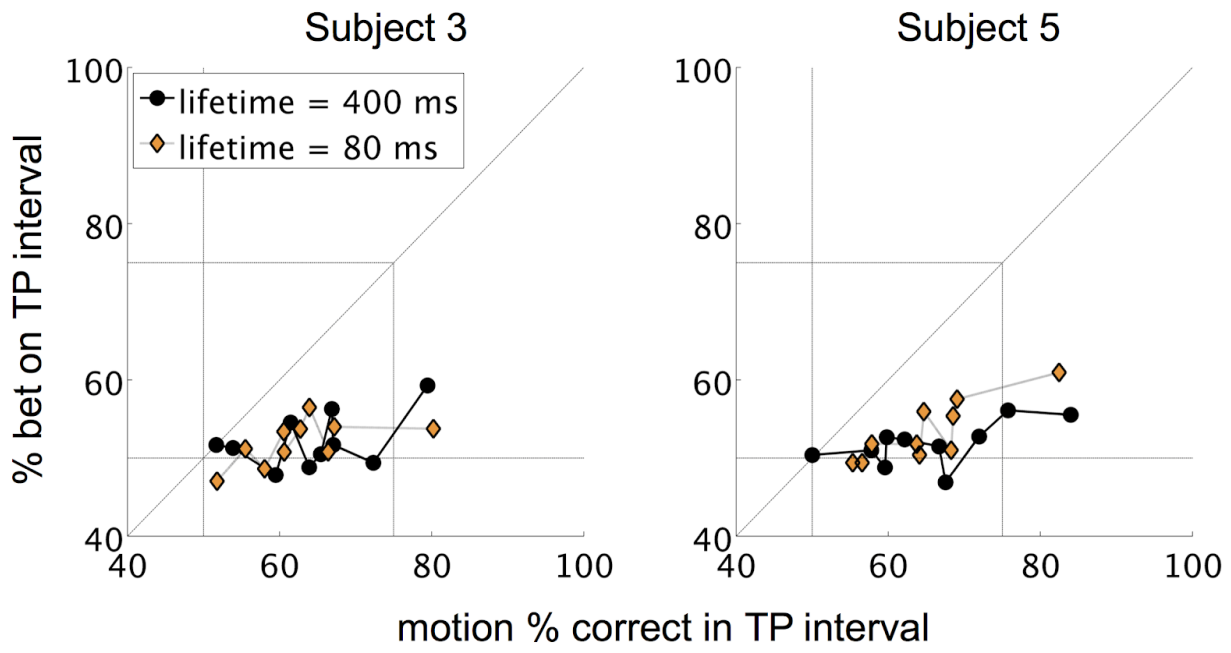

Figure B1. Peripheral 2IFC with null interval dot motion lifetime control results. The percentage of trials in which higher confidence was indicated for the target present interval is plotted as a function of motion discrimination percent correct scores in the target present interval (Figures 14b,c, 15b,c). All experimental conditions were identical to those in Experiment 3.2 except that the lifetime of the dot stimuli was reduced from 400 ms to 80 ms, and the total number of trials was reduced from 5,120 to 2,560.

Visual inspection of the absolute blindsight curves in Figure B1 suggests that reducing dot lifetime led to no clear improvements in Subject 3's ability to discriminate target absent and target present intervals, while it appears a slight improvement occurred for Subject 5. The current data are thus ambiguous, and more participants should be run on this task for a clearer picture to emerge. As such, we will have to wait to draw any conclusions about whether attention to local motion signals is responsible for the relatively high proportion of bets on the target absent interval in Experiments 3.2 and 3.3.

References

Achard, S., & Bullmore, E. (2007). Efficiency and Cost of Economical Brain Functional Networks. PLoS Computational Biology, 3(2), e17.

Almeida, J., Mahon, B. Z., & Caramazza, A. (2010). The Role of the Dorsal Visual Processing Stream in Tool Identification. Psychological Science, 21(6), 772–778.

Almeida, J., Mahon, B. Z., Nakayama, K., & Caramazza, A. (2008). Unconscious processing dissociates along categorical lines. Proceedings of the National Academy of Sciences of the United States of America, 105(39), 15214–15218.

Almeida, J., Pajtas, P. E., Mahon, B. Z., Nakayama, K., & Caramazza, A. (2013). Affect of the unconscious: Visually suppressed angry faces modulate our decisions. Cognitive, Affective & Behavioral Neuroscience, 13(1), 94–101.

Amano, K., Shibata, K., Kawato, M., Sasaki, Y., & Watanabe, T. (2016). Learning to Associate Orientation with Color in Early Visual Areas by Associative Decoded fMRI Neurofeedback. Current Biology: CB, 26(14), 1861–1866.

Azzopardi, P., & Cowey, A. (1993). Preferential representation of the fovea in the primary visual cortex. Nature, 361(6414), 719–721.

Balas, B., & Sinha, P. (2007). “Filling-in” colour in natural scenes. Visual Cognition, 15(7), 765–778.

Barthelmé, S., & Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system. PLoS Computational Biology, 5(9), e1000504.

Barthelmé, S., & Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. Proceedings of the National Academy of Sciences of the United States of America, 107(48), 20834–20839.

Blanke, O., Landis, T., & Seeck, M. (2000). Electrical cortical stimulation of the human prefrontal

cortex evokes complex visual hallucinations. Epilepsy and Behavior, 1(5), 356–361.

Block, N. (1995). On a confusion about a function of consciousness. The Behavioral and Brain Sciences, 18, 227–287.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. The Behavioral and Brain Sciences, 30(5-6), 481–499; discussion 499–548.

Block, N. (2011). Perceptual consciousness overflows cognitive access. Trends in Cognitive Sciences, 15(12), 567–575.

Block, N. (2014). Rich conscious perception outside focal attention. Trends in Cognitive Sciences, 18(9), 445–447.

Boly, M., Massimini, M., Tsuchiya, N., Postle, B. R., Koch, C., & Tononi, G. (2017). Are the Neural Correlates of Consciousness in the Front or in the Back of the Cerebral Cortex? Clinical and Neuroimaging Evidence. Journal of Neuroscience, 37(40), 9603–9613.

Braun, J., & Julesz, B. (1998). Withdrawing attention at little or no cost: detection and discrimination tasks. Perception & Psychophysics, 60(1), 1–23.

Breitmeyer, B. G. (2015). Psychophysical “blinding” methods reveal a functional hierarchy of unconscious visual processing. Consciousness and Cognition, 35, 234–250.

Breitmeyer, B. G., Koç, A., Öğmen, H., & Ziegler, R. (2008). Functional hierarchies of nonconscious visual processing. Vision Research, 48(14), 1509–1513.

Britten, K. H., & Heuer, H. W. (1999). Spatial summation in the receptive fields of MT neurons. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 19(12), 5074–5084.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 12(12), 4745–4765.

Bronfman, Z. Z., Brezis, N., Jacobson, H., & Usher, M. (2014). We See More Than We Can Report: “Cost Free” Color Phenomenality Outside Focal Attention. Psychological Science, 25(7), 1394–1403.

Buckner, R. L. (1998). Event-related fMRI and the hemodynamic response. Human Brain Mapping, 6(5-6), 373–377.

Carrasco, M. (2011). Visual attention: The past 25 years. Vision Research, 51(13), 1484–1525.

Chakravarthi, R., & Cavanagh, P. (2009). Recovery of a crowded object by masking the flankers: determining the locus of feature integration. Journal of Vision, 9(10), 4.1–9.

Cheesman, J., & Merikle, P. M. (1984). Priming with and without awareness. Perception & Psychophysics, 36(4), 387–395.

Cohen, M. A., Cavanagh, P., Chun, M. M., & Nakayama, K. (2012). The attentional requirements of consciousness. Trends in Cognitive Sciences, 16(8), 411–417.

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. Trends Cogn Sci., 15(8), 358–364.

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the Bandwidth of Perceptual Experience? Trends in Cognitive Sciences, 20(5), 324–335.

Cortese, A., Amano, K., Koizumi, A., Kawato, M., & Lau, H. (2016). Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. Nature Communications, 7, 1–18.

Curtis, C. E., & D’Esposito, M. (2004). The effects of prefrontal lesions on working memory performance and theory. Cognitive, Affective & Behavioral Neuroscience, 4(4), 528–539.

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

de Gardelle, V., & Mamassian, P. (2014). Does confidence use a common currency across two visual tasks? Psychological Science, 25(6), 1286–1288.

de Gardelle, V., Sackur, J., & Kouider, S. (2009). Perceptual illusions in brief visual presentations. Consciousness and Cognition, 18(3), 569–577.

Dehaene, S., Changeux, J.-P., Naccache, L., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends Cogn Sci., 10(5), 204–211.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? Science, 358(6362), 484–489.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. Cognition, 79(1-2), 1–37.

Dienes, Z. (2007). Subjective measures of unconscious knowledge. Progress in Brain Research, 168, 1–45.

Dienes, Z., & Seth, A. (2010). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. Consciousness and Cognition, 19(2), 674–681.

Dosher, B., & Lu, Z.-L. (2017). Visual Perceptual Learning and Models. Annual Review of Vision Science, 3(1), annurev – vision – 102016–061249.

Ehinger, B. V., Häusser, K., Ossandon, J., & König, P. (2017). Humans treat unreliable filled-in percepts as more real than veridical ones. eLife, 16(6). https://doi.org/10.7554/eLife.21761

Eriksen, C. W. (1960). Discrimination and learning without awareness: a methodological survey and evaluation. Psychological Review, 67, 279–300.

Faivre, N., Berthet, V., & Kouider, S. (2012). Nonconscious influences from emotional faces: A comparison of visual crowding, masking, and continuous flash suppression. Frontiers in Psychology, 3(MAY), 1–13.

Fazekas, P., & Overgaard, M. (2018). Perceptual consciousness and cognitive access: an introduction. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 373(1755). https://doi.org/10.1098/rstb.2017.0340

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., … Dale, A. M. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron, 33(3), 341–355.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. Frontiers in Human Neuroscience, 8(July), 1–9.

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. Brain, 137, 2811–2822.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. Nature Neuroscience, 14(9), 1195–1201.

Gau, R., & Noppeney, U. (2016). How prior expectations shape multisensory perception. NeuroImage, 124, 876–886.

Gaykema, R. P. A., Van Weeghel, R., Hersh, L. B., & Luiten, P. G. M. (1991). Prefrontal cortical projections to the cholinergic neurons in the basal forebrain. The Journal of Comparative Neurology, 303(4), 563–583.

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. Nature Reviews. Neuroscience, 14(5), 350–363.

Giles, N., Lau, H., & Odegaard, B. (2016). What Type of Awareness Does Binocular Rivalry Assess ? Developmental Topographical Disorientation. Trends in Cognitive Sciences, 20(10), 719–720.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. Annual Review of Neuroscience, 30, 535–574.

Goodale, M. A. (2011). Transforming vision into action. Vision Research, 51(13), 1567–1587.

Green DM, S. J. A. (1966). Signal detection theory and psychophysics. New York: Wiley.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. 455. Retrieved from https://psycnet.apa.org/fulltext/1967-02286-000.pdf

Gross, C. G. (2002). Genealogy of the “Grandmother Cell.” The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry, 8(5), 512–518.

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. From Perception to Consciousness. Retrieved from https://books.google.com/books?hl=en&lr=&id=Kw9pAgAAQBAJ&oi=fnd&pg=PA339&dq=Ariely,+Dakin+motion+perception&ots=a9clXm3ZDC&sig=6BXlV81ddX0b30dJ30nAb03eP7Q

Halpern, S. D., Andrews, T. J., & Purves, D. (1999). Interindividual variation in human visual performance. Journal of Cognitive Neuroscience, 11(5), 521–534.

Hannula, D. E., Simons, D. J., & Cohen, N. J. (2005). Imaging implicit perception: promise and pitfalls. Nature Reviews. Neuroscience, 6(3), 247–255.

Harman, G. H. (1965). The Inference to the Best Explanation. The Philosophical Review, 74(1), 88–95.

Harris, J. A., Wu, C.-T., & Woldorff, M. G. (2011). Sandwich masking eliminates both visual awareness of faces and face-specific brain activity through a feedforward mechanism. Journal of Vision, 11(7). https://doi.org/10.1167/11.7.3

Haun, A. M., Tononi, G., Koch, C., & Tsuchiya, N. (2018). Symposium on Haun, Tononi, Koch, and Tsuchiya:“ Are we underestimating the richness of visual experience?”: Response to Commentaries. Retrieved from

https://research.monash.edu/en/publications/symposium-on-haun-tononi-koch-and-tsuchiya-are-we-underestimating

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., … Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron, 72(2), 404–416.

Hebart, M. N., Schriever, Y., Donner, T. H., & Haynes, J.-D. (2014). The Relationship between Perceptual Decision Variables and Confidence in the Human Brain. Cerebral Cortex , (Dv), bhu181 – .

Hesselmann, G., Hebart, M., & Malach, R. (2011). Differential BOLD Activity Associated with Subjective and Objective Reports during “Blindsight” in Normal Observers. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 31(36), 12936–12944.

Izatt, G., Dubois, J., Faivre, N., & Koch, C. (2014). A direct comparison of unconscious face processing under masking and interocular suppression. Frontiers in Psychology, 5(JUL), 1–11.

Jiang, Y., Costello, P., & He, S. (2007). Processing of Invisible Stimuli. Psyc, 18(4), 349–355.

Jonas, E., & Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor? PLoS Computational Biology, 13(1), 1–24.

Kaplan, A. (1964). The conduct of scientific inquiry. Methodology for behavioral science. Scranton, PA: Chandler.

Kingdom, F. A. A., & Prins, N. (2010). Psychophysics: a practical introduction. Academic Press.

Knotts, J. D., Lau, H., & Peters, M. A. K. (2018). Continuous flash suppression and monocular pattern masking impact subjective awareness similarly. Attention, Perception & Psychophysics. https://doi.org/10.3758/s13414-018-1578-8

Knotts, J. D., Odegaard, B., Lau, H., & Rosenthal, D. (2018). Subjective inflation:

phenomenology's get-rich-quick scheme. Current Opinion in Psychology, 29, 49–55.

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. Nature Reviews Neuroscience, 17, 307–321.

Koch, C., & Tsuchiya, N. (2007). Attention and consciousness: two distinct brain processes. Trends in Cognitive Sciences, 11(1), 16–22.

Koizumi, A., Maniscalco, B., & Lau, H. (2015). Does perceptual confidence facilitate cognitive control? Attention, Perception, & Psychophysics, (March), 1295–1306.

Kok, P., Brouwer, G. J., van Gerven, M. A. J., & de Lange, F. P. (2013). Prior expectations bias sensory representations in visual cortex. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 33(41), 16275–16284.

Kok, P., Rahnev, D., Jehee, J. F. M., Lau, H. C., & De Lange, F. P. (2012). Attention reverses the effect of prediction in silencing sensory signals. Cerebral Cortex , 22(9), 2197–2206.

Kolb, F. C., & Braun, J. (1995). Blindsight in normal observers. Nature, 377, 336.

Komatsu, H. (2006). The neural mechanisms of perceptual filling-in. Nature Reviews. Neuroscience, 7(3), 220–231.

Kouider, S., & Dehaene, S. (2007). Levels of processing during non-conscious perception: a critical review of visual masking. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 362(1481), 857–875.

Kouider, S., & Dupoux, E. (2004). Partial awareness creates the "illusion" of subliminal semantic priming. Psychological Science, 15(2), 75–81.

Kouider, S., & Gardelle, V. D. (2010). How rich is consciousness? The partial awareness hypothesis. Trends in Cognitive Sciences, 14(7), 301–307.

Ko, Y., & Lau, H. (2012). A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. Philosophical Transactions of the Royal Society of

London. Series B, Biological Sciences, 367, 1401–1411.

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2018). Improving Inferences about Null Effects with Bayes Factors and Equivalence Tests. The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences. https://doi.org/10.1093/geronb/gby065

Lamme, V. A. F. (2003). Why visual attention and awareness are different. Trends Cogn Sci., 7(1), 12–18.

Lamme, V. A. F. (2010). How neuroscience will change our view on consciousness. Cognitive Neuroscience, 1(3), 204–220.

Lamy, D., Salti, M., & Bar-haim, Y. (2008). Neural Correlates of Subjective Awareness and Unconscious Processing : An ERP Study. 1435–1446.

Landau, A. N., & Fries, P. (2012). Report Attention Samples Stimuli Rhythmically. Current Biology: CB, 22(11), 1000–1004.

Landman, R., Spekreijse, H., & Lamme, V. A. F. (2003). Large capacity storage of integrated objects before change blindness. Vision Res., 43(2), 149–164.

Lau, H. (2008). Are we studying consciousness yet? Frontiers of Consciousness: Chichele Lectures. https://doi.org/10.1093/acprof:oso/9780199233151.003.0008

Lau, H., & Passingham, R. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. Proceedings of the National Academy of Sciences of the United States of America, 103(49), 18763–18768.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. Trends in Cognitive Sciences, 15(8), 365–373.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. Proceedings of the National Academy of Sciences of the United

States of America, 99(14), 9596–9601.

Li, M. K., Lau, H., & Odegaard, B. (2018). An investigation of detection biases in the unattended periphery during simulated driving. Attention, Perception & Psychophysics. https://doi.org/10.3758/s13414-018-1554-3

Linares, D., Aguilar-Lleyda, D., & López-Moliner, J. (2019). Decoupling sensory from decisional choice biases in perceptual decision making. eLife, 8. https://doi.org/10.7554/eLife.43994

Lloyd, D. A., Abrahamyan, A., & Harris, J. A. (2013). Brain-Stimulation Induced Blindsight : Unconscious Vision or Response Bias ? 8(12), 1–16.

Lu, J., & Itti, L. (2005). Perceptual consequences of feature-based attention. Journal of Vision, 5(7), 2–2.

Mack, A., & Rock, I. (1998). Inattentional blindness. MIT Press/Bradford Books Series in Cognitive Psychology. https://doi.org/10.1016/j.aorn.2010.03.011

Macknik, S. L., & Livingstone, M. S. (1998). Neuronal correlates of visibility and invisibility in the primate visual system. Nature Neuroscience, 1(2), 144–149.

MacMillan, N., & Creelman, C. D. (2004). Detection Theory: A User's Guide. Psychology Press.

Maniscalco, B., Peters, M. a. K., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. Attention, Perception & Psychophysics, 78(3), 923–937.

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. (V). https://doi.org/10.1038/nature12742

Matthews, J., Schröder, P., Kaunitz, L., van Boxtel, J. J. A., & Tsuchiya, N. (2018). Conscious access in the near absence of attention: critical extensions on the dual-task paradigm. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences,

373(1755). https://doi.org/10.1098/rstb.2017.0352

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. Perception & Psychophysics, 17(6), 578–586.

Merikle, P. M., Smilek, D., & Eastwood, J. D. (2001). Perception without awareness : perspectives from cognitive psychology. 79, 115–134.

Miconi, T., & VanRullen, R. (2016). A Feedback Model of Attention Explains the Diverse Effects of Attention on Neural Firing Rates and Receptive Field Structure. PLoS Computational Biology, 12(2), e1004770.

Milner, A. D., & Goodale, M. A. (2008). Two visual systems re-viewed. Neuropsychologia, 46(3), 774–785.

Morgan, M. J., Mason, A. J., & Solomon, J. A. (1997). Blindsight in Normal Subjects? Nature, 385, 401–402.

Naccache, L. (2018). Why and how access consciousness can account for phenomenal consciousness. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 373(1755). https://doi.org/10.1098/rstb.2017.0357

Nachmias, J., & Weber, A. (1975). Discrimination of simple and complex gratings. Vision Research, 15(2), 217–223.

Neisser, U., & Becklen, R. (1975). Selective looking: Attending to visually specified events. Cognitive Psychology, 7(4), 480–494.

Newsome, W. T., & Paré, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 8(6), 2201–2211.

Norman, E., & Price, M. C. (2015). Measuring consciousness with confidence ratings. In Behavioral Methods in Consciousness Research. Oxford: Oxford University Press.

Oblak, E. F., Sulzer, J. S., & Lewis-Peacock, J. A. (2018). A simulation-based approach to improve decoded neurofeedback performance (p. 450403). https://doi.org/10.1101/450403

Odegaard, B., Chang, M. Y., Lau, H., & Cheung, H. (2018). Inflation versus filling-in: why we feel we see more than we actually do in peripheral vision. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 373(1755). https://doi.org/10.1098/rstb.2017.0345

Odegaard, B., Grimaldi, P., Cho, S. H., Peters, M. A. K., Lau, H., & Basso, M. A. (2018). Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. Proceedings of the National Academy of Sciences of the United States of America, 115(7), E1588–E1597.

Odegaard, B., Knight, R. T., & Lau, H. (2017). Should a Few Null Findings Falsify Prefrontal Theories of Conscious Perception? The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 37(40), 9593–9602.

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. Current Opinion in Neurobiology, 14(4), 481–487.

Orne, M. T. (1969). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. American Psychologist, 17(11), 776–783.

Otten, M., Pinto, Y., Paffen, C. L. E., Seth, A. K., & Kanai, R. (2016). The Uniformity Illusion: Central Stimuli Can Determine Peripheral Perception. Psychological Science, 28(1), 56–68.

Overgaard, M. (2018). Phenomenal consciousness and cognitive access. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 373(1755). https://doi.org/10.1098/rstb.2017.0353

Overgaard, M., & Fazekas, P. (2016). Can No-Report Paradigms Extract True Correlates of

Consciousness? Trends in Cognitive Sciences, 20(4), 241–242.

Overgaard, M., Timmermans, B., Sandberg, K., & Cleeremans, A. (2010). Optimizing subjective measures of consciousness. Consciousness and Cognition, 19(2), 682–684; discussion 685–686.

Pal, D., & Mashour, G. (2018). Differential role of prefrontal and parietal cortices in controlling level of consciousness. Current Biology, 28(13), 2145–2152.

Panagiotaropoulos, T. I., Deco, G., Kapoor, V., & Logothetis, N. K. (2012). Neuronal discharges and gamma oscillations explicitly reflect visual consciousness in the lateral prefrontal cortex. Neuron, 74(5), 924–935.

Peremen, Z., & Lamy, D. (2014). Comparing unconscious processing during continuous flash suppression and meta-contrast masking just under the limen of consciousness. Frontiers in Psychology, 5(September), 969.

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. Nature Neuroscience, 10(2), 257–261.

Peters, M. A. K., Kentridge, R. W., Phillips, I., & Block, N. (2017). Does unconscious perception really exist? Continuing the ASSC20 debate. Neuroscience of Consciousness, 3(1), 1–11.

Peters, M. A. K., & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. eLife, 10.7554/eLife.09651.

Phillips, I. (2011). Perception and Iconic Memory: What Sperling Doesn't Show. Mind & Language, 26(4), 381–411.

Phillips, I. (2017). Unconscious perception reconsidered. Retrieved from http://www.ianbphillips.com/uploads/2/2/9/4/22946642/antwerp.pdf

Phillips, I. (2018). The methodological puzzle of phenomenal consciousness. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 373(1755).

https://doi.org/10.1098/rstb.2017.0347

Phillips, I., & Block, N. (2016). Block / Phillips Debate on Unconscious Perception. In B. Nanay (Ed.), Current controversies in philosophy of perception (pp. 165–192). Routledge.

Prins, N., & Kingdom, F. A. A. (2018). Applying the Model-Comparison Approach to Test Specific Research Hypotheses in Psychophysical Research Using the Palamedes Toolbox. Frontiers in Psychology, 9, 1250.

Quraishi, I. H., Benjamin, C. F., Spencer, D. D., Blumenfeld, H., & Alkawadri, R. (2017). Impairment of consciousness induced by bilateral electrical stimulation of the frontal convexity. Epilepsy and Behavior Case Reports, 8, 117–122.

Rahnev, D., Maniscalco, B., Graves, T., Huang, E., de Lange, F. P., & Lau, H. (2011). Attention induces conservative subjective biases in visual perception. Nature Neuroscience, 14(12), 1513–1515.

Rao, V., DeAngelis, G. C., & Snyder, L. H. (2012). Neural correlates of prior expectations of motion in the lateral intraparietal and middle temporal areas. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 32(29), 10063–10074.

Rausch, M., & Zehetleitner, M. (2016). Visibility Is Not Equivalent to Confidence in a Low Contrast Orientation Discrimination Task. Frontiers in Psychology, 7, 591.

Rosenthal, D. (2018). Consciousness and confidence. Neuropsychologia, (January). https://doi.org/10.1016/j.neuropsychologia.2018.01.018

Sàenz, M., Buraĉas, G. T., & Boynton, G. M. (2003). Global feature-based attention for motion and color. Vision Research, 43(6), 629–637.

Sahraie, A., Weiskrantz, L., & Barbur, J. L. (1998). Awareness and confidence ratings in motion perception without geniculo-striate projection. Behavioural Brain Research, 96(1-2), 71–77.

Sally, S. L., Vidnyánsky, Z., & Papathomas, T. V. (2009). Feature-based attentional modulation

increases with stimulus separation in divided-attention tasks. Spatial Vision, 22(6), 529–553.

Salti, M., Monto, S., Charles, L., & King, J.-R. (2015). Distinct cortical codes and temporal dynamics for conscious and unconscious percepts. https://doi.org/10.7554/eLife.05652

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: Is one measure better than the other? Consciousness and Cognition, 19(4), 1069–1078.

Scase, M. O., Braddick, O. J., & Raymond, J. E. (1996). What is Noise for the Motion System? Vision Research, 36(16), 2579–2586.

Schiff, N. D. (2010). Recovery of consciousness after brain injury: a mesocircuit hypothesis. Trends in Neurosciences, 33(1), 1–9.

Schlack, A., & Albright, T. D. (2007). Remembering visual motion: neural correlates of associative plasticity and motion recall in cortical area MT. Neuron, 53(6), 881–890.

Sergent, C. (2018). The offline stream of conscious representations. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 373(1755). https://doi.org/10.1098/rstb.2017.0349

Sergent, C., Wyart, V., Babo-rebelo, M., Cohen, L., Curie-paris, P. M., Recherche, C. D., & Inserm, U. (2013). Cueing Attention after the Stimulus Is Gone Can Retrospectively Trigger Conscious Perception. Current Biology: CB, 23(2), 150–155.

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

Simons, D. J. (2000). Attentional capture and inattentional blindness. Trends Cogn Sci., 4(4), 147–155.

Simons, D. J., & Rensink, R. A. (2005). Change blindness : past, present, and future. Trends Cogn Sci., 9(1), 16–20.

Simonson, E., & Brozek, J. (1952). Flicker fusion frequency; background and applications. Physiological Reviews, 32(3), 349–378.

Sivananda, R., Peters, M. A. K., Lau, H., & Odegaard, B. (2017). Subjective inflation of color saturation in the periphery under temporal overload. BioRxiv. https://doi.org/10.1101/227074

Sligte, I. G., Vandenbroucke, A. R. E., Scholte, H. S., & Lamme, V. A. F. (2010). Detailed sensory memory , sloppy working memory. Front Psychol, 1(175). https://doi.org/10.3389/fpsyg.2010.00175

Snodgrass, M., & Shevrin, H. (2006). Unconscious inhibition and facilitation at the objective detection threshold: replicable and qualitatively different unconscious perceptual effects. Cognition, 101(1), 43–79.

Solovey, G., Graney, G. G., & Lau, H. (2015). A decisional account of subjective inflation of visual perception at the periphery. Attention, Perception & Psychophysics, 77(1), 258–271.

Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. Consciousness and Cognition, 20(4), 1787–1792.

Sperling, G. (1960). The information available in brief visual presentations. Psychological Monographs: General and Applied, , 74, 1–29.

Sperling, G., & Dosher, B. A. (1986). Strategy and optimization in human information processing. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), Handbook of Perception and Performance (Vol. 1, pp. 2.1–2.65). Wiley-Interscience.

Stazicker, J. (2018). Partial report is the wrong paradigm. Philosophical Transactions of the

Royal Society of London. Series B, Biological Sciences, 373(1755). https://doi.org/10.1098/rstb.2017.0350

Stein, T., Hebart, M., & Sterzer, P. (2011). Breaking continuous flash suppression: A measure of unconscious processing during interocular suppression? Journal of Vision, 11(11), 315–315.

Strasburger, H., Rentschler, I., & Jüttner, M. (2011). Peripheral vision and pattern recognition: a review. Journal of Vision, 11(5), 13.

Suárez-Pinilla, M., Seth, A. K., & Roseboom, W. (2018). The Illusion of Uniformity Does Not Depend on the Primary Visual Cortex: Evidence From Sensory Adaptation. I-Perception, 9(5), 204166951880050.

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. (August), 403–409.

Szczepanowski, R., Traczyk, J., Wierzchoń, M., & Cleeremans, A. (2013). The perception of visual emotion: Comparing different measures of awareness. Consciousness and Cognition, 22(1), 212–220.

Tangen, J. M., Murphy, S. C., & Thompson, M. B. (2011). Flashed face distortion effect: grotesque faces from relative spaces. Perception, 40(5), 628–630.

Tapia, E., & Breitmeyer, B. G. (2011). Visual consciousness revisited: magnocellular and parvocellular contributions to conscious and nonconscious vision. Psychological Science, 22(7), 934–942.

Taschereau-Dumouchel, V., Cortese, A., Chiba, T., Knotts, J. D., Kawato, M., & Lau, H. (2018). Towards an unconscious neural reinforcement intervention for common fears. Proceedings of the National Academy of Sciences, 201721572.

Thomas, J. P., Gille, J., & Barker, R. A. (1982). and data. 72(12), 1642–1651.

Toscani, M., Gegenfurtner, K. R., & Valsecchi, M. (2017). Foveal to peripheral extrapolation of

brightness within objects. Journal of Vision, 17(9), 1–14.

Tsuchiya, N., & Koch, C. (2005). Continuous flash suppression reduces negative afterimages. Nature Neuroscience, 8(8), 1096–1101.

Tsuchiya, N., Wilke, M., Frässle, S., & Lamme, V. A. F. (2015). No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. Trends in Cognitive Sciences, 19(12), 757–770.

Usher, M., Bronfman, Z. Z., Talmor, S., Jacobson, H., & Eitam, B. (2018). Consciousness without report: insights from summary statistics and inattention “blindness.” Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 373(1755). https://doi.org/10.1098/rstb.2017.0354

Van Boxtel, J. J. A., Tsuchiya, N., & Koch, C. (2010). Consciousness and attention: On sufficiency and necessity. Frontiers in Psychology, 1(217). https://doi.org/10.3389/fpsyg.2010.00217

Vandenbroucke, A. R. E., Sligte, I. G., Barrett, A. B., Seth, A. K., Fahrenfort, J. J., & Lamme, V. A. F. (2014). Accurate Metacognition for Visual Sensory Memory Representations. Psychol Sci, 25(4), 861–873.

Vlassova, A., Donkin, C., & Pearson, J. (2014). Unconscious information changes decision accuracy but not confidence. Proceedings of the National Academy of Sciences of the United States of America, 111(45), 16214–16218.

Wallis, T. S. A., Funke, C. M., Alexander, S., Gatys, L. A., & Wichmann, F. A. (2018). Image content is more important than Bouma’s Law for scene metamers. BioRxiv. https://doi.org/378521

Wandell, B. a., & Winawer, J. (2011). Imaging retinotopic maps in the human brain. Vision Research, 51(7), 718–737.

Wang, L., Mruczek, R. E. B., Arcaro, M. J., & Kastner, S. (2015). Probabilistic maps of visual topography in human cortex. Cerebral Cortex , 25(10), 3911–3931.

Wang, M., Arteaga, D., & He, B. J. (2013). Brain mechanisms for simple perception and bistable perception. Proceedings of the National Academy of Sciences of the United States of America, 110(35), E3350–E3359.

Ward, E. J. (2018). Downgraded phenomenology: how conscious overflow lost its richness. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 373(1755). https://doi.org/10.1098/rstb.2017.0355

Ward, E. J., Bear, A., & Scholl, B. J. (2016). Can you perceive ensembles without perceiving individuals?: The role of statistical perception in determining whether awareness overflows access. Cognition, 152, 78–86.

Watamaniuk, S. N., Sekuler, R., & Williams, D. W. (1989). Direction perception in complex dynamic displays: the integration of direction information. Vision Research, 29(1), 47–59.

Watson, A. B., & Pelli, D. G. (1983). Quest: A Bayesian adaptive psychometric method. Perception & Psychophysics, 33(2), 113–120.

Watson, A. B., & Robson, J. G. (1981). Discrimination at threshold: labelled detectors in human vision. Vision Research, 21(7), 1115–1122.

Weiskrantz, L. (1986). Blindsight. A case study and implications. Oxford, U.K.: Oxford University Press.

Weiskrantz, L. (1999). Consciousness Lost and Found. Oxford: Oxford University Press.

Weiskrantz, L., Warrington, E. K., Sanders, M. D., & Marshall, J. (1974). Visual capacity in the hemianopic field following a restricted occipital ablation. Brain: A Journal of Neurology, 97(4), 709–728.

White, A. L., & Carrasco, M. (2011). Feature-based attention involuntarily and simultaneously

improves visual performance across locations. Journal of Vision, 11(6). https://doi.org/10.1167/11.6.15

Wierzchoń, M., Asanowicz, D., Paulewicz, B., & Cleeremans, A. (2012). Subjective measures of consciousness in artificial grammar learning task. Consciousness and Cognition, 21(3), 1141–1153.

Wijntjes, M. W. A., & Rosenholtz, R. (2018). Context mitigates crowding: Peripheral object recognition in real-world images. Cognition, 180, 158–164.

Williams, D. W., & Sekuler, R. (1984). Coherent global motion percepts from stochastic local motions. Vision Research, 24(1), 55–62.

Witt, J. K., Sugovic, M., & Wixted, J. T. (2012). Signal detection measures cannot distinguish perceptual biases from response biases 1. 1–13.

Witt, J. K., Taylor, J. E. T., Sugovic, M., & Wixted, J. T. (2015). Signal detection measures cannot distinguish perceptual biases from response biases. Perception, 44(3), 289–300.

Yamashita, O., Sato, M. A., Yoshioka, T., Tong, F., & Kamitani, Y. (2008). Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. NeuroImage, 42(4), 1414–1429.

Zaborszky, L., Gaykema, R. P., Swanson, D. J., & Cullinan, W. E. (1997). Cortical input to the basal forebrain. Neuroscience, 79(4), 1051–1078.

Zhang, R., Kwon, O.-S., & Tadin, D. (2013). Illusory movement of stationary stimuli in the visual periphery: evidence for a strong centrifugal prior in motion processing. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 33(10), 4415–4423.

Zinchenko, A., Conci, M., Müller, H. J., & Geyer, T. (2018). Predictive visual search: Role of environmental regularities in the learning of context cues. Attention, Perception & Psychophysics, 80(5), 1096–1109.