Modulation of perceptual brain processes through learning and attention

vorgelegt von

Dipl.-Phys., MSc
Matthias Guggenmos
geb. in München

von der Fakultät IV − Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
− Dr. rer. nat −

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Henning Sprekeler
Gutachter:	Prof. Dr. Klaus Obermayer
	Prof. Dr. Philipp Sterzer
	Prof. Dr. Stefan Pollmann

Tag der wissenschaftlichen Aussprache: 21. September 2015

Berlin 2015

Acknowledgements

Above all, I would like to thank my doctoral supervisor Philipp Sterzer for the opportunity to work on a variety of interesting research project, for his competent advice, creativity and encouragement. Under his supervision and through his expertise, I became acquainted and fond of the field of cognitive neuroscience. I am also highly grateful to my two co-supervisors, John-Dylan Haynes and Klaus Obermayer, who guided me throughout my PhD with helpful advice and fruitful discussions. Many thanks go to Stephan Pollmann, who kindly agreed to be the external reviewer for this thesis. I was lucky to be in a lab with an exceptionable group of people. I am particularly grateful to Gregor Wilbertz and Kiley Seymour for a personally and scientifically stimulating atmosphere in and outside our office, and to Marcus Rothkirch, who among many other things, introduced me to fMRI data collection and analysis. Many thanks go to Bianca van Kemenade, Katharina Schmack, Apoorva Rajiv Madipakkam, Karin Ludwig, James Kerr, Guido Hesselmann, Julia Lichte and Shea Karst, for thesis proofreading, many helpful discussions and a great time. I was lucky to have many great collaborators within the projects of this thesis, in particular Volker Thoma, Martin Hebart, Radoslaw Martin Cichy, and Alan Richardson-Klavehn, who took leading roles within these projects, and from whom I learned a lot about science. I thank the Bernstein Center of Computational Neuroscience in Berlin and the DFG for financial support, and Vanessa Cassagrande and Robert Martin for a highly professional—and personal—management of the BCCN PhD program. I also want to highlight that both the BCCN and the DFG were exemplary in their support for “parents in science”. Finally, and most importantly, I am indebted to Lena Janitzki and our daughter Cala Janitzki, for creating science-free spaces of cheerfulness, and to my parents, for their unconditional support.

Research Articles

This cumulative thesis is based on four studies, which are summarized in five research articles:

Study 1

Guggenmos M, Rothkirch M, Obermayer K, Haynes J-D, Sterzer P (2015). A hippocampal signature of perceptual learning in object recognition. *Journal of Cognitive Neuroscience* 27, 787–797.

Study 2

Guggenmos M, Thoma V, Cichy RM, Haynes JD, Sterzer P*, Richardson-Klavehn A* (2015). Non-holistic coding of objects in lateral occipital complex with and without attention. *NeuroImage* 107, 356–363.

Guggenmos M, Thoma V, Haynes J-D, Richardson-Klavehn A, Cichy RM*, Sterzer P* (2015). Spatial attention enhances object coding in local and distributed representations of the lateral occipital complex. *NeuroImage* 116, 149–157.

Study 3

Ostwald D, Spitzer B, Guggenmos M, Schmidt TT, Kiebel SJ, Blankenburg F (2012). Evidence for neural encoding of Bayesian surprise in human somatosensation. *NeuroImage* 62, 177–188.

Study 4

Guggenmos M, Wilbertz G, Hebart M*, Sterzer P*. Mesolimbic confidence signals guide perceptual learning in the absence of external feedback.

*contributed equally

Statement of contribution

I was the primary researcher in studies 1, 2 and 4, and as such responsible for all major aspects of these studies, including experimental design, implementation, data collection, analysis and writing. In study 3 I contributed to the design, was responsible for the implementation of the experiment, and for the collection and initial analysis of pilot data.

Table of contents

Abstract

To cope with the challenges and affordances in a dynamic environment, the brain needs mechanisms to adapt and improve perception. The present thesis addresses two such mechanisms—perceptual learning and attention. Whereas perceptual learning is known to be a slow process leading to stable and long-lasting changes in perception, attention affects perception instantaneously and transiently. The aim of this thesis is to contribute to a better understanding of the neural basis of these two mechanisms.

The first two studies investigated neural mechanisms underlying improvements in object perception, both through perceptual learning (study 1) and attention (study 2). Study 1 used functional magnetic resonance imaging (fMRI) to investigate the neural basis of object recognition learning under reduced sensory viewing conditions. The key result was that neural responses in the hippocampus to trained object categories were enhanced by perceptual learning. This finding is in line with a recently suggested hippocampal pattern completion signal, which may aid perception by generating more complete percepts from reduced sensory data. Study 2 used fMRI to examine quantitative and qualitative effects of attention on object representations in high-level visual cortex. The results demonstrated that attention led to the amplification of object-related neural activity, which resulted in object representations that were more informative about object identity and more reproducible across presentations. By contrast, the results did not yield evidence for qualitatively different neural representations of objects with and without attention.

The third and fourth study investigated the role of self-generated neural feedback signals in perceptual learning. Study 3 employed a somatosensory mismatch roving paradigm and electroencephalography (EEG) to search for neural evidence of "Bayesian perceptual learning", which refers to the process of updating an internal model of sensory input. The results showed that EEG responses in somatosensory and cingulate cortex were consistent with Bayesian surprise, a marker of Bayesian perceptual learning. Finally, study 4 used fMRI to investigate how the brain improves perception in the absence of external feedback. A correlation of brain activity with predictions of a novel learning model, which utilized internal confidence-based feedback, provided evidence that learning without feedback may be guided by some of the same brain structures—ventral striatum and ventral tegmental area—implicated in learning with external feedback.

Together, the results of this thesis provide novel neural evidence for an important role of self-generated modulatory and feedback signals underlying perceptual improvements: (i) a putative pattern completion signal in the hippocampus, (ii) top-down attentional modulation in high-level visual areas, (iii) a sensory surprise signal reflecting the update of internal stimulus models in somatosensory and cingulate cortex, and (iv) a confidence-based feedback signal in the ventral striatum and the ventral tegmental area.

Zusammenfassung

Um sich den Gegebenheiten einer dynamischen Umwelt anzupassen benötigt das Gehirn Mechanismen, die eigene Wahrnehmung zu verändern und zu verbessern. Die vorliegende Dissertation untersucht zwei solcher Mechanismen – perzeptuelles Lernen und Aufmerksamkeit. Während perzeptuelles Lernen ein langwieriger Prozess ist, der zu langanhaltenden Wahrnehmungsveränderungen führt, beeinflusst Aufmerksamkeit die Wahrnehmung instantan und vorrübergehend. Das Ziel dieser Dissertation liegt in einer Erweiterung des Verständnisses der neuronalen Grundlagen dieser beiden Mechanismen.

Die ersten beiden Studien befassten sich mit neuronalen Mechanismen zur Unterstützung von Objektwahrnehmung, sowohl durch perzeptuelles Lernen (Studie 1), als auch durch Aufmerksamkeit (Studie 2). In Studie 1 wurde mithilfe von funktioneller Magnetresonanztomographie (fMRT) die neuronale Basis von Objekterkennung unter erschwerten Sichtbarkeitsbedingungen untersucht. Als zentraler Befund zeigte sich, dass perzeptuelles Lernen von Objekterkennung zu einer verstärkten Antwort des Hippocampus auf trainierte Objektkategorien führte. Dieser Befund wies Übereinstimmung mit einem kürzlich postulierten Mustervervollständigungssignal im Hippocampus auf, welches Grundlage der Wahrnehmungsverbesserung sein könnte. In Studie 2 wurden mithilfe von fMRT die Effekte von Aufmerksamkeit auf Objektrepräsentationen im höheren visuellen Kortex untersucht. Es konnte nachgewiesen werden, dass Aufmerksamkeit zu einer verbesserten Enkodierung von Objekten führte. Dabei waren neuronale Objektrepräsentationen durch Aufmerksamkeit nicht nur informativer bezüglich der präsentierten Objekte, sondern auch besser reproduzierbar bei wiederholten Objektpräsentationen. Dagegen zeigten sich keine qualitativen Unterschiede hinsichtlich des Formats der Objektkodierung in Abhängigkeit von Aufmerksamkeit.

Die dritte und vierte Studie befassten sich mit der Rolle von selbsterzeugten neuronalen Feedbacksignalen bei perzeptuellem Lernen. In Studie 3 wurde mithilfe eines somatosensorischen Mismatchparadigmas und Elektroenzephalographie (EEG) nach neuronaler Evidenz für „Bayesianisches perzeptuelles Lernen“ gesucht, demzufolge perzeptuelles Lernen der Verbesserung eines internen Modells sensorischer Signale entspricht. Tatsächlich zeigte sich im somatosensorischen und cingulären Kortex eine Übereinstimmung von EEG-Signalen mit Bayesian Surprise, einem Marker von Bayesianischem perzeptuellem Lernen. In Studie 4 wurde schließlich mithilfe von fMRT untersucht, auf welche Weise das Gehirn perzeptuelle Sensitivität ohne externes Feedback verbessert. Eine Korrelation von Gehirnsignalen mit Vorhersagen eines neuartigen Lernmodells, beruhend auf internem konfidenzbasiertem Feedback, zeigte eine Beteiligung von Gehirnstrukturen (ventrales Striatum und ventrales Tegmentum), die zuvor primär mit externem Feedback in Verbindung gebracht wurden.

Zusammenfassend konstatieren die Ergebnisse dieser Dissertation neuronale Evidenz für eine wichtige Rolle von selbsterzeugten Modulations- und Feedbacksignalen bei Wahrnehmungsverbesserungen: (i) ein Mustervervollständigungssignal im Hippocampus, (ii) top-down Modulation durch Aufmerksamkeit im höheren visuellen Kortex, (iii) ein Bayesian-Surprise-Signal im somatosensorischen und cingulären Kortex, und (iv) ein konfidenzbasiertes Feedbacksignal im ventralen Striatum und ventralen Tegmentum.

1. INTRODUCTION

> "And so is any difference [..] more easily perceived when we carry in our mind to meet it a distinct image of what sort of a thing we are to look for." (James, 1890)

In the above excerpt, William James reflects upon some of the earliest studies in perceptual learning by Volkmann (1858), who investigated learning of tactile two-point discrimination. James proposed that an unobscured internal template of a stimulus may aid other, perceptually more challenging encounters with the stimulus. Today, the question of how the brain adapts and improves perception has attracted widespread interest in psychology and neuroscience, whereby such internal models of sensory input also have become one focal point. In the present thesis we investigated the neural basis of two mechanisms that are known to improve perception—the aforementioned *perceptual learning* and *attention*.

Perceptual learning is a form of implicit learning and describes our ability to change perception through training or repeated exposure. Classic examples of perceptual learning are expert professions such as radiologists, sommeliers and ornithologists, their mastery requiring years of perceptual training. Although laboratory examples of perceptual learning have been the subject of a considerable number of neuroscientific investigations, their neural mechanisms remain highly debated. One central controversy revolves around the question whether perceptual learning is based on a sharpening of sensory representations, or whether it occurs in higher brain areas, as a selective reweighting of otherwise unchanged sensory information (for review, see Watanabe and Sasaki, 2014). In addition, recent empirical studies and theories point to an important role of plastic internal models of sensory input, which may aid perception in situations of weak sensory evidence (for review, see Seriès and Seitz, 2013). In the present thesis, we conducted three functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) studies to investigate the neural mechanisms of perceptual learning in the context of these theories. In **study 1** we were interested in the neural correlates of perceptual learning in object recognition and probed the brain's response to object stimuli before and after extensive perceptual training. Two more studies were concerned with the role of self-generated neural feedback signals in perceptual learning. More specifically, **study 3** searched for neural evidence of signals associated with internal sensory model optimization, as introduced above, and **study 4** tested the hypothesis that perceptual learning without external feedback may be guided by a self-generated neural feedback signal based on perceptual confidence.

Attention is a cognitive process that enables us to focus on certain aspects of the environment for the benefit of improved perceptual performance (for review, see Carrasco, 2011) and speed (Posner, 1980). This beneficial effect of attention is thought to involve a temporary modulation of sensory activity, as opposed to long-term plastic changes. One key mechanism of such attentional modulation is the amplification of neural activity. Indeed, in the visual domain attentional amplification has been found throughout the visual processing hierarchy, from the earliest stage of visual neural processing in the lateral geniculate nucleus, primary visual cortex, up to high-level visual cortices and the frontal lobes (for review, see Treue, 2003). **Study 3** of this thesis investigated the neural basis of visual attention in object perception and pursued two separate research questions. First, we tested a specific model by Hummel (2001) which postulates a different representational formats for object representations with and without attention. Second, we investigated whether attentional modulation involved the amplification of neural responses or rather a mere baseline shift. Third and finally, we examined whether attentional gain enhanced the quality of object representations in high-level visual cortex, which may provide the basis for improved perceptual performance.

A unifying framework of various aspects of perception, including perceptual learning and attention, is provided by the predictive coding principle (Rao and Ballard, 1999) and its extensions (free energy principle: Friston, 2005; generalized predictive coding: Feldman and Friston, 2010). According to this framework, the brain maintains internal models of the world, which generate predictions about incoming sensory information at various stages of the sensory processing hierarchy. The (mis)match of predictions and observations elicits prediction error signals, which serve as update signals for internal models in order to improve future predictions. Within this framework, perceptual learning has been identified with the process of improving internal sensory models based on prediction error signals (Friston, 2009, 2005), and attention as a weighting of these prediction errors (Feldman and Friston, 2010). Although the predictive coding framework and its ensuing predictions were explicitly tested only in **study 3**, it guided the interpretation of important results in other studies. Towards the end of this thesis, we therefore attempt an integration of these results within the framework.

This thesis is organized as follows. Chapter 2 introduces the empirical and theoretical background that motivated the research, as well as two advanced methods of neuro-imaging data analysis that were keys to the evaluation of the studies. Chapter 3 provides a summary of all four studies, and chapter 4 discusses the implications of our results in a broader context and draws conclusions across studies. The publications and manuscripts associated with these studies are appended at the end of this thesis (chapter 6)

2. THEORETICAL AND EMPIRICAL BACKGROUND

2.1 Improving perception through perceptual learning

2.1.1 Definition and behavioral characteristics

Perceptual learning is a form of implicit learning that comprises "any relatively permanent and consistent change in the perception of a stimulus array, following practice or experience with this array" (Gibson, 1963). In experimental settings these perceptual changes are often reflected—and measured—as increased sensitivity in perceptual tasks (Tsodyks and Gilbert, 2004). Such improvements have been found across many perceptual tasks, from simple low-level discriminations (e.g., distinguishing the pitch of two musical tones, Burns and Ward, 1978) to complex categorizations in real-world scenarios (e.g., distinguishing subtle flavors in wine tasting, Walk, 1966), and for all five primary senses (Fahle and Poggio, 2002). Ensuing perceptual changes are relatively stable and can often be measured years after training (Karni and Sagi, 1993; Hussain et al., 2011), which is also a characteristic difference to short-term perceptual changes through allocation of attention or adaptation.

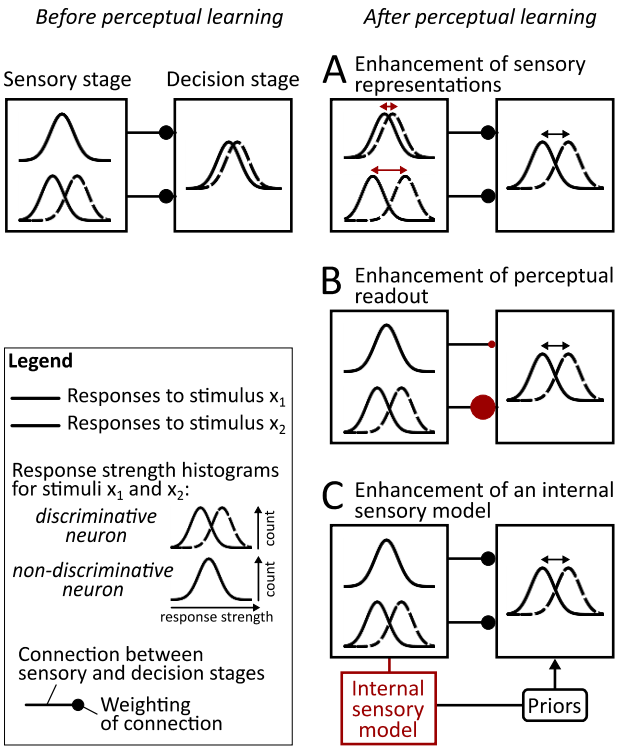
A hallmark of perceptual learning is the specificity of behavioral improvements to certain stimulus properties. Visual perceptual learning in particular has been reported to be specific for visual field location, eye and for many low-level features, including orientation, spatial frequency and motion direction (for review, see Fahle, 2005). While the specificity of such low-level perceptual learning has been fairly extensively studied, much less is known about the specificity of perceptual learning of more complex stimuli, such as objects. Paradigms of high-level visual perceptual learning typically involve the detection or identification of complex stimuli under challenging viewing conditions (Dolan et al., 1997; Furmanski and Engel, 2000; Grill-Spector et al., 2000; Schwiedrzik et al., 2009, 2011; Albrecht et al., 2010; Baeck and Op De Beeck, 2010). Here, the specificity of perceptual learning cannot easily be tested by variation along a single stimulus dimension. Instead, these previous studies presented sets of other, untrained stimuli after training. In doing so, they consistently found a *partial transfer* of learning; that is, the performance level for untrained stimuli was lower than for trained stimuli, but higher than before training. In addition, Furmanski and Engel (2000) showed that this partial transfer was identical for

different sizes of untrained stimuli, which suggests that learning was specific to high-level size-invariant features. However, these findings leave open the question which stimulus features of complex stimuli generalize and which not. To explore this question, **study 1** examined whether perceptual learning in object recognition involves a *category-level component*, that is, whether training on one set of exemplars of a category transfers to other (untrained) exemplars of the same category.

Another prominent characteristic of perceptual learning is that it occurs robustly in the absence of external feedback (McKee and Westheimer, 1978; Ball and Sekuler, 1987; Karni and Sagi, 1991; Shiu and Pashler, 1992; Fahle et al., 1995; Seitz and Watanabe, 2003; Seitz et al., 2009; but see Herzog and Fahle, 1997)—in contrast to many other types of learning, which strongly depend on cognitive or reward-based feedback (Goldstein and Rittenhouse, 1954; Goodman and Wood, 2004; Pavlov et al., 1928; Schultz, 2000). In apparent discrepancy to these behavioral observations, current perceptual learning models are fundamentally based on external feedback (Dosher et al., 2013; Petrov et al., 2005; Kahnt et al., 2011; Law and Gold, 2009) and therefore cannot account for behavioral improvements without feedback. A single notable exception is a model by Petrov et al. (2006), which assumes that observers acquire and exploit knowledge about the task structure, e.g., by learning about the frequency of different stimuli. However, such acquisition of explicit task knowledge can likely account only for a small fraction of perceptual learning and only in specific scenarios. One key aspect of **study 4** is therefore to fill this gap by devising a perceptual learning model guided by *internal*, confidence-based feedback.

2.1.2 Leading theories and neural mechanisms

A focal point of the debate around the neurobiological basis of perceptual learning is the question of whether perceptual learning happens in the encoding sensory cortices (“early selection”; Fahle, 2004), or further upstream as a selective reweighting of otherwise unchanged sensory information. In the following, the key ideas underlying these theories will be introduced. In addition, a third theory will be introduced, which proposes that perceptual learning reflects an enhancement of internal sensory models with the aim of improving perception through prior information. Figure 1 provides a conceptual overview of these three theories.



Please note that the neural response distributions do not represent tuning curves, but separate neural response strength histograms for stimuli x_1 and x_2.

Figure 1. Leading theories of perceptual learning.
Before perceptual learning: each model consists of a sensory stage with two exemplary stimulus-encoding neurons (one discriminative, and one non-discriminative for two arbitrary stimuli x_1, x_2), a decision stage with one choice-encoding neuron and connections between the two stages (lines). The weighting of each connection at the decision stage is indicated by the radius of an attached circle. After perceptual learning: **A** Enhancement of sensory representations: perceptual learning results from a greater separation of the stimulus-encoding neurons' response distributions at the sensory stage. **B** Enhancement of perceptual readout: perceptual learning results from a differential weighting of discriminative and non-discriminative sensory signals. **C** Enhancement of an internal sensory model: improved performance results from the integration of sensory data with prior information provided by an internal stimulus model. In this context, perceptual learning is the process of enhancing the internal model through repeated experience with the stimuli.

Enhancement of sensory representations

The way in which the sensory representation of a stimulus can be optimized is inseparably linked to the nature of the neural code. According to the most common view, neurons are tuned to a particular feature, whereby the response strength, or firing rate, indicates the neuron's degree of preference. In this view, stimulus representations can be enhanced in a number of ways: (1) by sharpening neuronal tuning curves, such that single neuronal responses more precisely signal the presence or absence of a particular feature; (2) by shifting neuronal tuning curves to achieve maximal sensitivity with respect to a particular range of a feature dimension; (3) by recruiting previously unresponsive neurons to increase the signal-to-noise ratio at the population level (*cortical recruitment*); and (4) by an increase of stimulus-driven firing rates which likewise leads to an improved signal-to-noise ratio (for review, see Op de Beeck et al., 2010).

In search of neuronal changes of this kind, a number of neurophysiological primate studies have characterized neuronal responses in sensory areas before and after perceptual

learning. However, although the perceptual tasks were often similar, the studies yielded inconsistent results. For instance, Yang and Maunsell (2004) found learning-related increases in stimulus-related response strength, whereas Ghose et al. (2002) found decreases; Schoups et al. (2001), Raiguel et al. (2006), and Yang and Maunsell (2004) found evidence for an optimization of tuning functions, whereas Ghose et al. (2002) and Crist et al. (2001) found none of the kind; and finally, Hua et al. (2010) and Zohary et al. (1994) provided evidence for increased contrast sensitivity, whereas Law and Gold (2008) reported the absence of any gain in contrast sensitivity of sensory neurons, despite a strong behavioral effect. The sources for these inconsistent results are unclear at present, but may be related to some of the disadvantages of primate neurophysiological studies, such as small sample sizes (often N=2), non-representative selection of neurons, or the lack of control for a top-down modulation of neural responses (e.g., attentional modulation). Several human neuroimaging studies, too, have investigated learning-related changes in sensory brain areas—with equally mixed evidence. A few studies found an increase of activation in sensory areas (Schwartz et al., 2002; Grill-Spector et al., 2000; Pleger et al., 2003; Dolan et al., 1997; Furmanski et al., 2004), whereas other studies found both increases and decreases (Op de Beeck et al., 2006; Kourtzi et al., 2005), consistent decreases (Schiltz et al., 1999; Mukai et al., 2007) or no change at all (Kahnt et al., 2011). Here, a possible source of discrepant results could be the duration of training. Yotsumoto (2008) found a stimulus-related increase of primary visual cortex (V1) activity in the initial phase of a low-level perceptual learning experiment, but a subsequent decrease with more training days. In addition, it was found that the strength of stimulus-related neural activity has a complicated relationship with the number of repeated presentations of a stimulus, irrespective of any learning involved (Müller et al., 2013). Thus, the duration of training and the number of repeated presentations of the same stimulus may influence the level of neural activity after training.

Going beyond the analysis of changes in overall response amplitude, recent fMRI studies have applied multivariate information-based methods (see section 2.4.2, below) to quantify the amount of stimulus-related information—conveyed by activation patterns in sensory cortices—as a function of learning (Kahnt et al., 2011; Shibata et al., 2012). However, the results of two studies investigating perceptual learning by means of multivoxel pattern analysis are again controversial: Shibata et al. (2012)—using random dot kinematograms in a motion detection task and focusing on motion-sensitive brain areas—found a sharpening of an information-based motion tuning function in motion-

sensitive areas. By contrast, Kahnt et al. (2011)—using Gabor patches in an orientation discrimination task and looking at V1 responses—found no such sharpening or other changes in information content. Overall, the plasticity of sensory areas in perceptual learning therefore remains highly debated. In the present thesis, we analyzed brain activity in functionally relevant sensory areas with respect to changes in both response amplitude and informativeness at the pattern level (**studies 1, 2 and 4**).

Enhancement of perceptual readout

Readout models propose that perceptual learning does not result from plastic sensory representations, but rather from a selective reweighting of otherwise unchanged sensory information at a decisional stage. The most common type of readout model is based on the idea of the Perceptron (Rosenblatt, 1957). The Perceptron takes inputs $\vec{x}$ and applies a binary classifier $\vec{w} \cdot \vec{x} + b$. The vector $\vec{w}$ and the bias b define angle and intercept of a decisional hyperplane in the N-dimensional space of inputs $\vec{x}$, so that new samples are classified according to which side of the hyperplane they fall. In the context of perception and perceptual learning, the vector $\vec{x}$ represents stimulus information signaled by a set of neural sensory units with different tuning functions (often referred to as sensory *channels*; Graham, 1989). The weights $\vec{w}$ determine how much the final decision is influenced by each sensory channel. Perceptual learning then corresponds to the optimization of these weights to achieve improved classification accuracy. To implement perceptual learning, weight changes may be governed by Hebbian learning, such that weight changes are proportional to the correlation of sensory channel activity and output signal (Petrov et al., 2005; Lu et al., 2010; Dosher et al., 2013), or by reinforcement learning, such that weight changes are proportional to reward prediction errors (Kahnt et al., 2011; Law and Gold, 2009). Direct neural evidence for a selective reweighting of sensory channels is missing. Nevertheless, first indirect evidence is provided by two recent neurophysiological (Law and Gold, 2008, 2009) and fMRI (Kahnt et al., 2011) studies, which found that perceptual learning affected the representation of stimulus information at a decisional but not a sensory stage. In **study 4** we devised a hybrid reinforcement and Hebbian learning model to search for neural evidence of a self-generated learning signal that may modulate the weighting of different sensory channels.

Enhancement of an internal sensory model

A third model of perceptual learning originates from the idea of the "Bayesian brain", which postulates that the brain represents and maintains internal models of the sensory

environment (*priors*). According to this view, perception is the process of making inferences about the current state of the environment by integrating both current sensory input and prior expectations provided by these internal sensory models. Prior expectations thus aid perception, especially under challenging perceptual conditions with weak or corrupted sensory signals. In the Bayesian framework, perceptual learning can be conceptualized as the process of improving these internal sensory models (Friston, 2005, 2009).

Little is known about how such priors are realized at the neural level (cf. Seriès and Seitz, 2013). Recently, the brain's response to unexpected stimuli has been suggested as an experimental window onto the neural dynamics of prior belief representations (Garrido et al., 2009; Friston and Stephan, 2007). The idea is that classical mismatch responses to deviant stimuli, such as mismatch negativity or P300 (i.e., a positive deflection of the EEG signal at 300ms), may, besides signaling surprise, also reflect the online updating of internal models of these sensory inputs. In Bayesian terms, this update of internal models corresponds to the trial-by-trial transition from prior to posterior belief distributions. It is precisely the aim of **study 3** to provide neural evidence for a link between mismatch responses and such belief transitions.

A seemingly separate line of research investigates the process of *pattern completion*, for which the hippocampus is believed to play a key role (Mizumori et al., 1989; Nakashiba et al., 2012; Dudukovic et al., 2011; Leutgeb and Leutgeb, 2007; Gold and Kesner, 2005). Instead of probabilistically combining prior expectations and sensory input, the idea of pattern completion is that weak or noisy sensory input can be sufficient to reinstate previous, more visible encodings of the percept, thereby completing the corrupted sensory input (Marr, 1971; for a recent review, see Rolls, 2013). However, the idea of pattern completion may be not so much different after all, as such previous encodings, too, can be regarded as models of sensory input. While pattern completion is of less use in low-level tasks like orientation discrimination, it could be a plausible mechanism for improving the recognition of multi-featured complex stimuli (scenes or objects) under impoverished viewing conditions. A hippocampal pattern completion signal, as described above, is discussed in **study 1** as a possible neural mechanism of perceptual learning in object recognition.

2.2 Improving perception through allocation of attention

2.2.1 Definition and impact on behavior

Whereas the mechanism of perceptual learning introduced in the previous section can lead to slow but long-lasting changes in perception, a more flexible mechanism to enhance perception is provided by attention. The mechanism of attention can be thought of as an allocation of limited cognitive processing resources to selected aspects of the sensory environment, for the benefit of enhanced perception and perceptual performance (Anderson, 2004). Visual attention, which is considered here, can be directed *overtly* to a stimulus or location by moving the eyes to point in that direction, or *covertly* by mentally shifting one's focus without eye movements. Although it is currently not clear whether the two types of attention ultimately rely on different neural mechanisms (Hunt and Kingstone, 2003), most visual perception studies employ covert attention paradigms (e.g., Posner's cueing paradigm, Posner, 1980) to avoid the potential confound of a shifted retinal image. Here too, we focus on covert attention, although we henceforth omit the name prefix "covert" for simplicity.

Regarding the behavioral relevance of attention, an enhancing effect of attention on visual perception has been demonstrated for many aspects of perception, including contrast sensitivity (Carrasco et al., 2000), spatial resolution (Yeshurun and Carrasco, 1998) and discrimination of complex stimuli (Spitzer et al., 1988). Furthermore, reaction times are improved, such that observers respond more rapidly to an event at an attended location than at another location (Posner, 1980). Another line of research has investigated the role of attention in object perception and provided evidence that attention enables *view-invariant* processing of objects (Stankiewicz et al., 1998; Thoma and Davidoff, 2007, 2006; Thoma et al., 2004). These studies investigated visual priming of objects under variations in view between primes and probes. The inherent experimental logic was that the occurrence of priming effects despite variations in view would also indicate view invariance of the underlying neural object representations. In manipulating object primes in terms of orientation, configuration (e.g., half-split, which too, is regarded as a view change) or mirror symmetry, these studies provided several examples of view-invariant visual priming with attention, but an abolishment of priming effects under view changes in the absence of attention. These observations fit remarkably well with the *hybrid model* of object recognition (Hummel, 2001), which poses that the representational format of objects in the brain fundamentally depends on the availability of attentional resources.

Specifically, it presumes that attention is necessary to create a structural description of objects based on segmented object parts (geons), which enables robust object recognition despite view changes, as long as the constituent object parts are visible. In contrast, object recognition without attention is limited to view-sensitive (*holistic)* object representations. Putting this hypothesis to the test by examining neural object representations with and without attention was one major goal of **study 2**.

2.2.2 Neural mechanisms

With regard to the neural basis of attention it is useful to distinguish neural mechanisms related to attentional selection (e.g., biased competition theory, Desimone, 1998) and attentional modulation per se, i.e., how attention affects neural processes of stimuli once they are in the "spotlight". Here we do not consider the case of multiple stimuli that compete for attentional selection, and instead focus on the case of a single stimulus within the attentional spotlight. A major way in which neural processing of stimuli is affected is through an increase in neural response strength (for review, see Treue, 2003). In the visual domain, these increases have been found throughout the visual processing hierarchy, from the earliest stage of visual neural processing in the lateral geniculate nucleus (O'Connor et al., 2002), primary visual cortex (Gandhi et al., 1999; Martínez et al., 1999; Somers et al., 1999), up to high-level visual cortices (Murray and Wojciulik, 2004; Serences et al., 2004; O'Craven et al., 1999) and the frontal lobes (Gitelman et al., 1999). The levels of enhancement have been found to be rather modest in V1 and V2 (typically < 5%) and more robust in higher visual areas (15%–20%) such as V4 and IT (cf. Roe et al., 2012).

Although there are other suggested mechanisms of how attention affects neural responses (e.g., reduction of variability, Mitchell et al., 2007; but see McAdams and Maunsell, 1999b), the most consistent finding to date is an increase in neural response strength. Yet, whether this increase involves an amplification of neural response or rather a mere baseline shift, has remained debated. A number of studies reported that attention leads to a multiplicative amplification of stimulus-driven neuronal responses only, but not to systematic changes in undriven activity (McAdams and Maunsell, 1999; Treue and Martínez Trujillo, 1999; Treue and Maunsell, 1999; Di Russo et al., 2001). In contrast, other studies reported results that violated the predictions of the multiplication hypothesis and implied an unspecific baseline shift in neural activity through visual attention: first, in systematically varying the stimulus contrast, a number of studies found that the amplification did not diminish as the stimulus contrast approached zero (Murray, 2008;

Williford and Maunsell, 2006; Buracas and Boynton, 2007); and second, it has been shown that spatial attention leads to increased neural responses in visual areas in the absence of any visual stimulation (Ress et al., 2000; Silver et al., 2007; Kastner et al., 1999; Luck et al., 1997). Whereas these studies focused on low-level visual areas, a few studies have investigated attentional modulation in high-level visual perception and found that an observed increase in neural activity was specific to coarse functional modules, such as parahippocampal place area or fusiform face area (Baldauf and Desimone, 2014; Serences et al., 2004; O'Craven et al., 1999). However, such a coarse analysis may not be sufficient, as it is increasingly apparent that complex stimuli are coded across distributed neuronal ensembles (Haxby et al., 2001), whereby individual ensembles may be involved in the representation of many different stimuli (Rice et al., 2014). Thus a more fine-grained level analysis of attentional modulation is desirable, taking into account the differential preference of neurons for various stimuli. In **study 2** of this thesis we thus characterized the voxel-wise modulation of responses in high-level visual cortex as a function of each voxel's object preference. If the strength of attentional modulation increased with object preference, a mere baseline shift of attention could be ruled out. Further, we went beyond an analysis of attentional modulation in terms of overall response amplitude by examining the effect of attention on the *informativeness* of responses.

Finally, the *view invariance* of object representations too, has been the subject of neuroscientific investigation (Eger et al., 2004; Thoma and Henson, 2011; Vuilleumier et al., 2005). These previous studies relied on an analysis of *repetition suppression*, whereby a drop in blood-oxygen-level dependent (BOLD) response between two subsequent stimulus presentations is thought to result from *adaptation*, and is therefore taken as evidence that the involved neurons responded to both stimuli or to the same stimulus in different views. The ensuing results were contradictory: whereas Thoma and Henson (2011) and Eger et al. (2004) found view invariance in high-level visual cortex specifically for attended but not unattended objects—in line with Hummel's hybrid model, Vuilleumier et al. (2005) reported view invariance only for unattended objects. However, repetition suppression as a mass-univariate technique misses out on any object-related information that only emerges at the pattern level as a distinct weighting of neurons or neuronal populations (see above). In **study 2** we therefore used a more sensitive multivariate approach to investigate the view invariance of object representations in

dependence of attention, which was able to pick up on object- or view-specific information coded across distributed voxels.

2.3 Perceptual modalities under investigation

2.3.1 Low-level visual perception

Early research in visual perceptual learning was almost exclusively focused on stimulus features that are processed at low levels of the visual processing hierarchy, such as spatial frequency, orientation or motion direction. The advantage of such low-level visual stimuli is that they permit a systematic variation of a single feature dimension. In addition, the experimenter can address and isolate specific functions of visual processing. An important class of such low-level stimuli are Gabor patches—two-dimensional Gaussian kernel functions modulated by a sinusoidal plane wave, because they are thought to optimally drive neurons in primary visual cortex (Mardelja, 1980; Hubel and Wiesel, 1959). For this reason, Gabor patches are the stimulus of choice for most perceptual learning studies interested in primary visual cortex plasticity (e.g., Schoups et al., 2001; Ghose et al., 2002). An additional experimental advantage is that the visibility of Gabor patches can be effectively interfered with by introducing noise of similar spatial frequencies. This property has enabled researchers to study perceptual performance of orientation discrimination with Gabor patches under systematically varied external noise conditions, while keeping the stimulus orientations constant (Petrov et al., 2005, 2006; Dosher and Lu, 1998; Dosher et al., 2013). **Study 4** used Gabor patches to investigate visual perceptual learning in orientation discrimination, and likewise exploited this property by embedding the Gabor patches in background noise of similar spatial frequency content.

2.3.2 Visual object perception

Although the majority of visual perceptual learning studies have focused on low-level visual stimuli, an important goal is to understand perceptual improvements of more realistic complex stimuli, such as objects or scenes. Classic examples for such learning come from the perceptual abilities of radiologists, ornithologists or luggage screeners, who after years of training are capable of subtle visual discriminations indiscernible to the untrained observer. These real-world examples show that the visual system can indeed improve perception of such complex stimuli. A number of studies have investigated improvements in object perception under challenging viewing conditions such as

backward masking (Baeck and Op De Beeck, 2010; Op de Beeck et al., 2007; Grill-Spector et al., 2000; Furmanski and Engel, 2000), noise degradation (Baeck and Op De Beeck, 2010) or metacontrast masking (Albrecht et al., 2010; Schwiedrzik et al., 2009, 2011). The most important behavioral findings are that (1) perceptual performance in object recognition can be improved through training (all of the above cited studies), (2) these improvements are partially specific to trained objects (all of the above cited studies), and (3) these improvements are likely based on non-retinotopic information (Furmanski and Engel, 2000). In investigating the neural correlates of such object recognition learning, a previous study (Grill-Spector et al., 2000) has linked improved object recognition performance to increased activity in *lateral occipital complex* (LOC)—a functionally defined region, which responds more to images of objects than their counterparts, stretching from lateral occipital cortex to posterior fusiform gyrus. The LOC is regarded as a key region in object recognition and object processing (Grill-Spector et al., 1999; Malach et al., 1995). Grill-Spector et al. (2000) found that BOLD activity in LOC was not only increased after 5-7 days of object recognition training with a large set of objects, but also correlated with recognition performance. They concluded that perceptual learning in object recognition is reflected in increased activity of LOC. However, a limitation of this study is that training was based on a single presentation of each image per session, which arguably is in closer resemblance to priming than to a perceptual learning (Fahle, 2004). In **study 1** we investigated object recognition learning with a more intensive training schedule, involving hundreds of presentations of each object category.

2.3.3 Somatosensory perception

Although the subject of some of the earliest published experiments in perceptual learning (Volkmann, 1858), the effects of both perceptual learning and attention in the somatosensory domain are currently much less well-studied than in the visual domain. A major reason is that studies in somatosensation often require tailored mechanical devices to present tactile or electric stimuli. Somatosensory stimuli are thus delivered through a variety of specialized devices, such as piezoelectric Braille displays (e.g., Wacker et al., 2011), disks with two needles for two-point discrimination (e.g., Pleger et al., 2001, 2003) or weak electric currents (e.g., Tamura et al., 2004; Pleger et al., 2009). In particular, electric stimulation has proven to be an effective way of studying the brain's response to "deviant" stimuli, because both deviant and "standard" stimuli can be optimized by adjusting the amperage to individual thresholds. Previous EEG studies examined the

neural correlates of deviant electric somatosensory stimuli in *mismatch* (or *oddball*) paradigms, in which repetitive sequences of identical stimuli are interspersed with a rare deviant (physically different) stimulus (Shinozaki et al., 1998; Restuccia et al., 2009; Akatsuka et al., 2007). The reported results indicate that somatosensory surprise is associated with analogous neural responses as previously reported for the auditory domain: an early mismatch negativity between 100 ms and 200 ms, as well as a later positive component at 300 ms (P300) related to attentional capture. The mismatch negativity in particular has recently attracted new interest as an experimental window to internal surprise signals (Friston, 2005; Garrido et al., 2008). The idea is that the mismatch negativity corresponds to the update of an internal sensory model through new information (here standard and deviant stimuli), and hence reflects "Bayesian perceptual learning" (cf. section 2.1.2). Given the high saliency of electric shocks and the potential for fine-tuning the intensity of stimuli, a somatosensory mismatch roving paradigm provided an effective tool in **study 3** to investigate such Bayesian perceptual learning.

2.4 Methodology

Apart from standard fMRI analysis based on cognitive subtraction (Harrison and Pantelis, 2010) of task conditions, and EEG analysis, based on event-related potentials, the studies in this thesis relied on two more specialized analysis techniques for neuroimaging data: model-based analysis of neural activation and multi-voxel pattern analysis (MVPA). In the following, these two techniques will be briefly introduced.

2.4.1 Model-based analysis of neuroimaging data

The common approach to investigate the neural basis of cognitive processes is to relate brain activation to either experimental manipulations (task conditions) or to behavioral reports (e.g., the level of awareness or an emotional state). However, not all cognitive variables are easily accessible through behavioral report or task manipulation. An instructive example is probabilistic reward learning, in which observers learn reward probabilities associated with different cues. During such an experiment, observers may estimate a number of hidden cognitive variables: (1) before the onset of the cue, they may anticipate a certain amount of reward based on the overall experience of reward frequency, (2) after the presentation of the cue, they may anticipate a certain amount of reward contingent on the cue (based on the conditional reward frequency), and (3) during

the outcome phase, they may be more or less surprised by the reward obtained or by the omission of an expected reward. In order to find neural correlates of these hidden variables (anticipated reward, conditional anticipated reward, surprise), first and foremost a model is required that is able to estimate these variables. This is the general idea behind model-based analyses. Such behavioral models take external experimental variables as input (e.g., cues and reward outcomes), compute hidden variables and predict behavior based on a set of mechanistic rules or equations. In addition, the models typically involve some unknown constants (e.g., risk proneness or learning rates), which are estimated by fitting model predictions (e.g., model-based choice probabilities to actual choices) to measured behavior. Finally, these estimated hidden variables can be used as regressors for brain signals, in order to identify neural correlates of the hidden variables (Figure 2). Thus, unlike more conventional neuroimaging approaches based on cognitive subtraction that permit inferences about *where* a given cognitive function is implemented, model-based fMRI provides insights not only into where but also as to *how* a specific function might be carried out. In **studies 3** and **4** we employed such a model-based approach to analyze brain activity measured with EEG and fMRI, respectively.

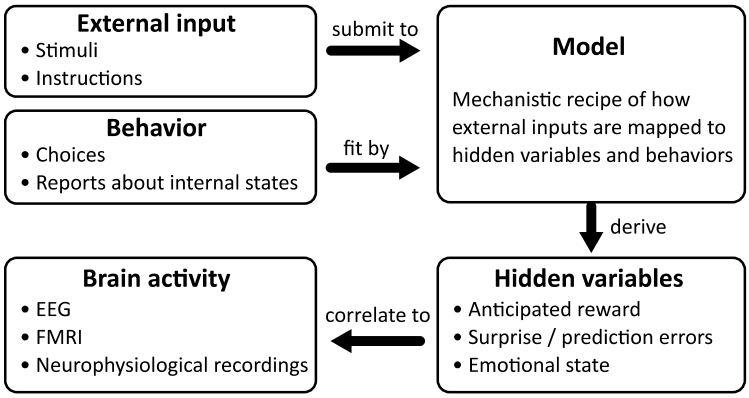


Figure 2. Model-based analyses of brain activity
A behavioral model maps inputs (e.g., stimulus intensity) to hidden cognitive variables and to behavior. After fitting the model's behavioral predictions to behavioral data, hidden variables can be derived. Finally, the hidden variables can be correlated to brain activity, enabling inference about their neural underpinnings.

Modeling behavior as Bayesian inference

A recent development in computational modeling is to describe human behavior, such as learning and perception, on the basis of Bayesian models, thereby accommodating evidence that humans optimally combine prior and new information in many contexts (den Ouden et al., 2010; Körding and Wolpert, 2004; Orbán et al., 2008; Behrens et al., 2007). Bayesian models assume that observers (1) maintain internal estimates of the likelihood $p(D|H)$ of observed data D (sensory input), conditional on their hypotheses H (or predictions) about these data, and (2) a priori probabilities $p(H)$ of these hypothesis. To make inferences about the world (i.e., to find the most probable hypothesis given data D),

the Bayesian observer combines the likelihood and the prior density to arrive at an optimal estimate of the posterior density p(H|D).

Recently, Bayesian models have been used to search for neural evidence of *surprise* signals, that is, the (un)expectedness of a current observation of data D given the prior probabilities of D. Such surprise signals are key variables of contemporary theories of learning and behavior (e.g., free energy principle). One way to define surprise in a Bayesian framework is "Bayesian surprise" (Itti and Baldi, 2009), defined as the Kullback-Leibler (KL) divergence between the probability densities of H before (prior) and after observation (posterior):

$$\text{BayesianSurprise} \coloneqq \text{KL}(\text{prior density} \,||\, \text{posterior density})$$

As the posterior density in a given trial becomes the prior density in the subsequent trial, Bayesian surprise can be estimated on a trial-by-trial basis. **Study 3** used such trial-by-trial Bayesian surprise sequences in a sensory mismatch roving paradigm to search for the neural correlates of Bayesian perceptual learning (whereby such internal sensory model updating is considered a form of perceptual learning, cf. section 2.1.2).

2.4.2 Multi-voxel pattern analysis

Conventional fMRI analysis seeks to quantify the involvement of brain areas in cognitive functions through a so-called massunivariate approach (Luo and Nichols, 2003): BOLD activity is recorded in a large amount of voxels and each voxel is analyzed separately. In addition, fMRI data are often spatially smoothed or averaged across a region of interest to increase the signal-to-noise and to find clusters of activation. While this approach is a reasonable choice in many cases, it ignores potential information about task conditions stored in spatial pattern of unsmoothed fMRI voxels. Overcoming this loss of information is the motivation for multi-voxel pattern analysis (MVPA), which rests on exactly the premise that the combined analysis of large voxel ensembles conveys additional information about task conditions. To assess the information stored in activation patterns, MVPA is often conjoined with machine learning techniques, such as support vector machine classification (Cortes and Vapnik, 1995). Instead of asking *which brain areas are significantly activated in a given task condition*, MVPA asks *in which brain areas do activation patterns significantly discriminate between task conditions*. These task conditions could be related to different stimuli presented to participants (e.g., *faces* versus *houses*) or to different instructions (e.g., *imagine being happy* versus *imagine being sad*).

The general idea is that a successful discrimination of task conditions indicates that a given brain area is involved in processing or representing these task conditions. The applicability of this approach was demonstrated by pioneering studies for activation patterns comprising regions of interest (Kamitani and Tong, 2005; Haynes and Rees, 2005a), and was subsequently extended to whole-brain information mapping (Kriegeskorte et al., 2006). The exact reasons for the effectiveness of MVPA are not well-understood, however. According to the most popular explanation—the *biased sampling hypothesis* (Haynes and Rees, 2005b; Kamitani and Tong, 2005), MVPA exploits the fact that each voxel samples a slightly different proportion of neurons tuned to different task conditions.

The usual methodological approach to MVPA is to split datasets into training and test sets, such that the predictive capacity of a machine learning classifier can be assessed without overfitting. The dependent variable is typically *decoding accuracy*, which quantifies the percentage of correctly classified samples in the test set. High decoding accuracies are an indicator of *reliable* and *distinct* activation patterns associated with different task conditions. In addition, MVPA can be used to *cross-classify* between classes of task conditions, such that the classifier is trained on one class and tested on another class. This approach is used to assess whether two classes of task conditions share a common representational basis. For instance, in a systematic assessment of whether semantic information is represented in a modality-independent manner, Simanova et al. (2014) trained a classifier on activation patterns of semantic categories of one sensory modality and tested the classifier on the same set of voxels in another modality. Here, cross-classification decoding (i.e., above-chance decoding accuracies) indicated that large clusters in frontal and temporal areas encoded supramodal semantic information. Multivariate cross-classification is a key method of **study 2**, in which we trained and tested a classifier on activation patterns associated with different configural appearances of objects. **Studies 1** and **4** applied standard within-class decoding analyses.

2.5 Brief overview of studies

The overall goal of this thesis is to elucidate the neural basis of two key mechanisms known to improve perception: attention and perceptual learning. In the first two of these studies we used fMRI in conjunction with conventional and multivariate (MVPA) data analyses to investigate the neural underpinnings of both mechanisms in visual object perception. In **study 1** we investigated behavioral characteristics and neural correlates of

perceptual learning in object recognition. A key goal, apart from identifying the brain structures that underlied behavioral improvements, was to examine whether such improvements generalized to new exemplars of object categories—both at the behavioral and the neural level. **Study 2** addressed effects of attention in object perception by probing object representations in LOC for quantitative and qualitative differences depending on the presence or absence of attention. Qualitatively, we tested the hypothesis that object processing relies on two different representations formats depending on the availability of attentional resources. Quantitatively, we tested whether attentional modulation involved the amplification of neural responses (rather than a mere baseline shift) and assessed whether such modulation enhanced the informativeness of object-related sensory signals.

The third and fourth study of this thesis were concerned with self-generated feedback signals that may underlie perceptual learning. **Study 3** searched for neural evidence of sensory model updating, as envisaged by the predictive coding framework, using a model-based analysis of EEG responses in a somatosensory mismatch roving paradigm. Finally, in **study 4** we aimed to elucidate the neural mechanisms of visual perceptual learning in the absence of external feedback. To this end, we devised a novel perceptual learning model based on confidence-based feedback signals to account for behavioral performance in an orientation discrimination paradigm and searched for neural evidence of the ensuing learning signals using fMRI in combination with model-based analyses. Note that a more detailed motivation as well as the formulation of specific research hypotheses is provided in the individual study summaries of the next chapter.

3. SUMMARY OF STUDIES

In this chapter we summarize the four studies constituting this thesis. In the first two of these studies, we investigated the neural mechanisms of enhanced object perception through training (**study 1**) and through allocation of attention (**study 2**). Subsequently, **studies 3** and **4** are presented, which were concerned with neural self-generated feedback signals that may underlie some forms of perceptual learning. Note that in order to draw conclusions across studies, a few additional analyses have been carried out specifically for this thesis (denoted by an asterisk * next to the corresponding figure letter). Only for these analyses statistical details are provided in this chapter.

3.1 Study 1: Improving object recognition through perceptual learning

Motivation and hypotheses

How do we learn to recognize objects under challenging viewing conditions? Previous behavioral studies demonstrated stimulus-specific perceptual learning in object recognition with training over several days (Furmanski and Engel, 2000; Schwiedrzik et al., 2009; Baeck and Op De Beeck, 2010; cf. section 2.3.2). However, several open questions have remained. First, it is not clear whether object recognition learning is specific to trained exemplars, or whether it generalizes to the category level and thus also leads to an improved recognition of other (untrained) exemplars pertaining to the same category. The answer to this question has important implications for whether learning involves low-level visual areas, or more high-level (potentially non-sensory) brain areas. A second and as yet entirely unaddressed question is whether perceptual learning in object recognition is susceptible to reward reinforcement in the same way as low-level perceptual learning (Seitz et al., 2009; Baldassi and Simoncini, 2011). It is thus unknown whether reward reinforcement specifically influences basic low-level visual function (e.g., contrast detection), or whether it could affect high-level processing of object exemplars or categories. Our goal at the neural level, first and foremost, was to identify the particular brain structures that underlay behavioral improvements in object recognition, both related to perceptual training and related to reward reinforcement. In addition, we were interested in whether the activity within these brain structures showed similar effects of within-category transfer as observed at the behavioral level (or absence thereof, respectively). Based on a previous study, which reported a correlation between object recognition performance and brain activity in LOC (Grill-Spector et al., 2000), we

hypothesized that response amplitude and object-related pattern information in LOC would be increased through both training and reward reinforcement.

Design

Human participants (N=18) performed a challenging object recognition task over the course of five consecutive days. The presented stimulus set consisted of animate and inanimate object categories, each comprising multiple exemplars. In the training phase (days 2−4), only a subset of categories was presented (*trained categories*). Trained categories were further subdivided into *trained* and *untrained exemplars* and were coupled with high or low monetary rewards during training. In the pre- and post-training sessions, the full stimulus set was presented and participants underwent fMRI scanning. Objects were presented for 17 ms and subsequently backward-masked with a variable SOA of 0ms (control condition with almost no visibility), 17ms (learning condition) and 33ms (control condition with high visibility). The stimulus display was followed by a response screen with two categorical choice options, among which participants chose per button press. While high or low rewarding feedback for correct responses was provided in the training phase, no feedback was given in the pre- and post-training sessions.

Results

Please note that, in the following, only results for the learning condition (SOA of 17ms) are reported. Behaviorally we found clear and continuous improvements of recognition performance across the five days (Figure 3A). Importantly, in the pre-/post-training comparison, performance improvements were significantly higher for trained relative to untrained categories, demonstrating *category-specific perceptual learning*. An analysis of the performance of untrained exemplars and (entirely) untrained categories revealed a stronger improvement for untrained exemplars, thus demonstrating *within-category transfer*. Finally, an enhancing effect of reward reinforcement was measured for trained exemplars, but this effect did not generalize to untrained exemplars.

At the neural level, we searched for changes in brain activation consistent with category-specific learning and within-category transfer. We found a strong and bilateral category-specific training effect in the hippocampus, that is, hippocampus showed greater signal enhancement for trained relative to untrained categories (Figure 3B,C). A more detailed anatomical analysis indicated that the extent of this training effect was largely confined to a specific subfield of the hippocampus, the *subiculum*. Remarkably, when we confined the analysis of the training effect to the subset of untrained exemplars, we found an equally

strong effect in the bilateral hippocampus (Figure 3D), thus reproducing the effect of within-category transfer observed at the behavioral level. Further, in relating brain activity to behavior, we found that the strength of the hippocampal category-specific training effect (based on individual beta values at the peak voxels) correlated with the category-specific improvements at the behavioral level (Figure 3E). Regarding the effects of reward reinforcement, we did not find a reward-by-session interaction in the hippocampus or elsewhere. However, a correlation between the neural and the behavioral reward effects showed a significant modulation of hippocampal activity by reward in bilateral hippocampus (Figure 3F). Just as observed behaviorally, this neural reward effect was constricted to trained exemplars.

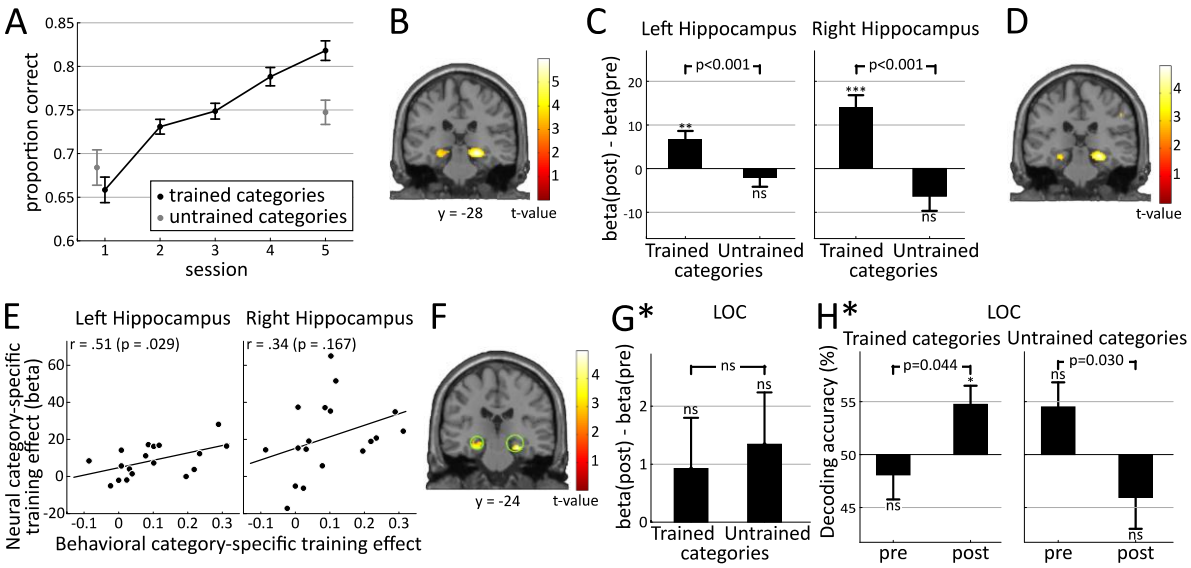


Figure 3. Neural correlates of perceptual learning in object recognition.

A Performance (proportion correct) for trained and untrained categories for all sessions. Note that between-subject variance was removed for illustration (Cousineau, 2005). **B** Whole-brain t-map for the category-specific training effect, based on the following contrast: (trained_categories_post − trained_categories_pre) > (untrained_categories_post − untrained_categories_pre). **C** The category-specific training effect was largely based on an increase in activity for trained categories. Untrained categories showed no significant change. **D** Correlation between the behavioral and neural category specific training effect, whereby the latter was based on individual beta values of the bilateral hippocampal group-level peak voxels. **E** Within-category transfer: bilateral hippocampal activation increased also for the subset of untrained exemplars (relative to untrained categories). **F** T-map for the correlation between the behavioral and neural reward effect (trained exemplars only). The reward effect was based on the contrast (high_reward_post – high_reward_pre) > (low_reward_post – low_reward_pre). Green circles indicate a hippocampal mask used for this contrast. **G*** Neither untrained nor trained categories showed an increase of activation in LOC and there was no significant interaction of training and session. **H*** A multivariate analysis indicated an increase of object-related information in LOC for trained categories and a decrease for untrained categories.

Whole-brain t-maps are thresholded at p<0.005, uncorrected , for illustration. Asterisks indicate statistical significance based on two-tailed t-tests relative to 0 (beta values) or to the chance-level decoding accuracy of 50%: * <0.05 ** <0.01 *** <0.001 ns: not significant.

However, contrary to our original hypothesis, there was no effect of training on activity in LOC: neither trained (p=0.30, t(17)=1.06, two-tailed t-test), nor untrained (p=0.15, t(17)=1.50) categories showed increased activation in the pre/post comparison, and there was no interaction of session and training (p=0.50, F(1,17)=0.47, repeated-measures ANOVA) (Figure 3G*). A multivariate decoding analysis between LOC activation patterns of object categories showed a weak pre/post increase of decoding accuracy for trained categories (p=0.044, t(17)=2.18) and a decrease for untrained categories (p=0.030, t(17)=2.36), leading to a significant interaction of session and training (p=0.018, F(1,17) = 6.80) (Figure 5H*). However, this interaction was in part driven by a baseline difference between trained and untrained categories in the pre-training session, and in part by the reduction for untrained categories. Since the source of these two additional effects is not clear, the overall interaction effect has to be treated with caution. An additional planned analysis to compare decoding accuracies of high- and low-rewarded trained categories (not shown) failed entirely, likely because of power issues after splitting trained categories according to reward.

Conclusion

We found that hippocampal activation before and after training was consistent with (1) category-specific performance improvements, (2) the generalization of learning to untrained exemplars and (3) an additional facilitatory effect of reward. Together these results suggest a key role of the hippocampus in (or after) perceptual learning of object recognition, which is in line with recent evidence for an involvement of the hippocampus beyond memory in perception and perceptual learning (Lee et al., 2012; Mundy et al., 2013; Aly et al., 2013). The fact that the category-specific training effect was largely confined to the subiculum is noteworthy in the light of recent evidence indicating a role of the subiculum in *pattern completion*, whereby partial cues reinstate information that was present during previous encodings (Dudukovic et al., 2011). Thus the increase in hippocampal activation after perceptual learning in object recognition might indicate a more prominent role of such top-down pattern completion signals. By contrast, LOC, which was implicated in object recognition learning by a previous study (Grill-Spector et al., 2000), showed no significant changes in activation, and only weak changes with regard to object-related information encoded in activation patterns.

3.2 Study 2: Improving object encoding through allocation of attention

Motivation and hypotheses

Whereas perceptual training as shown in **study 1** can lead to slow, but long-lasting changes in perception, allocation of attention to a stimulus provides a more flexible and transient mechanism for perceptual enhancement. The results of previous priming studies have indicated that directing attention to objects enables a flexible *view-invariant* neural coding scheme, in which the objects are represented in terms of their constituent parts (Stankiewicz et al., 1998; Thoma et al., 2004; Thoma and Davidoff, 2007; Thoma et al., 2007). These studies found that object primes facilitate subsequent object recognition even when primes and probes are presented in different views—but only if attention had been directed to the primes; in contrast, for unattended object primes such view-invariant facilitation is abolished (Stankiewicz et al., 1998; Thoma et al., 2004; Thoma and Davidoff, 2007; Thoma et al., 2007). Based on these findings, Hummel (2001, 2013) proposed that attention enables the activation of part-based (non-holistic) object representations, which are thought to enable robust object recognition even of unfamiliar object views, whereas object recognition without attention is limited to holistic view-sensitive object representations (*hybrid model*; cf. section 2.3.2). The present fMRI study was in part inspired by these behavioral observations and theoretical consideration. Our research agenda was two-fold. First, we specifically designed an experiment to test Hummel's hypothesis of different representational formats of object representations in dependence of attention. Our region of interest was LOC due to a large body of evidence supporting its pivotal role in object processing (Malach et al., 1995; Grill-Spector et al., 1998) and object recognition (Grill-Spector et al., 2000). Based on the results of these previous priming studies, we hypothesized to find *similar activation patterns* associated with objects in different views, when *attention is directed to objects,* and *distinct activation patterns* for different views in the *absence of attention*. Second, we investigated the attentional modulation of object representations in high-level visual cortex in more detail. Given the evidence for an enhancing effect of attention on perceptual performance, we hypothesized that attention leads to an amplification of neural responses (rather than a mere baseline shift) and to a more robust and informative object code in LOC.

Design

Human participants (N=18) were presented with real-world objects located left or right of fixation, while undergoing fMRI scanning. The objects were presented in either intact or half-split configuration and were either attended or unattended. The half-split manipulation, while preserving the constituent object parts, distorted the holistic view. The attentional manipulation involved an additional noise stimulus on the contralateral side of fixation and a brightness change detection task. As indicated by a visual cue at the beginning of each trial, a brightness change had to be detected either on the object (attended condition) or on the noise stimulus (unattended condition). Participants were instructed to fixate at the center of the screen at all times. The design enabled us to test whether intact and split objects shared a common neural code, which would be indicative of a non-holistic object code, and to assess whether this commonality depended on attention.

Results

In a first step we showed that object identity could be reliably decoded in a *within-configuration* decoding analysis, that is, training and testing of the classifier on activation patterns of *either* intact *or* split objects. The critical next step was the *cross-configuration* analysis, in which the classifier was trained on patterns of intact objects and tested on patterns of split objects, and vice versa. Remarkably, we found that cross-configuration decoding accuracy was not different from within-configuration decoding accuracy—both for attended and for unattended objects (Figure 4A). This result provides evidence for part-based (non-holistic) representations irrespective of attention—in disagreement with Hummel's hypothesis. A supplementary analysis, in which the classifier was trained and tested on object presentations on contralateral sides of the fixation cross, confirmed these findings (Figure 4B), thereby demonstrating that the finding of complete cross-configuration generalization persisted for high-level neuronal populations with receptive fields encompassing a horizontal area of at least 11.4 degree in visual angle. Finally, a *cross-attention* analysis, in which the classifier was trained on patterns of attended and tested on patterns of unattended objects (and vice versa), showed above-chance classification, providing evidence that attended and unattended objects share a common neural basis (Figure 4C).

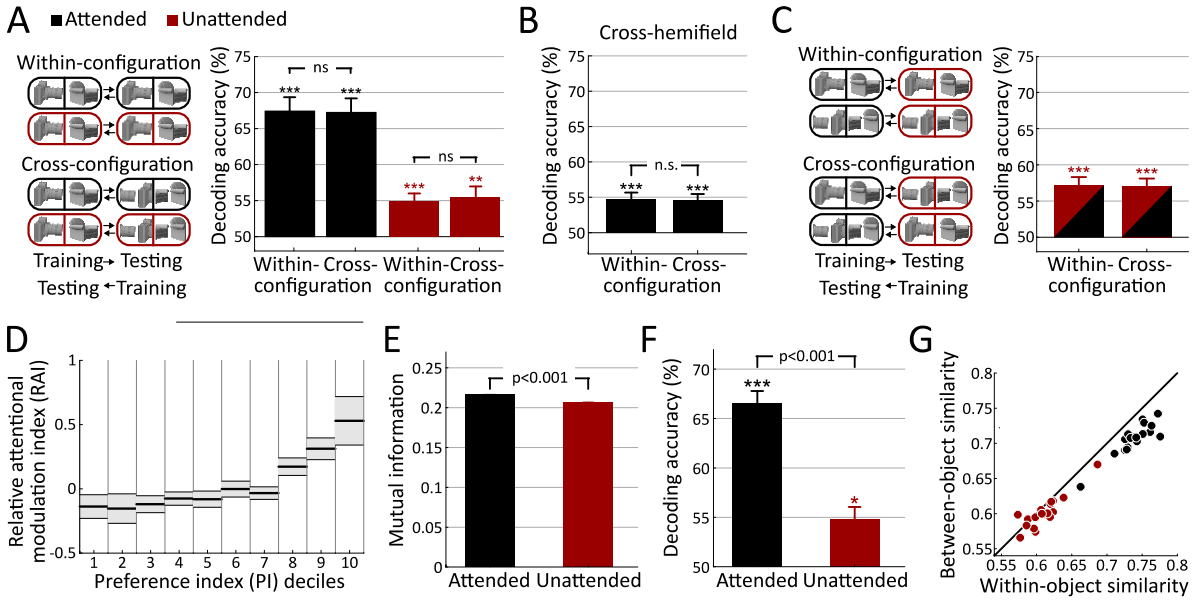


Figure 4. High-level visual object representations in dependence of attention

A Within- and cross-configuration decoding accuracies are statistically undistinguishable both with and without attention, providing evidence for a non-holistic code in LOC irrespective of attention. **B** Within- and cross-configuration decoding under a cross-hemifield decoding scheme. **C** Successful cross-classification between activation patterns of attended and unattended objects indicates a common neural code irrespective of attention. **D** Relative attentional modulation index (RAI) as a function of a voxel's preference index (PI) (see text for definitions of RAI and PI). **E** Attention increased the mutual information between single-voxel BOLD responses and object identity. **F** Attention led to an improved readout of object-related information from LOC, as indicated by greater decoding accuracies for attended relative to unattended objects. **G** Attention increased the similarity between activation patterns of different objects (*between-object similarity*), and between activation patterns of repeated presentations of the same objects (*within-object similarity*). Note that between-subject variance was removed for illustration.
Color code: black – attended; red – unattended.

Our second objective was to probe the nature of attentional modulation and to examine to what extent such modulatory effects of attention resulted in enhanced object representations in LOC. Note that this analysis was limited to intact objects only. We reasoned that if attentional modulation involved an amplification of neural responses, rather than a mere baseline shift, the overall amplification should depend on a voxel's *preference* for an object in the absence of attention. To compute a voxel's preference for a given object, we subtracted the average of its responses to all other objects from its response to the given object, whereby responses were computed as the average of run-wise beta values of the respective object regressors in the unattended condition. In a similar vein, we computed a *relative attentional modulation index* as the difference between the attention contrast (attended responses minus unattended responses) of the given object and the average attention contrasts of all other objects. When we sorted the

voxels in LOC according to object preference, we found a positive relationship between the relative attentional modulation index and object preference (Figure 4D), thus providing evidence for a voxel-wise object-specific attentional modulation.

We next assessed how such attentional modulation affected object representations. Our analyses revealed that attention (1) increased the overall BOLD response, (2) increased the mutual information between BOLD response and object identity (Figure 4E), (3) increased the informativeness of activation *patterns* about object identity (Figure 4F), and (4) increased the similarity between activation patterns of repeated presentations of the same object (*reproducibility*), as well as the similarity of patterns of different objects (Figure 4G). Concerning the latter point, a direct comparison of within- relative to between-object pattern similarity showed a significantly greater increase of within-object pattern similarity. This result indicates that attention led to a *functionally relevant* improvement of representational reproducibility compared to a case, in which the increase in reproducibility had been outweighed by an equivalent increase in ambiguity between different objects.

Conclusion

Our results show that LOC encodes objects non-holistically even in the absence of attention and therefore do not provide evidence for Hummel's hybrid model. Rather, we found that, beyond sharing a similar representational *format*, attended and unattended objects exhibit a *common neural basis*. Together these results do not support the notion that a behavioral advantage of attention for view-invariant object perception is based on a non-holistic versus holistic representational code for attended and unattended objects, respectively.

If objects are coded in a common format irrespective of attention, how are the behavioral advantages of attention to be explained? Our quantitative analyses and ensuing results might provide an answer to this question by showing that the attentional amplification of neural responses led to enhanced object representations in terms of response strength, informativeness and reproducibility.

3.3 Study 3: In search for neural evidence of internal sensory model updating

Whereas the first two studies were specifically investigating neural mechanisms supporting object perception, the third and fourth study were concerned with more general, potentially modality-independent, perceptual learning signals. Thus, the specific sensory modalities probed in these two studies were more experimental tools, rather than primary foci of interest.

Motivation and hypotheses

How does the brain adapt to regularities in its sensory environment? The Bayesian brain hypothesis and its derivatives (predictive coding principle: Rao and Ballard, 1999; free energy principle: Friston, 2005) state that the brain maintains internal generative models of its sensory input, based on which it predicts—and thereby reduces surprise about—new sensory events. In this context, the process of perceptual learning refers to a continuous update of the internal sensory model based on new sensory observations (Friston, 2005). Recently, *Bayesian surprise* (cf. section 2.4.1) has been suggested as a marker of such internal model updating, quantifying the difference of the model's internal probability densities before (prior) and after (posterior) an observation (Friston and Stephan, 2007; Feldman and Friston, 2010). The present study addressed the hypothesis that sensory processing exhibits dynamics that are consistent with such Bayesian perceptual learning, thereby using the somatosensory modality as a model case.

Design

Human participants (N=15) took part in a somatosensory mismatch roving paradigm (Baldeweg et al., 2004), in which brief electrical stimuli were delivered to the left median nerve via adhesive electrodes attached to the wrist. Two intensity levels (low/high stimulus amplitude) were adjusted on an individual basis to account for participant-specific sensory thresholds. The low stimulus intensity was determined to be close to detection threshold but clearly noticeable for every stimulus presentation. The high stimulus intensity was chosen to be markedly distinguishable from the low stimulus intensity, but not painful and below the motor threshold. Stimuli were delivered in consecutive trains of alternating stimulus intensity. Since the experiment was based on a roving paradigm, both stimuli (low/high amplitude) took on the role of deviant and standard stimuli. By averaging over deviant and standard potentials, the differential responses to the physical stimulus could be discounted.

Models

The tested Bayesian surprise models were based on participant- and session-specific trial-by-trial Bayesian surprise sequences, as quantified by the Kullback-Leibler divergence between the prior density before observation of a new stimulus, and the posterior density after observation. To account for the assumption that the brain uses finite time-windows to dynamically update its generative model, five Bayesian surprise models were tested with different forgetting time constants, including a model without forgetting. Additionally, three less complex but widely used models to account for trial-by-trial source amplitude variations were tested as control models: (1) a stimulus change model (model SC), with regressors indexing deviant stimuli as 1's and standard stimuli as 0's; (2) a linearly modulated stimulus change model (model LIN), implementing a linear relationship between the expression of the evoked source activity and the number of standards preceding a deviant stimulus; and (3) a constant null model (model M0) with a regressor comprising a vector of 1's, which served as a baseline.

Results

Inspection of the grand mean somatosensory evoked potential at the channel level confirmed the presence of well-established SEP components (Figure 5A). The largest differences between deviant and standard somatosensory evoked potentials were observed in time windows around 140, 200 and 350 ms post-stimulus (Figure 5B). To localize sources, time-windows of interest were selected based on the peak times of the most prominent deflections in both the non-differential (Figure 5A) and differential (Figure 5B) somatosensory evoked potentials. Overall, six equivalent current dipole sources were determined, including bilateral secondary somatosensory cortex and medial cingulate cortex (Figure 5C). In a next step, the peri-stimulus trial-by-trial electrode data were projected onto the identified set of dipole sources and the ensuing peri-stimulus trial-by-trial activity was modeled with Bayesian surprise models and control models. The analyses of group model log-evidences indicated two time windows in which Bayesian surprise models with forgetting were statistically superior to the model without forgetting and to control models: an early period in right secondary somatosensory cortex around 140 ms and a later period in medial cingulate cortex around 360 ms (Figure 5D). This result indicates the manifestation of perceptual learning at different temporal stages and spatial locations.

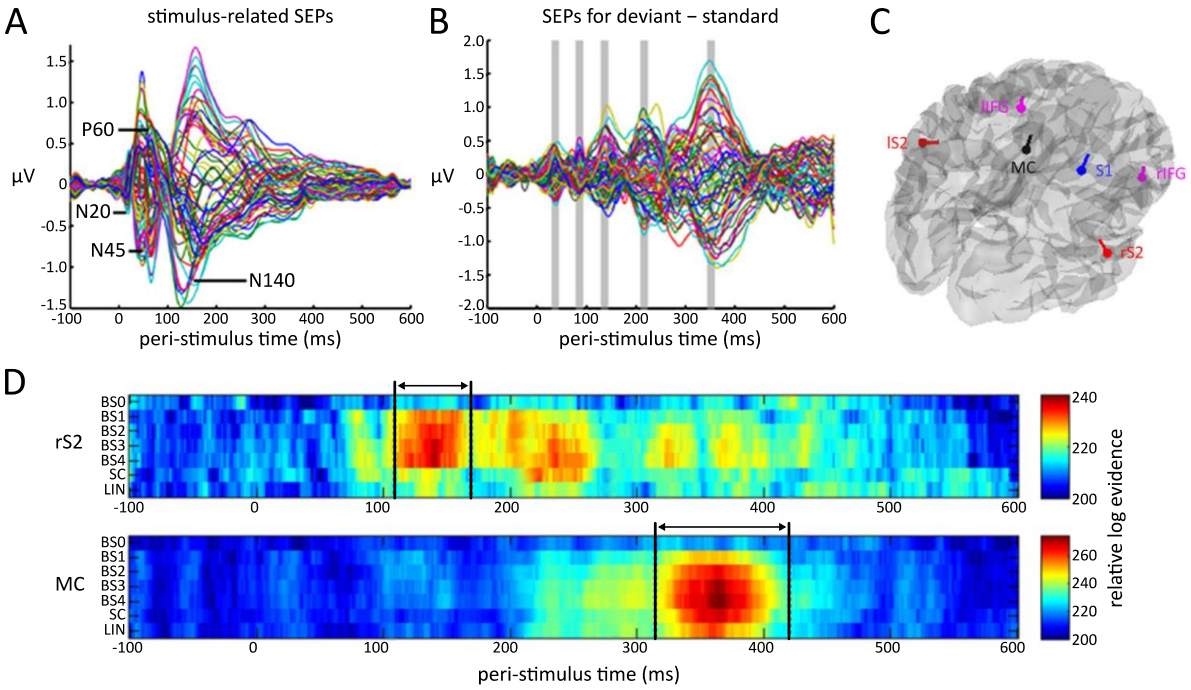


Figure 5. Computational modeling of Bayesian surprise.
A Grand mean somatosensory evoked potentials (SEPs) across all stimuli and participants, for all electrodes. The classical SEP peaks (N20, N45/P60 and N140) are labeled. **B** Grand mean difference (deviant−standard) waveform for all electrodes. The largest differences between deviant and standard SEPs are observed in time windows around 140, 200 and 350 ms post-stimulus, while smaller differences are found around 40 and 85 ms post-stimulus. **C** The trial-by-trial electrode space EEG data was projected onto a set of six oriented equivalent current dipole sources, consisting of contralateral primary (S1), bilateral secondary somatosensory cortex (lS2, rS2), bilateral inferior frontal gyrus (lIFG, rIFG), and medial cingulate cortex (MC). **D** Relative group model log evidences for two dipole sources (rS2, MC). Each row within the two panels corresponds to a specific model, relative to a constant null model (B0: Bayesian surprise model without forgetting; BS1–BS4: Bayesian surprise models with different time constants for the forgetting constant; SC/LIN: control models). Black bars mark the beginning and end of periods, which include statistically superior model evidence for Bayesian surprise models with forgetting (BS1-BS4).

Conclusion

The results of the model-based approach demonstrated spatiotemporal encoding of Bayesian surprise in somatosensory-evoked EEG responses. The finding of early-processing/low-level perceptual learning in right secondary somatosensory cortex is in line with results of previous imaging and electrophysiological studies that implicated secondary somatosensory cortex in fast perceptual learning (Pleger et al., 2003; Romo et al., 2003). By contrast, additional late-processing/high-level perceptual learning attributable to medial cingulate cortex may reflect learning-induced updating of top-down attentional control mechanisms (Gilbert et al., 2001). Taken together, these results provide neural evidence for perceptual learning processes consistent with the Bayesian brain hypothesis.

3.4 Study 4: Perceptual learning guided by confidence-based neural feedback signals

Motivation and hypotheses

The previous **study 3** has provided evidence for neural learning signals associated with the enhancement of an internal sensory model. As introduced in section 2.1.2, another major class of perceptual learning models is based on an enhancement of sensory readout. A major drawback of previous models based on plastic sensory readout (Dosher et al., 2013; Petrov et al., 2005; Kahnt et al., 2011) is that they were explicitly based on external feedback, although perceptual learning in particular is known to occur without such feedback (cf. section 2.1.1). In the present study we tested the idea that *self-generated teaching signals* based on perceptual confidence may fine-tune perceptual readout filters in the absence of external feedback. Our rationale was that the brain may improve perception by *strengthening* neural circuitry giving rise to a percept with *higher-than-expected confidence*, and by *weakening* circuitry leading to percepts with *lower-than-expected* confidence. To this end, we devised a hybrid reinforcement and Hebbian learning model, in which a confidence prediction error—the difference between actual and expected confidence (Daniel and Pollmann, 2012)—provided internal feedback. To test the model at the behavioral and neural level, we designed an fMRI experiment in which participants judged the orientation of noisy Gabor stimuli without feedback. Based on recent evidence of confidence signals in the ventral striatum (Daniel and Pollmann, 2012; Hebart et al., 2014; Schwarze et al., 2013), we hypothesized to find neural correlates of confidence prediction errors in the ventral striatum, but also in ventral tegmental area, given a possible dopaminergic origin of these striatal signals (Daniel and Pollmann, 2014). Since confidence prediction errors acted as a teaching signal in our model, we further hypothesized that the strength of striatal modulation should be predictive of individual performance improvements in orientation discrimination.

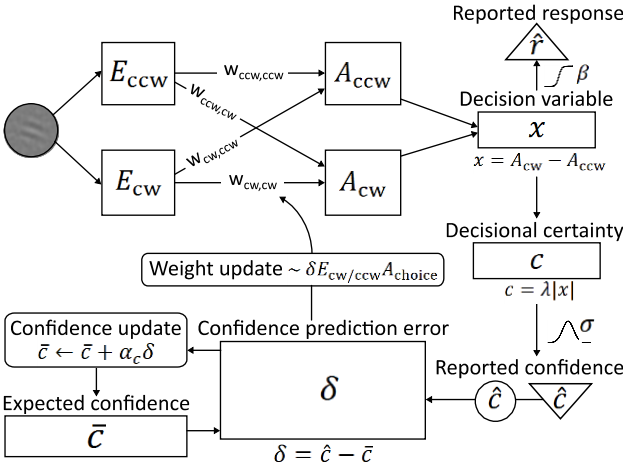
Design

Human participants (N=29) learned to detect the orientation of peripheral noise-embedded Gabor patches while undergoing fMRI scanning. In addition, the experiment comprised three sessions outside the MRI scanner: an initial session to establish the pre-training baseline for two reference axes, as well as two post-experimental test sessions to examine both short-term (1 day after training) and long-term (10 weeks after training) stimulus-specific training effects. The Gabor patches were flashed briefly in the upper right

quadrant and participants had to judge their orientation with respect to a horizontal or a vertical reference axis (participants were randomly assigned to one of the reference axes). Importantly, they did not receive external feedback during the entire experiment, except for an initial familiarization with the stimuli. Rather, in addition to their choice, the participants reported their confidence about the stimulus orientation on a visual analogue scale. The experiment consisted of two conditions in randomly interleaved presentation: a *constant performance* condition, in which the contrast of the Gabor patches was continuously adapted to a performance level of 80.35%, and based on which we measured perceptual learning as a change in contrast threshold; and a *constant contrast* condition with a contrast threshold set to the measured threshold of the pre-test, based on which we assessed stimulus-related neural activity in visual areas without a potential stimulus confound.

Model

We devised a hybrid Hebbian and reinforcement learning model to account for perceptual learning in the absence of external feedback (Figure 6). The essential idea was that neural circuitry underlying the readout of information from sensory representations should be strengthened when leading to percepts with higher-than-expected confidence, and weakened when leading to percepts with lower-than-expected confidence. For this purpose, the model estimated its *expected confidence* with regard to percepts as a running weighted average of previous confidence experiences. The comparison between current and expected confidence resulted in a *confidence prediction error* (Daniel and Pollmann, 2012), with positive or negative confidence prediction errors leading to a strengthening or weakening of connections, respectively. The confidence prediction error served as a learning signal and augmented an intrinsic Hebbian learning component. The rationale of the Hebbian learning component was that those connections should be updated more strongly, which contributed more to the final choice (i.e., based on the strength of the sensory signal they were conveying). Figure 6 provides a more detailed description of the model's mechanics.



Weight update

If *cw* chosen:

$$w_{cw,cw} \leftarrow w_{cw,cw} + \alpha_w \delta E_{cw} A_{cw}$$
$$w_{ccw,cw} \leftarrow w_{ccw,cw} + \alpha_w \delta E_{ccw} A_{cw}$$

If *ccw* chosen:

$$w_{ccw,ccw} \leftarrow w_{ccw,ccw} + \alpha_w \delta E_{ccw} A_{ccw}$$
$$w_{cw,ccw} \leftarrow w_{cw,ccw} + \alpha_w \delta E_{cw} A_{ccw}$$

Parameters

α_c: learning rate for confidence update
α_w: learning rate for weight update
β: inverse temperature
λ: scaling parameter for confidence
σ: uncertainty of reported confidence

Nodes

Circles: input variables
Rectangles: latent variables
Rounded rectangles: update rules
Triangles: observed variables (to which the model is fitted, indicated by *dotted arrows*)

Note that reported confidence $\hat{c}$ serves both as an observed variable and as a model input variable

Figure 6. Perceptual learning model with confidence-based feedback.
Incoming stimulus information is analyzed by two orientation detectors measuring the clockwise (E_{cw}) and counter-clockwise (E_{ccw}) orientation energy. The orientation detectors are connected to a decisional stage (A_{ccw}, A_{cw}) through *signal weights*, which link sensory orientation information with decision units of the same orientation, as well as through *noise weights*, which link sensory orientation information to decision units with an opposing orientation. The probability for a specific decision is computed from a decision variable *x*, which is based on the difference between clockwise (A_{cw}) and counter-clockwise (A_{ccw}) output activities. The absolute value of *x* corresponds to the model's decisional certainty *c* and reflects the model's prediction for the reported confidence $\hat{c}$. In each trial, a confidence prediction error δ is computed based on the difference between the current reported confidence $\hat{c}$ and an expected confidence $\bar{c}$. To estimate the expected confidence level, the model computes a running weighted average of previous confidence experiences using a Rescorla-Wagner update rule. The update of weights comprises both a reinforcement-like ($\sim \delta$) and a Hebbian ($\sim E_{cw/ccw} A_{choice}$) learning component. Through the Hebbian component, weight updates are proportional to the correlated activity of input and output units (note that only the output unit corresponding to the participant's choice is considered for the update). The sign of the confidence prediction error δ sets the direction of the update, with negative or positive values of δ corresponding to a weakening or strengthening of weights, respectively.

Results

Over the course of the training session, participants showed a decrease of contrast thresholds for the pre-defined constant performance level (80.35 percent correct). A comparison of pre-and post-training performance revealed a stronger learning effect for trained relative to untrained stimulus orientations (Figure 7A). This orientation-specific

training effect could still be detected in the long-term test. Next, we fitted the learning model to participants' choices and confidence reports during the training session. The model fitted well with the data and accounted for participants' threshold improvements through increasing signal weights and decreasing noise weights (Figure 7B, top and middle panel). The decision variable (x) remained constant, in accordance with the constant performance of participants due to the underlying staircase procedure (Figure 7B, bottom panel). An initial *model-free* analysis of the event-related BOLD time course in an anatomical mask of the ventral striatum showed a clear division by confidence, both in terms of a positive deflection for high-confidence trials and a negative deflection for low-confidence trials (Figure 7C). Statistical analyses confirmed a main effect of confidence at stimulus onset in the bilateral ventral striatum. Notably, the BOLD time courses in the ventral striatum indicated an initial response around trial onset. We reasoned that this response corresponded to an anticipatory signal, triggered by the trial onset screen and representing the expected level of confidence. An analysis of activation in the ventral striatum at trial onset indeed confirmed the presence of an above-baseline response (Figure 7D).

Testing whether this response was related to anticipation was the first objective of the subsequent *model-based* analysis. By correlating expected confidence from the model to brain activity at trial onset, we found a significant bilateral striatal modulation by expected confidence (Figure 7E). Next, we tested our a priori hypothesis of a significant striatal modulation by confidence prediction errors at stimulus onset. As predicted, this modulation was significant in the bilateral ventral striatum (Figure 7F). Notably, we also found a modulation by confidence prediction errors in the ventral tegmental area (Figure 7G), pointing to a dopaminergic nature of the signal. A significant correlation between individual levels of learning (i.e., improvements from pre- to post-session) and the strength of striatal modulation by confidence prediction errors confirmed its relevance for perceptual learning (Figure 7H). An attempt to track the amount of orientation information, stored in activation patterns during learning, turned out to be unfeasible, because decoding of orientation information failed within visual functional regions of interest (not shown).

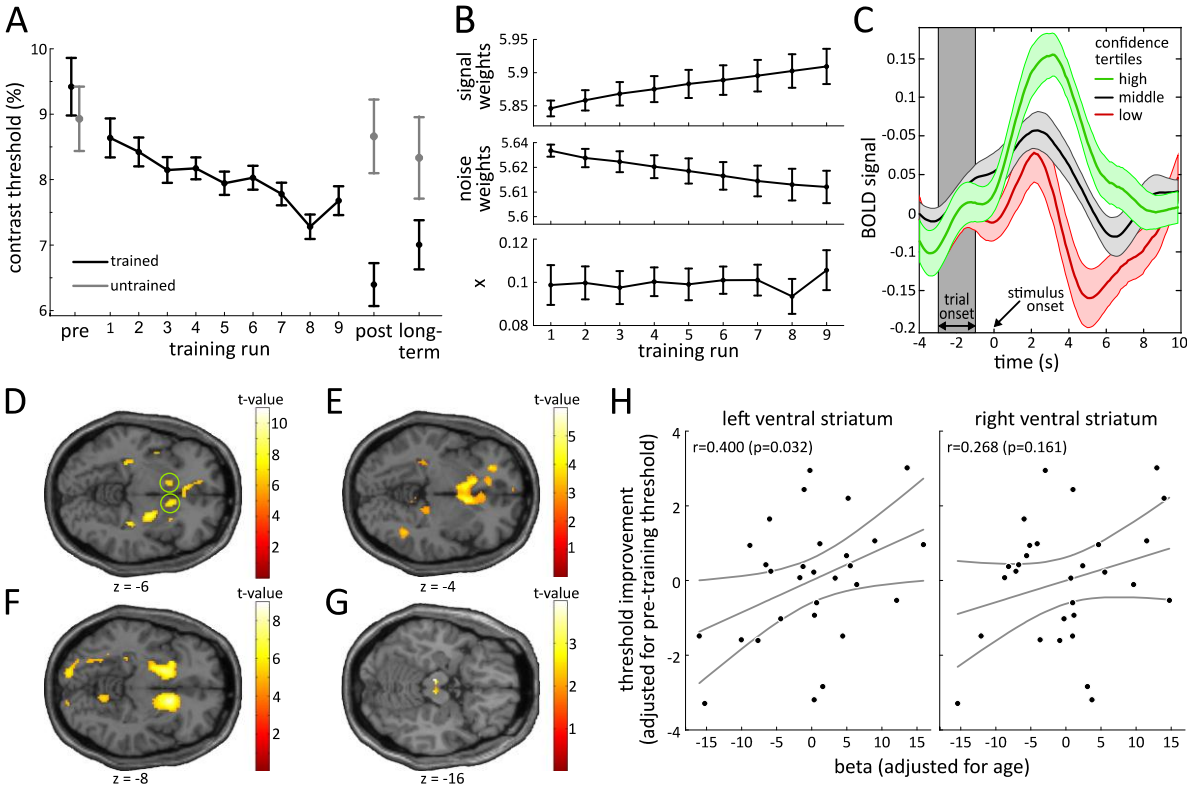


Figure 7. The role of a striatal confidence prediction error in perceptual learning

A Contrast thresholds for trained and untrained Gabor orientations across the runs of the training session and in the three test-sessions (pre/post/long-term). **B** Values of signal and noise weights (averaged across orientations), and the decision value x across training runs. **C** Event-related BOLD time course in the ventral striatum for tertiles of the behavioral confidence reports, corresponding to "high", "middle" and "low" confidence. **D** Above-baseline activation at trial onset (t-contrast; $p<0.05$, family-wise error corrected). Green circles mark significant voxels in the ventral striatum. **E** Parametric modulation of BOLD activity at trial onset by expected confidence (t-contrast; $p<0.001$, uncorrected). **F** Parametric modulation of BOLD activity by confidence prediction errors at stimulus onset (t-contrast; $p<0.05$, family-wise error corrected). **G** Same as F, but for a horizontal plane through the ventral tegmental area ($p<0.001$, uncorrected). **H** Scatter plot for the relation between individual threshold improvements and strength of striatal modulation by confidence prediction errors (beta values at group-level peaks). Statistics are based on Pearson's correlation coefficients. Pre-training thresholds were regressed out from threshold improvements and age effects were regressed out from beta values.

Conclusion

We devised and tested a novel model of perceptual learning in the absence of external feedback, utilizing an internal confidence prediction error to guide the learning process. We found that two brain areas previously implicated in learning with external feedback, ventral striatum and ventral tegmental area, correlated with model-based confidence prediction errors. In addition, the striatal confidence prediction error modulation was correlated to individual learning success. Together, these results provide behavioral and neural evidence for self-reinforced perceptual learning and suggest the mesolimbic pathway as a candidate neural substrate.

4. GENERAL DISCUSSION

4.1 Category-level perceptual learning of object recognition in the hippocampus

The key finding of **study 1** was that hippocampal activation before and after object recognition training was highly consistent with several observed behavioral characteristics. First, superior perceptual improvements for trained relative to untrained object categories were paralleled by a stronger increase of hippocampal activation for the trained categories. Second, just as there was a behavioral transfer of learning from trained to untrained exemplars within trained categories, there was an equally strong increase of hippocampal activation for both trained *and* untrained exemplars within trained categories. And third, the boosting effect of reward reinforcement was reflected in an analogous modulation of hippocampal responses, whereby both the behavioral and the neural reward effects were specific to trained exemplars. Together these results suggest an involvement of the hippocampus in object recognition in or after perceptual learning.

As such, our findings lend support to recent evidence of a hippocampal role in high-level perception and perceptual learning (Mundy et al., 2013; Aly et al., 2013; Graham et al., 2006; Lee et al., 2005a, 2005b). Yet, the *precise role* of the hippocampus in perception is currently not well understood. Mundy et al. (2013) found that activity in the posterior hippocampus was stronger for correct (versus incorrect) responses in a difficult scene discrimination task and suggested that hippocampal function is involved in perceptual discriminations between highly similar complex stimuli. Aly et al. (2013) provided complementary evidence from patients with hippocampal lesions and from fMRI data of healthy participants for a hippocampal involvement in the processing of *spatial relations* between features—as opposed to a detailed processing of these features themselves. This latter finding could match well with our finding of category- rather than exemplar-level perceptual learning, as categories might be defined on the basis of a specific spatial configuration of features. Another salient observation in our study was that functional hippocampal changes could be largely attributed to a specific subfield of the hippocampus, the subiculum. While the subiculum has previously mainly been associated with the retrieval of memories (Eldridge et al., 2005; Gabrieli, 1997), recent evidence suggests that stimulus-related subicular activation reflects a *match-enhancement signal* (Duncan et al., 2009; Dudukovic et al., 2011), caused by a neuronal firing rate increase after presentation of a target that matches an actively retained sample (Otto and Eichenbaum, 1992).

Further, there is evidence that such a match-enhancement signal may reflect *pattern completion*, a putative process which is considered to reinstate perceptual information of previous encodings based on partial sensory cues (Dudukovic et al., 2011). Applied to the case of object recognition, a pattern completion signal could be generated when the current sensory information about an object matches acquired templates of trained object categories, thereby leading to richer and more complete percepts. Bearing in mind the observed within-category transfer of perceptual learning (behaviorally and neurally), object perception might be enhanced through category-specific information, corresponding to a bias towards prototypical percepts of these categories.

4.2 The challenges of investigating plasticity in sensory brain areas with fMRI

As outlined in section 2.1.2, the role of sensory cortex plasticity in perceptual learning is a matter of great controversy and previous studies have yielded inconsistent or contradictory results. Here too, we conducted analyses of fMRI data to search for experience-dependent neural changes in visual cortex. In **study 1** we tested for training-related activation changes in LOC, inspired by a previous study reporting neural correlates of object recognition learning in this brain area (Grill-Spector et al., 2000). However, we neither found increased activity levels in LOC for trained categories in a direct pre/post-training comparison, nor an interaction with untrained object categories (i.e., correcting for unspecific training effects). Unfortunately, such discrepant results between seemingly similar perceptual learning studies are no exception in the literature (cf. section 2.1.2). A potentially illuminating point in this regard was made by Yotsumoto et al. (2008), who found an increase in sensory brain activity in the initial training phase of a perceptual learning experiment and a subsequent leveling off to baseline following additional training, with no negative effect on performance. They reasoned that the initial increase in activity may have been related to a synaptic strengthening, whereas the reduction in sensory activity in the later training phase could have been due to *synaptic downscaling*; i.e., the overall number or strength of synapses may be reduced or downscaled after training, while the synapses most critical to the task are preserved (cf., Censor et al., 2006; Tononi and Cirelli, 2003). Such a mechanism may reconcile the discrepancy between the results of study 1 and Grill-Spector et al. (2000). Whereas participants in Grill-Spector et al. (2000) had only a single exposure to each object per training session, object categories in study 1 were presented hundreds of times. Under the assumption that the degree of synaptic down-scaling is a function of experience with a stimulus (Yotsumoto et al., 2008),

associated neural changes in our study might have been more consolidated (*after* training) compared to the effects in the study of Grill-Spector et al. (2000).

Whereas the above analyses were based on an assessment of the overall response amplitude in visual cortex, another approach is to examine stimulus-related information encoded at the level of multivoxel activation patterns. This approach has, amongst others, (cf. section 2.4.2), the advantage that it is more robust with respect to processes that affect the overall amplitude of the BOLD response, such as synaptic downscaling (which may or may not be effective) or repetition suppression. We therefore conducted multivariate analyses of learning-related changes in **studies 1** and **4**. In **study 1** we assessed the informativeness of activation patterns for trained and untrained object categories before and after training, analogous to the univariate analysis. The results showed a stronger increase in decoding accuracy for trained, relative to untrained object categories. However, the significance and robustness of this result is disputable, as the effect was partly driven by a baseline difference between trained and untrained categories—likely related to the high variability between participants and between categories, and by a drop of decoding accuracies for untrained categories, the causes of both being not readily apparent. An additional analysis of decoding accuracies, for which we further split trained categories according to high versus low reward, failed entirely—likely due to insufficient statistical power. In **study 4** we originally aimed to track the amount of orientation information stored in activation patterns in visual functional regions of interest *during* the fMRI training session. However, a run-by-run multivariate analysis turned out to be unfeasible, because stimulus orientation could not be decoded, even when the data of all runs were pooled. Altogether the results of our multivariate decoding approach were therefore inconclusive in terms of results and unsatisfactory in terms of methodological feasibility. This realization also highlights a general challenge in pursuing the neural basis of perceptual learning with fMRI: since the difficulty of perceptual tasks has to be close to the threshold of awareness to leave room for learning, the associated neural signals are necessarily weak or have little discriminative power between stimuli. This problem is particularly prominent in fMRI and other non-invasive neuroimaging techniques, because of the high noise level inherent to the measurement process itself. It therefore remains to be shown whether with current fMRI technology consistent and reproducible learning-related neural changes in sensory areas can be detected. The experience in our studies and the inconsistent results of previous studies seem to suggest

that either fMRI is not sensitive enough, or perceptual learning factually does not involve systematic changes in sensory areas.

Finally, when stimulus-related sensory signals are too weak to measure learning, why might we have found a neural correlate of learning in another brain area, the hippocampus (**study 1**)? A possible answer to this question is that, while initial sensory representations in visual cortex might have been incomplete or noisy (even after perceptual training), stimulus-related activity in higher non-sensory brain areas was nevertheless measurable, because these representations might have already benefited from high-level supportive mechanisms, such as pattern completion or selective sensory readout (cf. 2.1.2). Notably, two recent studies found a similar pattern of results, i.e., representational changes in higher non-sensory, but not in sensory brain areas (Kahnt et al., 2011; Law and Gold, 2008).

4.3 Quantitative but not qualitative effects of attention on object encoding

In **study 2** we investigated the effects of attention on object representations in LOC, both qualitatively, by testing a specific hypothesis regarding different representational formats with and without attention (Hummel, 2001), and quantitatively, by assessing the effects of attention on neural response amplitude, as well as on the informativeness and reproducibility of distributed activation patterns. Previous behavioral studies have indicated that the view-point invariance of priming effects crucially depends on attention (Hummel and Stankiewicz, 1998; Thoma et al., 2004; Thoma and Davidoff, 2007; Thoma et al., 2007). These observations led to the hypothesis that attention is required to activate part-based (non-holistic) object representations, leading to a more flexible object code with greater robustness to view changes (Hummel, 2001, 2013); without attention, object recognition may be limited to holistic view-sensitive object representations. However, the results of our study did not provide evidence for a qualitative effect of attention on the format of object representations in LOC. Instead we found that LOC adheres to a non-holistic object format irrespective of attention. This finding was further corroborated by the fact that we were able to predict the attended objects based on the activation patterns of unattended objects (and vice versa). Thus, not only appears the LOC to adhere to a part-based format irrespective of attention, but the underlying neural representations additionally seem to be shared between attended and unattended objects. Importantly, we found this cross-attention generalization also under a cross-configuration decoding

scheme, demonstrating that the activation patterns of attended and unattended objects shared non-retinotopic, high-level information. Thus any behavioral advantage of attention for object perception under view changes is unlikely to be based on a non-holistic versus holistic representational code for attended versus unattended objects.

In contrast, we found that attention entailed an increase in neural response strength, leading to the amplification of object-related neural activity as opposed to a mere (unspecific) baseline shift. More specifically, we identified a positive relationship between attentional gain and voxel-wise object preference, such that the strength of attentional gain could be predicted from a voxel's preference for the object in the absence of attention. The consistent increase of the relative attentional modulation index across preference levels suggests that subtle difference in preference in the absence of attention became amplified, as attention was directed to the objects. In contrast, in the case of an unspecific baseline shift neural activity would have shown an equal increase in activity for preferred and non-preferred objects. Previous studies had already reported that attentional modulation in high-level visual cortex was specific to coarse functional modules, such as parahippocampal place area or fusiform face area (Murray and Wojciulik, 2004; Serences et al., 2004; O'Craven et al., 1999; Connor et al., 1997; Baldauf and Desimone, 2014). Our results go beyond this level of description by showing that the level of attentional gain varies as a function of the tuning of neuronal populations within these high-level visual areas. Finally, we found clear and enhancing effects of attention on the strength and informativeness of object-related responses in LOC. In particular, the fact that attentional modulation increased the informativeness of neuronal responses about object identity may entail an improved information readout by high-level executive cortices and thus benefit the perceptual decision process. Finally, an additional analysis of pattern similarity showed that attention increased the reproducibility of activation patterns of the same object. Such an increase in reproducibility would be expected on the assumption of a content-specific attentional scaling mechanism, where neuronal responses become amplified without an equivalent increase of the noise. Another possibility is that the increase in reproducibility with attention is the result of more discrete neural processing with attention, as proposed for conscious relative to non-conscious percepts (Schurger et al., 2010; Sackur and Dehaene, 2009). When discrete decisions are reached at each (object) processing stage, before they are dispatched to the next stage, the resulting activation patterns might become more stereotypical and reproducible.

4.4 The role of attention in perceptual learning

What is the relation between attention and perceptual learning? Both mechanisms improve perception, but whereas attention is a mechanism that enables flexible and transient modulation of neural activity, perceptual learning typically induces long-lasting changes in neuronal processing. However, despite these differences, attention and perceptual learning seem to share similar neural mechanisms. In a recent review, Byers and Serences (2012) highlight evidence that both top-down attention and perceptual learning have been found to operate by increasing the signal-to-noise ratio of sensory signals (e.g., Desimone, 1998; Reynolds et al., 2000; Schoups et al., 2001; Bao et al., 2010), by modulating intrinsic neural variability (e.g., Mitchell et al., 2007, 2009; Adab and Vogels, 2011), and by optimizing the readout of sensory information by decision mechanisms (Dosher and Lu, 1998, 1999; Palmer and Moore, 2009; Pestilli et al., 2011). They suggest that, whereas top-down attentional control serves to almost instantaneously modulate sensory responses, perceptual learning might be a mechanism to optimize the effectiveness of this attentional control in the long run. This effectiveness may be improved by optimizing attentional search strategies, or by changing the gain or readout of particularly discriminative or behaviorally relevant stimulus features.

Other researchers, too, have argued for the importance of a mechanism for adjusting top-down (attentional) control in perceptual learning (Dosher and Lu, 2005, 2007). Dosher and Lu provided empirical and theoretical evidence that, although enhanced perceptual readout suffices for perceptual learning in low noise environments, learning in high noise environments can only be achieved through a top-down differential adjustment of informative and uninformative signals at the sensory level. Thus, in high noise environments perceptual learning may result from an improvement in the efficiency with which top-down attentional control selectively modulates sensory gain enhancement of stimulus and noise signals. First evidence for plastic attentional gain control is provided by a recent study from Byers and Serences (2014), who observed enhanced early visual BOLD responses in voxels tuned to a task-relevant (and thus attended) grating stimulus after perceptual learning, but attenuated responses in voxels tuned away from the stimulus. Although we did not investigate perceptual learning in **study 2,** our finding of enhanced object representations with attention lends general plausibility to the hypothesis that a (long-term) fine-tuning of attentional top-down control could improve perception.

4.5 Confidence-based reinforcement signals and their role in perceptual learning

In **study 4** we were interested in neural mechanisms supporting perceptual learning with *self-generated feedback*, when neither external reward-based nor external cognitive feedback is available. Our key idea was that observers' subjective confidence about their percepts could serve as a learning signal in terms of a strengthening of neural circuitry giving rise to high-confidence percepts, and weakening of circuitry giving rise to low-confidence percepts. In this way the confidence about future percepts should become increased. More precisely, this adjustment of neural circuitry should depend on *expected* confidence, i.e., circuitry should be strengthened when the confidence was *higher than expected*, since a rather low confidence level could still present an advance, if the average confidence of preceding percepts was even lower. This supposition led to the concept of a confidence prediction error (Daniel and Pollmann, 2012), which in our study corresponded to the difference between current confidence and a running weighted average of previous confidence experiences (expected confidence). In devising a hybrid reinforcement and Hebbian learning model based on this idea and applying it to the behavioral and fMRI data of study 4, we found a correlate of these confidence prediction errors in ventral striatum and ventral tegmental area. Our result thus suggests an extension of a neural network previously implicated in reward-based and cognitive feedback (Aron et al., 2004; Daniel and Pollmann, 2010) to the domain of self-generated feedback, in line with a similar result for categorization learning (Daniel and Pollmann, 2012). The common neural basis of confidence- and reward-based learning might lend support to the notion of confidence as a form of *internal reward*, reflecting the positive evaluation of self-monitored decisions and actions (including percepts, which too, are susceptible to reinforcement learning; Wilbertz et al., 2014). Such a link between confidence and reward signals is plausible also at the level of subjective experience: Clos et al. (2014) found that—in the absence of any external rewards—trial-wise ratings for confidence in a memory task were highly correlated to those of pleasantness. Thus the subjective experience of (high) confidence seems to comprise pleasantness as an important phenomenological property, similar to the cases of receiving external reward (Berridge and Kringelbach, 2008) and cognitive feedback (Litman, 2005).

A striking finding with regard to (external) reward reinforcement in perceptual learning is that it leads to perceptual improvements even when the reinforced stimuli are presented outside of awareness (Seitz et al., 2009). Of interest in the present context is the

observation of highly similar effects when the subliminal stimuli are paired with *targets* of a parallel distractor task, instead of rewards (*task-irrelevant perceptual learning*; Seitz and Watanabe, 2003). Seitz and Watanabe (2005) suggested that the successful recognition of targets triggers a self-generated diffuse neuromodulatory signal (they likewise use the term *internal reward*), which accidently also strengthens neural circuitry responding to the background stimulus. Given the evidence for an involvement of the dopaminergic system in such self-generated feedback (see above) it is tempting to speculate that similar neural mechanisms are involved in task-irrelevant perceptual learning. This would lead to the following prediction for future studies: task-irrelevant perceptual learning should be specific to (or at least stronger for) trials, in which the target was recognized with *above-average confidence*.

More generally, behavioral reports of confidence may provide a convenient experimental window onto putative internal reward signals. It is important to note that the report of confidence might not be a direct report about these internal reward signals; rather, the idea is that confidence reports are highly correlated to such signals and thus present a behaviorally accessible proxy. As shown in our model and the model of Daniel and Pollmann (2012), confidence reports can be used to model behavior in the reinforcement learning framework equivalent to external rewards. Instead of maximizing external reward, the observer's goal in this context is to maximize confidence—or internal reward.

4.6 A common framework for neural signals in support of perception

In three studies of this thesis we provided neural evidence for signals associated with perceptual learning: a putative pattern completion signal in the hippocampus (**study 1**), a Bayesian surprise signal in somatosensory and cingulate cortex (**study 3**), and a confidence prediction error signal in ventral striatum and ventral tegmental area (**study 4**). In the following, we argue that these three signals could be described within a common framework based on the predictive coding principle (Rao and Ballard, 1999) and its extensions (free energy principle: Friston, 2005; generalized predictive coding: Feldman and Friston, 2010), which rest on the concept of *internal generative models*.

Formally, an internal model can be described as a mapping of variables of interest *Z*, which may be variables of the external world or self-monitored internal variables of the observer. This mapping is realized either by representing a full probability distribution over

observations $\hat{Z}$ of Z, as assumed in Bayesian inference, or by representing expected values of observations $\hat{Z}$, as in most reinforcement learning models (note that whereas Z is the true value of the variable, $\hat{Z}$ is the neurally encoded or *observed* value). In either case, the internal representation $\hat{Z}$ enables the model to make *predictions* about future encounters of Z. In addition, *prediction errors* convey (mis)matches between predictions for $\hat{Z}$ and observations of Z. These prediction errors are learning signals, carrying the information as to how internal models should be updated to improve future predictions. In the following we outline how the identified neural signals in studies 1, 3 and 4 may fit into the prediction/prediction error dichotomy.

Prediction signals

In **study 1** we interpreted hippocampal activation after perceptual learning with a *pattern completion signal* in object recognition, with the idea that such a signal complements weak sensory evidence $\hat{Z}$ of objects Z with information from internal object category templates. In the present framework this pattern completion signal corresponds to a prediction signal, which is integrated with the object observations to form more complete percepts (Figure 8A). The measured increase of hippocampal activity after training could thus be related to a more dominant role of such prediction signals, not least because the internal category templates themselves might have been improved during learning.

Prediction error signals

The model used in **study 3** captured how the brain internalizes the probabilistic structure of somatosensory stimulus sequences consisting of low- and high- intensity currents Z. Here, prediction errors correspond to the surprise of making observations $\hat{Z}$ of low or high intensity stimuli Z relative to their predicted probabilities (Figure 8B). The optimization of predictions—and thus the minimization of the model's surprise by the stream of incoming stimuli—is based on a continuous refinement of the internal model through prediction errors, whereby *Bayesian surprise* (cf. section 2.4.1) served as a marker of these optimization steps.

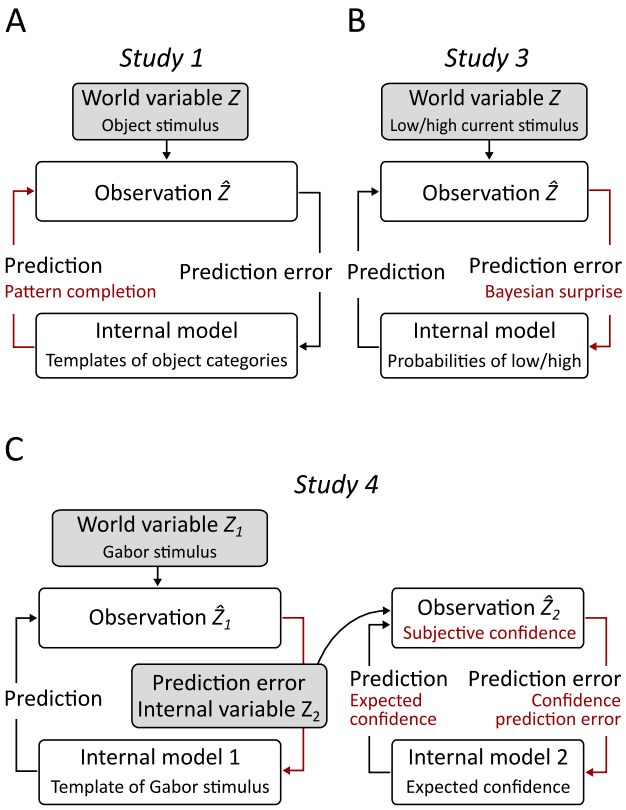


Figure 8. Suggested framework for perceptual learning signals

An agent is considered to maintain internal models of variables Z, which may either be variables of the external *world variables* or self-monitored *internal variables*. The role of these internal models is to generate predictions about observations Ẑ of these variables. The comparison of predictions and (actual) observations results in the generation of prediction errors, which serve as learning signals to narrow the gap between future predictions and observations. Note that in study 4 (panel C) both a world variable and an internal variable are observed and represented as internal models. The observed internal variable is the prediction error between the observation of the Gabor stimulus and its prediction by the internal stimulus model. The observation of this internal variable (or more precisely, its inverse) is regarded as the observer's subjective confidence, such that the confidence about the observation is assumed to be high when it closely matches the internal model. Highlighted in red are the specific signals considered in studies 1, 3 and 4.

Finally, in the perceptual learning model of **study 4,** the key observation is *subjective confidence*, which corresponds to a self-monitoring $\hat{Z}$ of an internal state Z in relation to the stimulus. The integration of confidence in the predictive coding framework, as suggested in Figure 8C, entails the assumption that this internal state Z corresponds to the prediction error associated with the comparison of observations and predictions of the Gabor stimuli. More specifically, if Gabor stimulus observations closely match with internal predicted templates, confidence is high; if these observations are far apart from the templates, e.g., because only noise is perceived, confidence is low. The internal model associated with the high-level observation of subjective confidence corresponds to a weighted running average of previous confidence experiences (*expected confidence*) and gives rise to *confidence prediction errors*—the difference between actual and expected confidence. Of note, these confidence prediction error signals in the learning model of study 4 served not only as self-contained learning signals to update the internal model, but also as outward learning signals to optimize perceptual performance (not included in Figure 8C). While it is beyond the scope of this thesis to integrate diffuse reinforcement learning signals in the predictive coding framework, it can nevertheless be noted that such

outward learning signals based on confidence prediction errors are entirely consistent with the basic principle of prediction error minimization: as the outward learning signals serve to optimize stimulus observations, rendering these observations more closely to internal templates of Gabor stimuli, they naturally lead to the minimization of prediction errors arising from the comparison between predictions and observations of the Gabor stimuli.

Attention and predictive coding

How does the role of top-down attentional modulation (**study 2**) fit in this picture? Attention has previously been posed as a challenge for predictive coding theories, given its assumed function as a top-down *enhancement* of neuronal responses to stimulus information with (predicted) behavioral relevance (Spratling, 2008; Rao, 2005). By contrast, predictive coding theories propose that top-down predictive feedback acts to *suppress* ("explain away") expected sensory information. A recent theoretical advancement termed *generalized predictive coding* (Feldman and Friston, 2010; Brown and Friston, 2013; Friston and Kiebel, 2009) has resolved this conflict by conceptualizing top-down attentional modulation as a precision weighting of prediction error signals through synaptic gain modulation. In this way, attentional gain modulation could boost the influence of behaviorally more relevant prediction errors on processing in the next level of the cortical hierarchy. This notion has received experimental backup by two recent fMRI studies, which found that attention reversed the suppressive effect of top-down prediction signals on neuronal responses of expected stimuli, such that *expected and attended* stimuli yielded the largest responses in sensory cortices (Jiang et al., 2013; Kok et al., 2012). The results of the present study 2 likewise showed an enhancing effect of attention in terms of increased signal strength, informativeness and reproducibility in object-selective visual cortex, concurring well with the notion of enhanced and more precise feed-forward prediction error signals in sensory cortices. An open question is whether such attentional precision weighting of prediction errors itself is something that can be optimized with perceptual learning (cf. section 4.4). The framework of generalized predictive coding would suggest so, casting perceptual learning as both the minimization of prediction errors through perceptual learning *and* the optimization of their precision through attention (Feldman and Friston, 2010; Brown and Friston, 2013; Friston and Kiebel, 2009).

Taken together, the above considerations lend support to the view that the (generalized) predictive coding principle provides a unifying framework for the neural signals underlying a modulation of perceptual brain processes, both through learning and attention. The key premises of the neural signals identified in this thesis are in accordance with the predictive coding principle, which casts the brain as a machinery that tries to minimize the difference between observations and internal representations of the world (here also internal states of the observer). An extension to the framework, that may be worthwhile to consider, is the idea that prediction error-based reinforcement learning signals arising from internal (as in **study 4**) or external (as in reward-based learning) feedback may operate as diffuse learning signals also outside their own prediction/prediction error loops. In this way, the framework could become inclusive of all three introduced theories of perceptual learning in section 2.1.2, as these learning signals might serve to minimize the mismatch between observations and internal models of observations, not only by an adjustment of the internal models, but also by improving observations itself through an enhancement of sensory representations or sensory readout.

4.7 Final remarks and outlook

Overall, the results of this thesis point to an important role of self-generated modulatory and feedback signals in support of perception: (i) a putative pattern completion signal in the hippocampus, (ii) a sensory surprise signal reflecting the update of internal stimulus models in somatosensory and cingulate cortex, (iii) a confidence-based feedback signal in ventral striatum and ventral tegmental area, and (iv) top-down attentional modulation in high-level visual areas. In contrast, our results provided little evidence for plasticity in sensory areas.

We introduced this thesis with William James' intuition that perception may be facilitated "[..]when we carry in our mind to meet it a distinct image of what sort of a thing we are to look for[..]". Our findings, viewed in the predictive coding framework, provide conceptual and empirical grounds to James' idea, although in a more generalized fashion. Instead of representing internal models of sensory information content only, to which James was alluding, the models in this thesis also mapped more abstract properties in relation to sensory information, such as stimulus probability and subjective confidence. This view of perception and cognition is in line with the more general notion formulated by the free energy principle, stating that the brain is fundamentally a hierarchical and generative (or

predictive) model of the world, with multiple levels of representational hierarchies abstracting away from initial stimulus representations (Friston, 2012).

Future work could more systematically design experiments to test the hypothesis that neural signals associated with perceptual improvements reflect a prediction/prediction error dichotomy. Such an approach to the analysis of neural signals may also clarify previous discrepant findings. For instance, the unintuitive finding of decreased activity in sensory cortex with perceptual learning in some studies (Op de Beeck et al., 2006; Kourtzi et al., 2005; Mukai et al., 2007; Schiltz et al., 1999) might turn out be a sign of reduced prediction errors and thus successful perceptual learning. Also the dynamics of perceptual learning as reported by Yotsumoto et al. (2008) could emerge naturally from such a framework. It may be that the initial phase of perceptual learning is dominated by large prediction errors, as observations not yet match internal templates of the observations. However, with learning these prediction errors might become smaller and thus lead to decreased neural responses. Along these lines, the investigation of perceptual learning and attention may entail powerful paradigms to study the dynamics of neural systems outside their equilibrium state, thereby providing a window on to some of the core principals of neural information processing.

5. BIBLIOGRAPHY

Adab HZ and Vogels R (2011). Practicing coarse orientation discrimination improves orientation signals in macaque cortical area v4. Current biology 21, 1–6.

Akatsuka K, Wasaka T, Nakata H, Kida T, Hoshiyama M, Tamura Y, and Kakigi R (2007). Objective examination for two-point stimulation using a somatosensory oddball paradigm: an MEG study. Clinical Neurophysiology 118, 403–11.

Albrecht T, Klapötke S, and Mattler U (2010). Individual differences in metacontrast masking are enhanced by perceptual learning. Consciousness and cognition 19, 656–66.

Aly M, Ranganath C, and Yonelinas AP (2013). Detecting changes in scenes: the hippocampus is critical for strength-based perception. Neuron 78, 1127–37.

Anderson JR (2004). Cognitive psychology and its implications. 6th ed. London: Worth Publishers.

Aron AR, Shohamy D, Clark J, Myers C, Gluck MA, and Poldrack RA (2004). Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. Journal of neurophysiology 92, 1144–52.

Baeck A and Op De Beeck HP (2010). Transfer of object learning across distinct visual learning paradigms. Journal of Vision 10, 1–9.

Baldassi S and Simoncini C (2011). Reward sharpens orientation coding independently of attention. Frontiers in neuroscience 5, 13.

Baldauf D and Desimone R (2014). Neural Mechanisms of Object-Based Attention. Science 344, 424–8.

Baldeweg T, Klugman A, Gruzelier J, and Hirsch SR (2004). Mismatch negativity potentials and cognitive impairment in schizophrenia. Schizophrenia Research 69, 203–17.

Ball K and Sekuler R (1987). Direction-specific improvement in motion discrimination. Vision Research 27, 953–65.

Bao M, Yang L, Rios C, He B, and Engel SA (2010). Perceptual learning increases the strength of the earliest signals in visual cortex. The Journal of neuroscience 30, 15080–4.

Op de Beeck HP, Baker CI, and de Beeck HPO (2010). The neural basis of visual object learning. Trends in cognitive sciences 14, 22–30.

Op de Beeck HP, Baker CI, DiCarlo JJ, and Kanwisher NG (2006). Discrimination training alters object representations in human extrastriate cortex. The Journal of Neuroscience 26, 13025–36.

Op de Beeck HP, Wagemans J, and Vogels R (2007). Effects of perceptual learning in visual backward masking on the responses of macaque inferior temporal neurons. Neuroscience 145, 775–89.

Behrens TEJ, Woolrich MW, Walton ME, and Rushworth MFS (2007). Learning the value of information in an uncertain world. Nature neuroscience 10, 1214–21.

Berridge KC and Kringelbach ML (2008). Affective neuroscience of pleasure: Reward in humans and animals. Psychopharmacology 199, 457–80.

Brown HR and Friston KJ (2013). The functional anatomy of attention: a DCM study. Frontiers in human neuroscience 7, 1–10.

Buracas GT and Boynton GM (2007). The effect of spatial attention on contrast response functions in human visual cortex. The Journal of neuroscience 27, 93–7.

Burns EM and Ward WD (1978). Categorical perception—phenomenon or epiphenomenon: evidence from experiments in the perception of melodic musical intervals. The Journal of the Acoustical Society of America 63, 456–68.

Byers A and Serences JT (2014). Enhanced attentional gain as a mechanism for generalized perceptual learning in human visual cortex. Journal of neurophysiology 112, 1217–27.

Byers A and Serences JT (2012). Exploring the relationship between perceptual learning and top-down attentional control. Vision research 74, 30–9.

Carrasco M (2011). Visual attention: the past 25 years. Vision research 51, 1484–525.

Carrasco M, Penpeci-Talgar C, and Eckstein M (2000). Spatial covert attention increases contrast sensitivity across the CSF: support for signal enhancement. Vision research 40, 1203–15.

Censor N, Karni A, and Sagi D (2006). A link between perceptual learning, adaptation and sleep. Vision research 46, 4071–4.

Clos M, Schwarze U, Gluth S, Bunzeck N, and Sommer T (2014). The Role of the Ventral Striatum in Recognition Memory. Conference of the Human Brain Mapping Organization, Hamburg, Germany.

Connor CE, Preddie DC, Gallant JL, and Essen DC Van (1997). Spatial Attention Effects in Macaque Area V4. The Journal of Neuroscience 17, 3201–14.

Cortes C and Vapnik V (1995). Support-Vector Networks. Machine Learning 20, 273–97.

Cousineau D (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. Tutorial in Quantitative Methods for Psychology 1, 42–5.

Crist RE, Li W, and Gilbert CD (2001). Learning to see: experience and attention in primary visual cortex. Nature neuroscience 4, 519–25.

Daniel R and Pollmann S (2014). A universal role of the ventral striatum in reward-based learning: Evidence from human studies. Neurobiology of learning and memory 114, 90–100.

Daniel R and Pollmann S (2010). Comparing the neural basis of monetary reward and cognitive feedback during information-integration category learning. The Journal of neuroscience 30, 47–55.

Daniel R and Pollmann S (2012). Striatal activations signal prediction errors on confidence in the

absence of external feedback. NeuroImage 59, 3457–67.

Desimone R (1998). Visual attention mediated by biased competition in extrastriate visual cortex. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 353, 1245–55.

Dolan RJ, Fink GR, Rolls ET, Booth M, Holmes A, Frackowiak RS, and Friston KJ (1997). How the brain learns to see objects and faces in an impoverished context. Nature 389, 596–9.

Dosher BA, Han S, and Lu Z-L (2010). Perceptual learning and attention: Reduction of object attention limitations with practice. Vision research 50, 402–15.

Dosher BA, Jeter PE, Liu J, and Lu Z-L (2013). An integrated reweighting theory of perceptual learning. Proceedings of the National Academy of Sciences of the United States of America 110, 13678–83.

Dosher BA and Lu Z (1999). Mechanisms of perceptual learning. Vision research 39, 3197–221.

Dosher BA and Lu Z (2005). Perceptual learning in clear displays optimizes perceptual expertise : Learning the limiting process. Proceedings of the National Academy of Sciences of the United States of America 102, 5286–90.

Dosher BA and Lu ZL (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. Proceedings of the National Academy of Sciences of the United States of America 95, 13988–93.

Dosher BA and Lu Z-L (2007). The functional form of performance improvements in perceptual learning: learning rates and transfer. Psychological science 18, 531–9.

Dudukovic NM, Preston AR, Archie JJ, Glover GH, and Wagner AD (2011). High-resolution fMRI reveals match enhancement and attentional modulation in the human medial temporal lobe. Journal of cognitive neuroscience 23, 670–82.

Duncan K, Curtis C, and Davachi L (2009). Distinct memory signatures in the hippocampus: intentional States distinguish match and mismatch enhancement signals. The Journal of Neuroscience 29, 131–9.

Eger E, Henson RNA, Driver J, and Dolan RJ (2004). BOLD repetition decreases in object-responsive ventral visual areas depend on spatial attention. Journal of neurophysiology 92, 1241–7.

Eldridge LL, Engel SA, Zeineh MM, Bookheimer SY, and Knowlton BJ (2005). A dissociation of encoding and retrieval processes in the human hippocampus. The Journal of Neuroscience 25, 3280–6.

Fahle M (2004). Perceptual learning: A case for early selection. Journal of vision 4, 879–90.

Fahle M (2005). Perceptual learning: specificity versus generalization. Current opinion in neurobiology 15, 154–60.

Fahle M, Edelman S, and Poggio T (1995). Fast perceptual learning in hyperacuity. Vision research 35, 3003–13.

Fahle M and Poggio T (2002). Perceptual learning. Manfred Fahle and Tomaso Poggio (eds) . Cambridge, USA: MIT Press.

Feldman H and Friston KJ (2010). Attention , uncertainty , and free-energy. Frontiers in human neuroscience 4, 215.

Friston KJ (2012). A Free Energy Principle for Biological Systems. Entropy 14, 2100–21.

Friston KJ (2005). A theory of cortical responses. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 360, 815–36.

Friston KJ (2009). The free-energy principle: a rough guide to the brain? Trends in cognitive sciences 13, 293–301.

Friston KJ and Kiebel S (2009). Predictive coding under the free-energy principle. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 364, 1211–21.

Friston KJ and Stephan KE (2007). Free-energy and the brain. Synthese 159, 417–58.

Furmanski CS and Engel SA (2000). Perceptual learning in object recognition: object specificity and size invariance. Vision research 40, 473–84.

Furmanski CS, Schluppeck D, Engel SA, and Angeles L (2004). Learning Strengthens the Response of Primary Visual Cortex to Simple Patterns. Current Biology 14, 573–8.

Gabrieli JD (1997). Separate Neural Bases of Two Fundamental Memory Processes in the Human Medial Temporal Lobe. Science 276, 264–6.

Gandhi SP, Heeger DJ, and Boynton GM (1999). Spatial attention affects brain activity in human primary. Proceedings of the National Academy of Sciences of the United States of America 96, 3314–9.

Garrido MI, Friston KJ, Kiebel SJ, Stephan KE, Baldeweg T, and Kilner JM (2008). The functional anatomy of the MMN: a DCM study of the roving paradigm. NeuroImage 42, 936–44.

Garrido MI, Kilner JM, Stephan KE, and Friston KJ (2009). The mismatch negativity: a review of underlying mechanisms. Clinical neurophysiology 120, 453–63.

Ghose GM, Yang T, and Maunsell JHR (2002). Physiological correlates of perceptual learning in monkey V1 and V2. Journal of neurophysiology 87, 1867–88.

Gibson EJ (1963). Perceptual Learning. Annual Review of Psychology 14, 29–56.

Gilbert CD, Sigman M, and Crist RE (2001). The Neural Basis of Perceptual Learning. Neuron 31, 681–97.

Gitelman DR, Nobre AC, Parrish TB, LaBar KS, Kim YH, Meyer JR, and Mesulam M (1999). A large-scale distributed network for covert spatial attention: further anatomical delineation based on stringent behavioural and cognitive controls. Brain 122, 1093–106.

Gold AE and Kesner RP (2005). The role of the CA3 subregion of the dorsal hippocampus in spatial pattern completion in the rat. Hippocampus 15, 808–14.

Goldstein M and Rittenhouse CH (1954). Knowledge of results in the acquisition and transfer of a gunnery skill. Journal of Experimental Psychology 48, 187–96.

Goodman JS and Wood RE (2004). Feedback specificity, learning opportunities, and learning. The Journal of applied psychology 89, 809–21.

Graham KS, Scahill VL, Hornberger M, Barense MD, Lee ACH, Bussey TJ, and Saksida LM (2006). Abnormal categorization and perceptual learning in patients with hippocampal damage. The Journal of Neuroscience 26, 7547–54.

Graham NVS (1989). Visual pattern analyzers. New York, NY: Oxford University Press.

Grill-Spector K, Kushnir T, Edelman S, Avidan G, Itzchak Y, and Malach R (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. Neuron 24, 187–203.

Grill-Spector K, Kushnir T, Hendler T, Edelman S, Itzchak Y, and Malach R (1998). A sequence of object-processing stages revealed by fMRI in the human occipital lobe. Human brain mapping 6, 316–28.

Grill-Spector K, Kushnir T, Hendler T, and Malach R (2000). The dynamics of object-selective activation correlate with recognition performance in humans. Nature neuroscience 3, 837–43.

Harrison BJ and Pantelis C (2010) Cognitive subtraction. *Encyclopedia of Psychopharmacology*. Berlin Heidelberg: Springer.

Haxby J V, Gobbini MI, Furey ML, Ishai A, Schouten JL, and Pietrini P (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–30.

Haynes J-D and Rees G (2005a). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nature neuroscience 8, 686–91.

Haynes J-D and Rees G (2005b). Predicting the stream of consciousness from activity in human visual cortex. Current biology 15, 1301–7.

Hebart MN, Schriever Y, Donner TH, and Haynes J-D (2014). The Relationship between Perceptual Decision Variables and Confidence in the Human Brain. Cerebral Cortex.

Herzog MH and Fahle M (1997). The role of feedback in learning a vernier discrimination task. Vision research 37, 2133–41.

Hua T, Bao P, Huang C-B, Wang Z, Xu J, Zhou Y, and Lu Z-L (2010). Perceptual Learning Improves Contrast Sensitivity of V1 Neurons in Cats. Current biology 20, 887–94.

Hubel DH and Wiesel TN (1959). Receptive fields of single neurons in the cat's striate cortex. Journal of Phyiology 148, 574–91.

Hummel J (2013) Object recognition. In: D Reisberg (ed) *Oxford Handbook of Cognitive Psychology*. Oxford, UK: Oxford University Press.

Hummel JE (2001). Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. Visual Cognition 8, 489–517.

Hummel JE and Stankiewicz BJ (1998). Two roles for attention in shape perception: A structural description model of visual scrutiny. Visual Cognition 5, 49–79.

Hunt AR and Kingstone A (2003). Covert and overt voluntary attention: Linked or independent? Cognitive Brain Research 18, 102–5.

Hussain Z, Sekuler AB, and Bennett PJ (2011). Superior identification of familiar visual patterns a year after learning. Psychological science 22, 724–30.

Itti L and Baldi P (2009). Bayesian surprise attracts human attention. Vision research 49, 1295–306.

James W (1890). The Principles of Psychology. New York, NY: Dover Publications.

Jiang J, Summerfield C, and Egner T (2013). Attention Sharpens the Distinction between Expected and Unexpected Percepts in the Visual Brain. Journal of Neuroscience 33, 18438–47.

Kahnt T, Grueschow M, Speck O, and Haynes J (2011). Perceptual learning and decision-making in human medial frontal cortex. Neuron 70, 549–59.

Kamitani Y and Tong F (2005). Decoding the visual and subjective contents of the human brain. Nature neuroscience 8, 679–85.

Karni A and Sagi D (1993). The time course of learning a visual skill. Nature 365, 250–2.

Karni A and Sagi D (1991). Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity. Proceedings of the National Academy of Sciences of the United States of America 88, 4966–70.

Kastner S, Pinsk MA, De Weerd P, Desimone R, and Ungerleider LG (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. Neuron 22, 751–61.

Kok P, Rahnev D, Jehee JFM, Lau HC, and de Lange FP (2012). Attention reverses the effect of prediction in silencing sensory signals. Cerebral cortex 22, 2197–206.

Körding KP and Wolpert DM (2004). Bayesian integration in sensorimotor learning. Nature 427, 244–7.

Kourtzi Z, Betts LR, Sarkheil P, and Welchman AE (2005). Distributed neural plasticity for shape learning in the human visual cortex. PLoS biology 3, e204.

Kriegeskorte N, Goebel R, and Bandettini P (2006). Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America 103, 3863–8.

Law C and Gold JI (2008). Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. Nature neuroscience 11, 505–13.

Law C-T and Gold JI (2009). Reinforcement learning can account for associative and perceptual learning on a visual-decision task. Nature neuroscience 12, 655–63.

Lee ACH, Buckley MJ, Pegman SJ, Spiers H, Scahill VL, Gaffan D, Bussey TJ, Davies RR, Kapur N, Hodges JR, and Graham KS (2005a). Specialization in the medial temporal lobe for processing of objects and scenes. Hippocampus 15, 782–97.

Lee ACH, Bussey TJ, Murray EA, Saksida LM, Epstein RA, Kapur N, Hodges JR, and Graham KS (2005b). Perceptual deficits in amnesia: challenging the medial temporal lobe 'mnemonic' view. Neuropsychologia 43, 1–11.

Lee ACH, Yeung L-K, and Barense MD (2012). The hippocampus and visual perception. Frontiers in human neuroscience 6, 1–17.

Leutgeb S and Leutgeb JK (2007). Pattern separation, pattern completion, and new neuronal codes within a continuous CA3 map. Learning & memory 14, 745–57.

Litman J (2005). Curiosity and the pleasures of learning: Wanting and liking new information. Cognition & Emotion 19, 793–814.

Lu Z-L, Liu J, and Dosher BA (2010). Modeling mechanisms of perceptual learning with augmented Hebbian re-weighting. Vision research 50, 375–90.

Luck SJ, Chelazzi L, Hillyard SA, and Desimone R (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. Journal of neurophysiology 77, 24–42.

Luo W-L and Nichols TE (2003). Diagnosis and exploration of massively univariate neuroimaging models. NeuroImage 19, 1014–32.

Malach R, Reppas JB, Benson RR, Kwong KK, Jiang H, Kennedy WA, Ledden PJ, Brady TJ, Rosen BR, and Tootell RB (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proceedings of the National Academy of Sciences of the United States of America 92, 8135–9.

Mardelja S (1980). Mathematical description of the responses of simple cortical cells. Journal of the Optical Society of America 70, 1297–300.

Marr D (1971). Simple Memory : A Theory for Archicortex. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 262, 23–81.

Martínez A, Anllo-Vento L, Sereno MI, Frank LR, Buxton RB, Dubowitz DJ, Wong EC, Hinrichs H, Heinze HJ, and Hillyard SA (1999). Involvement of striate and extrastriate visual cortical areas in spatial attention. Nature neuroscience 2, 364–9.

McAdams CJ and Maunsell JHR (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. The Journal of neuroscience 19, 431–41.

McKee SP and Westheimer G (1978). Improvement in vernier acuity with practice. Perception & psychophysics 24, 258–62.

Mitchell JF, Sundberg KA, and Reynolds JH (2007). Differential attention-dependent response modulation across cell classes in macaque visual area V4. Neuron 55, 131–41.

Mitchell JF, Sundberg KA, and Reynolds JH (2009). Spatial Attention Decorrelates Intrinsic Activity Fluctuations in Macaque Area V4. Neuron 63, 879–88.

Mizumori SJY, McNaughton BL, Barnes CA, and Fox KB (1989). Preserved Spatial Coding in Hippocampal CA1 Pyramidal Cells During Reversible Suppression of CA3c Output: Evidence for Pattern Completion in Hippocampus. Journal of Neuroscience 9, 3915–28.

Mukai I, Kim D, Fukunaga M, Japee S, Marrett S, and Ungerleider LG (2007). Activations in visual and attention-related areas predict and correlate with the degree of perceptual learning. The Journal of neuroscience 27, 11401–11.

Müller NG, Strumpf H, Scholz M, Baier B, and Melloni L (2013). Repetition suppression versus enhancement—it's quantity that matters. Cerebral cortex 23, 315–22.

Mundy ME, Downing PE, Dwyer DM, Honey RC, and Graham KS (2013). A critical role for the hippocampus and perirhinal cortex in perceptual learning of scenes and faces: complementary findings from amnesia and FMRI. The Journal of Neuroscience 33, 10490–502.

Murray SO (2008). The effects of spatial attention in early human visual cortex are stimulus independent. Journal of Vision 8, 1–11.

Murray SO and Wojciulik E (2004). Attention increases neural selectivity in the human lateral occipital complex. Nature neuroscience 7, 70–4.

Nakashiba T, Cushman JD, Pelkey KA, Renaudineau S, Buhl DL, McHugh TJ, Rodriguez Barrera V, Chittajallu R, Iwamoto KS, McBain CJ, Fanselow MS, and Tonegawa S (2012). Young dentate granule cells mediate pattern separation, whereas old granule cells facilitate pattern completion. Cell 149, 188–201.

O'Connor DH, Fukui MM, Pinsk MA, and Kastner S (2002). Attention modulates responses in the human lateral geniculate nucleus. Nature neuroscience 5, 1203–9.

O'Craven KM, Downing PE, and Kanwisher N (1999). FMRI evidence for objects as the units of attentional selection. Nature 401, 584–7.

Orbán G, Fiser J, Aslin RN, and Lengyel M (2008). Bayesian learning of visual chunks by human observers. Proceedings of the National Academy of Sciences of the United States of America 105, 2745–50.

Otto T and Eichenbaum H (1992). Neuronal activity in the hippocampus during delayed non-match to sample performance in rats: evidence for hippocampal processing in recognition memory. Hippocampus 2, 323–34.

den Ouden HEM, Daunizeau J, Roiser J, Friston KJ, and Stephan KE (2010). Striatal prediction error modulates cortical coupling. The Journal of neuroscience 30, 3210–9.

Palmer J and Moore CM (2009). Using a filtering task to measure the spatial extent of selective attention. Vision Research 49, 1045–64.

Pavlov IP, Gantt WH, Volborth G, and Cannon WB (1928). Lectures on Conditioned Reflexes

Twenty. New York: International Publishers.

Pestilli F, Carrasco M, Heeger DJ, and Gardner JL (2011). Attentional enhancement via selection and pooling of early sensory responses in human visual cortex. Neuron 72, 832–46.

Petrov AA, Dosher BA, and Lu Z-L (2006). Perceptual learning without feedback in non-stationary contexts: data and model. Vision research 46, 3177–97.

Petrov AA, Dosher BA, and Lu Z-L (2005). The dynamics of perceptual learning: an incremental reweighting model. Psychological review 112, 715–43.

Pleger B, Dinse HR, Ragert P, Schwenkreis P, Malin JP, and Tegenthoff M (2001). Shifts in cortical representations predict human discrimination improvement. Proceedings of the National Academy of Sciences of the United States of America 98, 12255–60.

Pleger B, Foerster A, Ragert P, Dinse HR, Schwenkreis P, Malin J, Nicolas V, and Tegenthoff M (2003). Functional Imaging of Perceptual Learning in Human Primary and Secondary Somatosensory Cortex. Neuron 40, 643–53.

Pleger B, Ruff CC, Blankenburg F, Klöppel S, Driver J, Dolan RJ, and Klo S (2009). Influence of dopaminergically mediated reward on somatosensory decision-making. PLoS biology 7, 1–10.

Posner MI (1980). Orienting of attention. Quarterly Journal of Experimental Psychology 32, 3–25.

Raiguel S, Vogels R, Mysore SG, and Orban GA (2006). Learning to see the difference specifically alters the most informative V4 neurons. The Journal of neuroscience 26, 6589–602.

Rao RPN (2005). Bayesian inference and attentional modulation in the visual cortex. Neuroreport 16, 1843–8.

Rao RPN and Ballard DH (1999). Predictive coding in the visual cortex : a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience 2, 79–87.

Ress D, Backus BT, and Heeger DJ (2000). Activity in primary visual cortex predicts performance in a visual detection task. Nature neuroscience 3, 940–5.

Restuccia D, Zanini S, Cazzagon M, Del Piero I, Martucci L, and Della Marca G (2009). Somatosensory mismatch negativity in healthy children. Developmental medicine and child neurology 51, 991–8.

Reynolds JH, Pasternak T, and Desimone R (2000). Attention increases sensitivity of V4 neurons. Neuron 26, 703–14.

Rice GE, Watson DM, Hartley T, and Andrews TJ (2014). Low-Level Image Properties of Visual Objects Predict Patterns of Neural Response across Category-Selective Regions of the Ventral Visual Pathway. Journal of Neuroscience 34, 8837–44.

Roe AW, Chelazzi L, Connor CE, Conway BR, Fujita I, Gallant JL, Lu H, and Vanduffel W (2012). Toward a unified theory of visual area V4. Neuron 74, 12–29.

Rolls ET (2013). The mechanisms for pattern completion and pattern separation in the hippocampus. Frontiers in systems neuroscience 7, 74.

Romo R, Hernández A, Zainos A, and Salinas E (2003). Correlated Neuronal Discharges that Increase Coding Efficiency during Perceptual Discrimination. Neuron 38, 649–57.

Rosenblatt F (1957). The Perceptron - a perceiving and recognizing automaton. Report 85-460-1, Cornell Aeronautical Laboratory.

Di Russo F, Spinelli D, and Morrone MC (2001). Automatic gain control contrast mechanisms are modulated by attention in humans: evidence from visual evoked potentials. Vision Research 41, 2435–47.

Sackur J and Dehaene S (2009). The cognitive architecture for chaining of two mental operations. Cognition 111, 187–211.

Schiltz C, Bodart JM, Dubois S, Dejardin S, Michel C, Roucoux A, Crommelinck M, and Orban GA (1999). Neuronal Mechanisms of Perceptual Learning: Changes in Human Brain Activity with Training in Orientation Discrimination. NeuroImage 9, 46–62.

Schoups AA, Vogels R, Qian N, and Orban G (2001). Practising orientation identification improves orientation coding in V1 neurons. Nature 412, 549–53.

Schultz W (2000). Multiple reward signals in the brain. Nature Reviews Neuroscience 1, 199–207.

Schurger A, Pereira F, Treisman A, and Cohen JD (2010). Reproducibility distinguishes conscious from nonconscious neural representations. Science 327, 97–9.

Schwartz S, Maquet P, and Frith C (2002). Neural correlates of perceptual learning: a functional MRI study of visual texture discrimination. Proceedings of the National Academy of Sciences of the United States of America 99, 17137–42.

Schwarze U, Bingel U, Badre D, and Sommer T (2013). Ventral Striatal Activity Correlates with Memory Confidence for Old- and New-Responses in a Difficult Recognition Test. PLoS ONE 8, e54324.

Schwiedrzik CM, Singer W, and Melloni L (2009). Sensitivity and perceptual awareness increase with practice in metacontrast masking. Journal of Vision 9, 1–18.

Schwiedrzik CM, Singer W, and Melloni L (2011). Subjective and objective awareness dissociate in space and time. Proceedings of the National Academy of Sciences of the United States of America 108, 4506–11.

Seitz AR, Kim D, and Watanabe T (2009). Rewards evoke learning of unconsciously processed visual stimuli in adult humans. Neuron 61, 700–7.

Seitz AR and Watanabe T (2005). A unified model for perceptual learning. Trends in cognitive sciences 9, 329–34.

Seitz AR and Watanabe T (2003). Is subliminal learning really passive? Nature 422, 2003–2003.

Serences JT, Schwarzbach J, Courtney SM, Golay X, and Yantis S (2004). Control of object-based attention in human cortex. Cerebral cortex 14, 1346–57.

Seriès P and Seitz AR (2013). Learning what to expect (in visual perception). Frontiers in human

neuroscience 7, 668.

Shibata K, Chang L-H, Kim D, Náñez JE, Kamitani Y, Watanabe T, and Sasaki Y (2012). Decoding Reveals Plasticity in V3A as a Result of Motion Perceptual Learning. PLoS ONE 7, e44003.

Shinozaki N, Yabe H, Sutoh T, Hiruma T, and Kaneko S (1998). Somatosensory automatic responses to deviant stimuli. Cognitive Brain Research 7, 165–71.

Shiu LP and Pashler H (1992). Improvement in line orientation discrimination is retinally local but dependent on cognitive set. Perception & psychophysics 52, 582–8.

Silver MA, Ress D, and Heeger DJ (2007). Neural correlates of sustained spatial attention in human early visual cortex. Journal of neurophysiology 97, 229–37.

Simanova I, Hagoort P, Oostenveld R, and van Gerven MAJ (2014). Modality-independent decoding of semantic information from the human brain. Cerebral cortex 24, 426–34.

Somers DC, Dale AM, Seiffert AE, and Tootell RBH (1999). Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. Proceedings of the National Academy of Sciences of the United States of America 96, 1663–8.

Spitzer H, Desimone R, and Moran J (1988). Increased attention enhances both behavioral and neuronal performance. Science 240, 338–40.

Spratling MW (2008). Predictive coding as a model of biased competition in visual attention. Vision research 48, 1391–408.

Stankiewicz BJ, Hummel JE, and Cooper EE (1998). The role of attention in priming for left-right reflections of object images: evidence for a dual representation of object shape. Journal of experimental psychology. Human perception and performance 24, 732–44.

Tamura Y, Hoshiyama M, Inui K, Nakata H, Wasaka T, Ojima S, Inoue K, and Kakigi R (2004). Cognitive processes in two-point discrimination: an ERP study. Clinical neurophysiology 115, 1875–84.

Thoma V and Davidoff J (2007) Object recognition: attention and dual routes. In: Naoyuki Osaka, Ingo Rentschler, and Irving Biederman (eds) *Object Recognition, Attention, and Action*. Tokyo: Springer.

Thoma V and Davidoff J (2006). Priming of Depth-Rotated Objects Depends on Attention and Part Changes. Experimental psychology 53, 31–47.

Thoma V, Davidoff J, and Hummel JE (2007). Priming of plane-rotated objects depends on attention and view familiarity. Visual Cognition 15, 179–210.

Thoma V and Henson RN (2011). Object representations in ventral and dorsal visual streams: fMRI repetition effects depend on attention and part-whole configuration. NeuroImage 57, 513–25.

Thoma V, Hummel JE, and Davidoff J (2004). Evidence for holistic representations of ignored images and analytic representations of attended images. Journal of experimental psychology.

Human perception and performance 30, 257–67.

Tononi G and Cirelli C (2003). Sleep and synaptic homeostasis: a hypothesis. Brain Research Bulletin 62, 143–50.

Treue S (2003). Visual attention: the where, what, how and why of saliency. Current Opinion in Neurobiology 13, 428–32.

Treue S and Martínez Trujillo JC (1999). Feature-based attention influences motion processing gain in macaque visual cortex. Nature 399, 575–9.

Treue S and Maunsell JHR (1999). Effects of attention on the processing of motion in macaque middle temporal and medial superior temporal visual cortical areas. The Journal of neuroscience 19, 7591–602.

Tsodyks M and Gilbert CD (2004). Neural networks and perceptual learning. Nature 431, 775–81.

Volkmann AW (1858). Über den Einfluss der Uebung auf das Erkennen räumlicher Distanzen. Berichte über die Verhandlungen der Sächsischen Gesellschaft der Wissenschaft zu Leipzig, mathematische und physische Abtheiltung 10, 38–69.

Vuilleumier P, Schwartz S, Duhoux S, Dolan RJ, and Driver J (2005). Selective attention modulates neural substrates of repetition priming and "implicit" visual memory: suppressions and enhancements revealed by FMRI. Journal of cognitive neuroscience 17, 1245–60.

Wacker E, Spitzer B, Lützkendorf R, Bernarding J, and Blankenburg F (2011). Tactile motion and pattern processing assessed with high-field FMRI. PloS one 6, e24860.

Walk RD (1966). Perceptual learning and the discrimination of wines. Psychonomic Science 5, 57–8.

Watanabe T and Sasaki Y (2014). Perceptual Learning: Toward a Comprehensive Theory. Annual review of psychology 1–25.

Wilbertz G, van Slooten J, and Sterzer P (2014). Reinforcement of perceptual inference: reward and punishment alter conscious visual perception during binocular rivalry. Frontiers in Psychology 5, 1377.

Williford T and Maunsell JHR (2006). Effects of spatial attention on contrast response functions in macaque area V4. Journal of neurophysiology 96, 40–54.

Yang T and Maunsell JHR (2004). The effect of perceptual learning on neuronal responses in monkey visual area V4. The Journal of neuroscience 24, 1617–26.

Yeshurun Y and Carrasco M (1998). Attention improves or impairs visual performance by enhancing spatial resolution. Nature 396, 72–5.

Yotsumoto Y, Watanabe T, and Sasaki Y (2008). Different dynamics of performance and brain activation in the time course of perceptual learning. Neuron 57, 827–33.

Zohary E, Celebrini S, Britten K, and Newsome W (1994). Neuronal plasticity that underlies improvement in perceptual performance. Science 263, 1289–92.

6. RESEARCH ARTICLES

6.1 A hippocampal signature of perceptual learning in object recognition

Journal: Journal of Cognitive Neuroscience

Acceptance date: 08 September 2014

URL: http://dx.doi.org/10.1162/jocn_a_00735

A Hippocampal Signature of Perceptual Learning in Object Recognition

Matthias Guggenmos[1,2], Marcus Rothkirch[2], Klaus Obermayer[1], John-Dylan Haynes[1], and Philipp Sterzer[1,2]

Abstract

■ Perceptual learning is the improvement in perceptual performance through training or exposure. Here, we used fMRI before and after extensive behavioral training to investigate the effects of perceptual learning on the recognition of objects under challenging viewing conditions. Objects belonged either to trained or untrained categories. Trained categories were further subdivided into trained and untrained exemplars and were coupled with high or low monetary rewards during training. After a 3-day training, object recognition was markedly improved. Although there was a considerable transfer of learning to untrained exemplars within categories, an enhancing effect of reward reinforcement was specific to trained exemplars. fMRI showed that hippocampus responses to both trained and untrained exemplars of trained categories were enhanced by perceptual learning and correlated with the effect of reward reinforcement. Our results suggest a key role of hippocampus in object recognition after perceptual learning. ■

INTRODUCTION

Intuitively, the recognition of objects around us seems like an easy task. Most objects have one or several characteristic features, we can use contextual cues, and we often have enough time to enable successful recognition. In many real-life situations, however, object recognition is complicated by impoverished sensory evidence, for example, through poor illumination, occlusion, or brief appearance. Yet, recognition performance in such challenging situations can be improved through training (Baeck & Op De Beeck, 2010; Furmanski & Engel, 2000; Grill-Spector, Kushnir, Hendler, & Malach, 2000). Although such perceptual learning and its neural underpinnings have been studied extensively for low-level visual features, such as contrast or orientation (see Fahle & Poggio, 2002, for a review), less is known about the mechanisms of learning in higher-level perceptual tasks, such as object recognition.

A hallmark of perceptual learning is the specificity of its effects for the features that have been learned, as reported for a number of low-level visual stimulus features (Fahle & Poggio, 2002). Previous work on object recognition learning showed improved performance specifically for those objects that were presented during training (Baeck & Op De Beeck, 2010; Furmanski & Engel, 2000; Grill-Spector et al., 2000). At the neural level, improved object recognition after training is associated with increased activation in object-selective lateral occipital complex (LOC; Grill-Spector et al., 2000).

From these findings, the important question arises whether perceptual learning of object recognition is confined to specific object exemplars or generalizes to other stimuli pertaining to the same category. Such a generalization at the category level would imply that learning is not solely based on specific low-level features of a particular object but involves a higher-level process that relies on the extraction of the features common to—and therefore constitutive of—a given category of objects.

Another important aspect of perceptual learning is the role of reward reinforcement, because of its implications for the underlying neural mechanisms, in particular with regard to an involvement of global neuromodulatory systems (e.g., the dopaminergic system; Roelfsema, van Ooyen, & Watanabe, 2010; Seitz & Dinse, 2007). At the behavioral level, it is evident that we need to be able to prioritize the learning of important stimuli in an environment with abundant sources of sensory information. One might expect, for example, that perceptual learning will be enhanced if the successful recognition of an object is associated with favorable consequences such as reward. In line with this notion, enhancing effects of reward on perceptual learning were reported for low-level orientation discrimination tasks (Baldassi & Simoncini, 2011; Seitz, Kim, & Watanabe, 2009), although it has remained unclear whether reward can also boost higher-level perceptual learning. We reasoned that, if reinforcement affected only early processing stages, the effect of reward should be limited to the particular object exemplars that are used for perceptual training, whereas a higher-level category-related mechanism should enhance performance also

[1]Bernstein Center for Computational Neuroscience, Berlin, Germany, [2]Charité - Universitätsmedizin, Berlin, Germany

© 2015 Massachusetts Institute of Technology

Journal of Cognitive Neuroscience 27:4, pp. 787–797
doi:10.1162/jocn_a_00735

for untrained object exemplars belonging to the same category.

Here, we studied the behavioral and neural effects of category-related object recognition learning and their dependence on reward. Recognition of objects presented with backward masking and associated neural responses were examined with fMRI before and after extensive training. Object categories were coupled with either high or low reward during training. To assess the specificity of learning, the pretest and posttest additionally included a set of untrained exemplars of the same categories and an additional completely untrained object category.

As the aim of our study was to investigate the learning of object recognition under challenging viewing conditions rather than the learning of new categories or boundaries between them, we used object categories that were familiar to the participants beforehand. Objects were presented very briefly and with backward masking to reduce visibility (Baeck & Op De Beeck, 2010). In coping with those challenging viewing conditions, the brain may either optimize processing at the sensory stage, as reported in a previous study (Grill-Spector et al., 2000), or improve the interpretation of the sensory evidence by means of high-level mechanisms such as the integration of sensory priors (Seriès & Seitz, 2013) or perceptual completion (Pessoa & De Weerd, 2003).

We hypothesized that category-related learning should be reflected by a specific enhancement of posttraining recognition performance for trained object categories and, importantly, that it should extend to those exemplars of the trained categories that were not presented during training. Furthermore, we predicted enhanced learning for high-rewarded as opposed to low-rewarded trained categories and reasoned that whether the effects of reward generalized to untrained exemplars would be informative in regard to the processing level at which reward reinforcement takes effect. Furthermore, we reasoned that, if reinforcement affected only early processing stages (cf. Seitz et al., 2009), the effect of reward should be limited to trained exemplars, whereas a higher-level mechanism should enhance performance for the whole object category associated with reward. Finally, we hypothesized that category-related learning effects should lead to enhanced responses to trained object categories in the LOC (Grill-Spector et al., 2000) or in medial-temporal lobe structures, such as the perirhinal cortex and the hippocampus, that have been implicated in high-level perception (Aly, Ranganath, & Yonelinas, 2013; Mundy, Downing, Dwyer, Honey, & Graham, 2013; Lee, Yeung, & Barense, 2012; Graham, Barense, & Lee, 2010; Lee, Buckley, et al., 2005; Lee, Bussey, et al., 2005).

METHODS

Participants

Twenty healthy individuals participated in this fMRI experiment. Two participants were excluded from the analysis because of excessive head movement in the scanner (several shifts of more than 3 mm), leaving 18 valid participants (10 women, mean age ± *SEM* = 25.7 ± 1.1 years) for the subsequent behavioral and fMRI data analysis. Additional to a fixed allowance (€26), participants received monetary rewards proportional to the points they earned in the five test sessions (two fMRI sessions and three behavioral sessions). All participants gave their informed written consent, and the study was approved by the local ethics committee.

Task and Experimental Setup

Each participant came in on 5 successive days. Training took place during Days 2–4 in a darkened room, in which the participants sat in front of a 17-in. LCD monitor (LG Flatron L1750S, 60 Hz, at 1024 × 768; Englewood Cliffs, NJ) with a standard computer keyboard. The pretraining and posttraining sessions on Days 1 and 5 took place in the fMRI scanner with a beamer/projection screen setup (Sanyo PLC-XT21L, 60 Hz, at 1024 × 768; Moriguchi, Japan) and a response box.

In each trial of the object recognition task (Figure 1A), the participants were briefly presented with an object (17 msec), followed by either an ISI of 17 msec (learning condition) or an ISI of 0 or 33 msec (stimulating conditions) and a pattern mask (150 msec). The condition of interest was the learning condition with an intermediate ISI of 17 msec. The 0- and 33-msec conditions were included to stimulate learning in the 17-msec condition by providing participants with the occasional experience of higher visibility (33 msec), although concurrently challenging them at the threshold of awareness (0 msec). Indeed, a number of previous studies have demonstrated the benefit of training at high- and low-accuracy levels (Liu, Lu, & Dosher, 2012; Petrov, Dosher, & Lu, 2005, 2006). In addition, it has been shown that a mix of hard and easy trials is important to foster both specificity and generalizability of learning (Ahissar & Hochstein, 1997), which was of particular importance in our paradigm. In the framework of their reverse hierarchy model, Ahissar and Hochstein (2004) further conjecture that initial easy conditions (as in our first fMRI session) may lead to a "Eureka effect," which in turn enables learning in more difficult conditions.

After a variable delay (2000 ± 1000 msec), in which a white fixation cross was shown, a response screen appeared, offering two response options left and right of the fixation cross (in randomized order). At the same time, the fixation cross turned green, indicating the beginning of the response phase. Participants had 1700 msec to make a choice. Different keys (buttons) were assigned to the options left and right of the fixation cross. After the response phase—and only during training—a feedback screen was shown for 1000 msec, indicating whether the participants responded correctly and how much reward they received (see below). During the intertrial interval (2000 ± 1000 msec), a white fixation cross was presented.

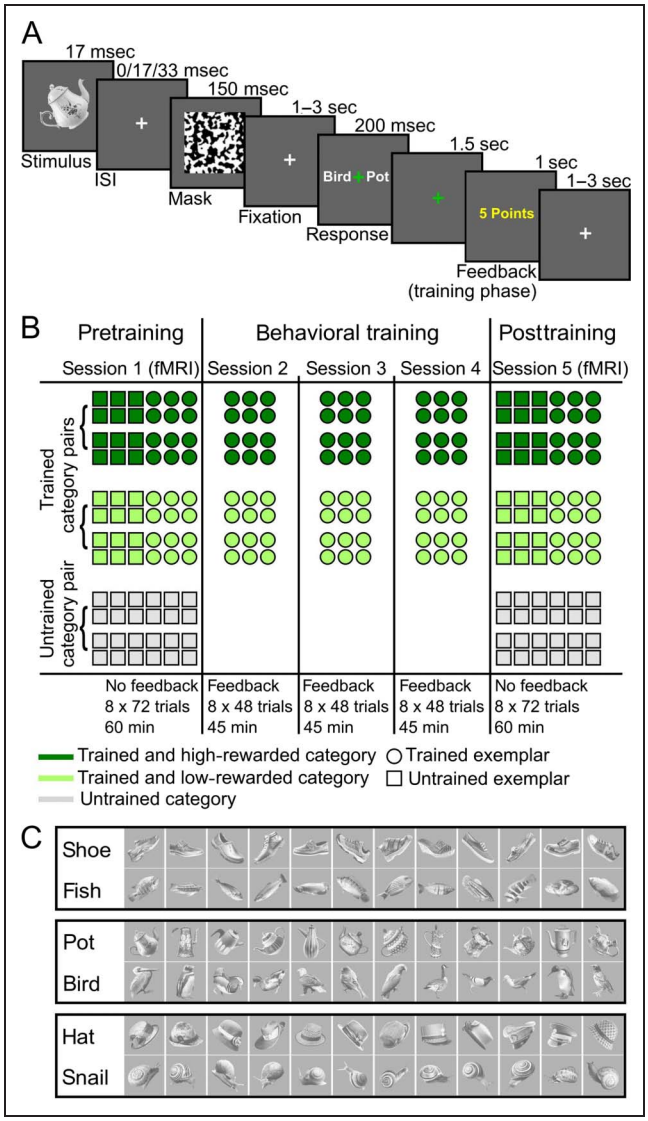


Figure 1. Experimental procedure. (A) Schematic trial of the object recognition task. (B) Experimental design. For each participant, the category pairs were split into a trained pair with high reward, a trained pair with low reward, and an untrained pair. Trained categories were additionally halfsplit into a set of trained and untrained exemplars. (C) The stimulus set consisted of six categories, each comprising 12 exemplars. Each category was assigned to one of three category pairs.

In total, there were three category pairs (shoe–fish, pot–bird, and hat–snail), with each category consisting of 12 exemplars. Participants always had to discriminate within a category pair, for example, when presented with a fish, the response screen offered the options shoe and fish (in counterbalanced order). The categories within a pair were carefully chosen to make the discrimination task challenging in the sense that the rough shape of the stimuli would not be sufficient to infer the category. Thus, by selecting categories with similar shapes and presenting them in random orientations, we forced the participants to recruit more fine-grained features of the stimuli.

For each participant, the category pairs were divided into two trained category pairs and one untrained category pair. Trained category pairs were presented both in the fMRI sessions and the training sessions, whereas the untrained category pair was only shown in the fMRI sessions. Across participants, we ensured counterbalancing of trained and untrained category pairs. The trained categories were further divided into six trained exemplars (presented in fMRI and training sessions) and six untrained exemplars (presented only in the fMRI sessions) of each category (see Figure 1B for a schematic overview).

Each fMRI and behavioral session consisted of eight runs. In each fMRI run, every exemplar of every category was presented once (adding up to 72 trials per run), whereas in runs of a behavioral session, each trained exemplar was presented twice (48 trials per run). Within each run, categories were counterbalanced across ISIs.

As a final manipulation, the trained category pairs were randomly assigned to high- and low-reward conditions. Participants received 5 points for correct responses to high-reward categories and 1 point in case of low-reward categories (0 points for incorrect responses). This assignment was stable for one participant throughout the experiment and counterbalanced across participants. Because participants received feedback only in the training phase, untrained categories were not associated with any reward. After the experiment, the sum of all collected points was converted to Euro cents (1:1 ratio, i.e., 1 point = 1 cent) and paid out to the participants.

Stimuli

The stimulus set (Figure 1C) consisted of 72 grayscale objects (450 × 450 pixels) subtending 10° × 10° of visual angle. We equalized the gray value histogram of all stimuli to match low-level properties as well as possible. The masks were generated on a trial basis by sampling a 450 × 450 pixel matrix of uniformly distributed values between 0 and 1, low-pass filtering with a Butterworth filter (cutoff frequency = 0.025 1/pixel), and thresholding at 0.5. To ensure equal physical luminance for stimulus presentation during fMRI, the parameters of all object and mask stimuli were adapted to the projection screen in the fMRI scanner based on luminance measurements. The timing of the stimulus presentation was inspected with a high-speed digital camera (1000 frames/s) and was confirmed as correct.

ROI Procedures

At the end of one of the fMRI sessions, we additionally performed a localizer run to map object-responsive regions of each participant (Malach et al., 1995). The localizer run consisted of 10 blocks of intact images and 10 blocks of scrambled images in a randomized order. Images were presented for 600 msec followed by a 200-msec blank screen. To hold participants' attention, they had to press a button whenever the same image appeared twice in a row.

We generated an ROI for the LOC (872 ± 4 voxels) based on the intersection of object-responsive voxels of

the localizer ($p_{FWE} < .05$) and an anatomical composite mask composed of the inferior and middle occipital gyrus, inferior temporal gyrus, and fusiform gyrus (derived from the Anatomical Automatic Labeling atlas; Tzourio-Mazoyer et al., 2002).

fMRI Data Acquisition and Preprocessing

Functional MRI data were acquired on a 3-T Siemens Trio (Siemens, Erlangen, Germany) scanner using a gradient EPI sequence and a 12-channel head coil. The experiment was composed of two fMRI sessions with eight runs of the main experiment on Days 1 and 5. In each run of the main experiment, 214 whole-brain volumes were acquired (repetition time = 2 sec, echo time = 25 msec, flip angle = 78°, 33 slices, descending acquisition, resolution = 3 mm *isotropic*, interslice gap = 0.75 mm). In addition, a functional LOC localizer run (233 volumes) was performed. On both sessions, a high-resolution T1-weighted magnetization prepared rapid gradient echo image was acquired as an anatomical reference (repetition time = 1.9 sec, echo time = 2.51 msec, flip angle = 9°, 192 slices, resolution = 1 mm *isotropic*) as well as a standard fieldmap (Hutton et al., 2002). Preprocessing was performed by using SPM8 (www.fil.ion.ucl.ac.uk/spm) and included realignment, field map correction, smoothing (8-mm FWHM), and spatial normalization to a standard Montreal Neurological Institute template.

Statistical Inference in the fMRI Data Analysis

Statistical analyses of the fMRI data was composed of two steps. At the first level, a general linear model was estimated for both fMRI sessions of each participant. The general linear models contained a regressor for the onset of the response screen and six motion regressors from the realignment analysis. Regressors for the experimental conditions modeled the effects of training (trained exemplars of trained categories, untrained exemplars of trained categories, and untrained categories), ISI (0, 17, and 33 msec), and reward (high reward and low reward). Although all three ISIs were modeled to account for associated variance in the data, only the learning condition (17 msec), as our condition of interest, was considered for all further statistical analyses. Regressors for the experimental conditions as well as one regressor for motor responses were modeled as stick functions convolved with a canonical hemodynamic response function. Contrast maps from each participant were submitted to a second-level random effects analysis, using a flexible factorial ANOVA with repeated measures. We considered effects statistically significant if FWE-corrected p values passed a significance threshold of $<.05$ at the cluster level (denoted as p_{cFWE}; using a cluster-defining threshold of $p < .001$). If the search space was restricted to a small volume, we applied a threshold of $p < .05$ at the voxel level to FWE-corrected p values (denoted as p_{SVC}). If appropriate, we reported effects with uncorrected p values (denoted as p_{unc}) for responses in the homologous contralateral area.

Category-specific Training Effect

Estimated beta maps for trained categories (both trained and untrained exemplars; TC) and untrained categories (UC) of the learning condition were submitted to a repeated-measures group-level ANOVA with factors Training and Session (pretraining/posttraining). To obtain the category-specific training effect, we computed the interaction of Training and Session as follows: $(TC_{post} - TC_{pre}) - (UC_{post} - UC_{pre})$.

Within-category Transfer

Estimated beta maps for untrained exemplars of trained categories (UE) and untrained categories (UC) of the learning condition were submitted to a repeated-measures group-level ANOVA with factors Training and Session (pretraining/posttraining). The within-category transfer was computed analogous to the category-specific training effect: $(UE_{post} - UE_{pre}) - (UC_{post} - UC_{pre})$.

Reward Effect

Because of the exemplar specificity of reward reinforcement at the behavioral level, we restricted the analysis at the group level to trained exemplars of trained categories (TE). Two different analyses were performed. In the first analysis, using repeated-measures ANOVA with factors Session (pretraining/posttraining) and Reward (low reward and high reward), we tested for an interaction of reward and session.

In a second post hoc analysis, we tested for a correlation of the behavioral and neural reward effects. At the neural level, we computed first-level contrasts of the reward effect as follows: $(TE_{high,post} - TE_{high,pre}) - (TE_{low,post} - TE_{low,pre})$. Analogously, we computed the behavioral reward effect for each participant as the performance improvement for high-rewarded trained exemplars minus the improvement for low-rewarded trained exemplars. This allowed us to correlate behavioral reward effects with neural reward effects. Because we were interested in modulatory effects of reward reinforcement on top of the training effects, we restricted this analysis to regions that showed a category-specific training effect. To this end, we generated post hoc spherical ROIs with 12-mm radius, centered at the peak voxels of the category-specific training effect in the bilateral inferior hippocampus.

RESULTS

Category-specific Perceptual Learning

Figure 2A shows the recognition performance for trained exemplars over the course of five sessions. A two-way

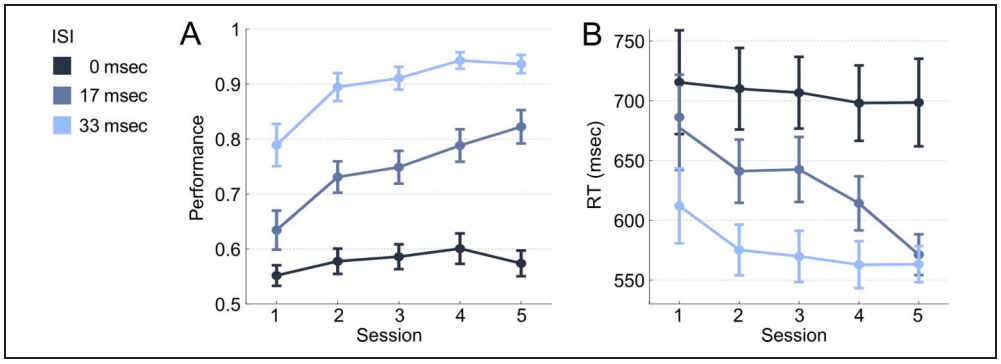


Figure 2. Perceptual learning across five sessions. Error bars represent *SEM*. (A) Learning curves for trained exemplars based on proportion of correct responses (referred to as performance). (B) RTs for trained exemplars.

ANOVA with repeated measures on the proportion of correct responses revealed a significant main effect of Session ($F(4, 68) = 32.8$, $p < .001$) and ISI ($F(2, 34) = 166.5$, $p < .001$) and a significant interaction ($F(8, 136) = 5.9$, $p < .001$). We found similar effects for response times (Figure 2B), which showed a main effect of Session ($F(4, 68) = 3.8$, $p = .008$), a main effect of ISI ($F(2, 34) = 33.9$, $p < .001$), and a Session × ISI interaction ($F(8, 136) = 2.1$, $p = .043$).

To assess whether improvements in object recognition were category specific, we compared the pretraining and posttraining performance of trained and untrained categories. Performance improvements were significantly higher for trained categories, evidenced by a Training (trained categories and untrained categories) × Session (pre, post) interaction for recognition performance ($F(1, 17) = 7.7$, $p = .013$). This demonstrates that the improvements were, to a large extent, specific to trained categories, henceforth referred to as the *category-specific training effect*. This effect was also present for RTs ($F(1, 17) = 18.2$, $p < .001$). However, the training effects were dependent on the ISI as indicated by a three-way interaction effect Session × Training × ISI for both recognition performance ($F(2, 34) = 4.1$, $p = .024$) and RT ($F(2, 34) = 4.1$, $p = .025$). This was owed to our design, in which we intended to stimulate perceptual learning by including both an easy condition (33 msec) and a difficult condition (0 msec; Liu et al., 2012). Because there were no or only marginal category-specific training effects in those conditions—both in terms of recognition performance (0 msec: $F(1, 17) = 0.006$, $p = .94$; 33 msec: $F(1, 17) = 3.5$, $p = .079$) and RT (0 msec: $F(1, 17) = 1.1$, $p = .31$; 33 msec: $F(1, 17) = 3.1$; $p = .094$)—we confined all further analyses of neural and behavioral learning effects to the 17-msec condition. As expected, the 17-msec ISI showed a robust category-specific training effect for recognition performance ($F(1, 17) = 13.3$, $p = .002$; Figure 3A) and RT ($F(1, 17) = 19.3$, $p < .001$). In direct comparison with the other two ISIs, the 17-msec condition showed a stronger category-specific training effect both in terms of recognition performance (0 msec: $F(1, 17) = 9.7$, $p = .006$; 33 msec: $F(1, 17) = 3.7$, $p = .070$) and RT (0 msec: $F(1, 17) = 6.7$, $p = .019$; 33 msec: $F(1, 17) = 4.9$, $p = .041$).

Within-category Transfer

To determine whether the category-specific training effect in the 17-msec condition was limited to trained exemplars or whether it generalized to untrained exemplars of the same category, we directly compared the improvements for untrained exemplars and untrained categories. As shown in Figure 3B, there was an advantage for untrained exemplars of trained categories relative to untrained categories ($t(17) = 2.39$, $p = .029$). Thus, perceptual learning was not limited to the particular object exemplars that were shown during training, indicating a transfer of learning to untrained exemplars within a category and, thus, a true category-specific (rather than exemplar-specific) training effect. Hereinafter, we refer to this effect as *within-category transfer*.

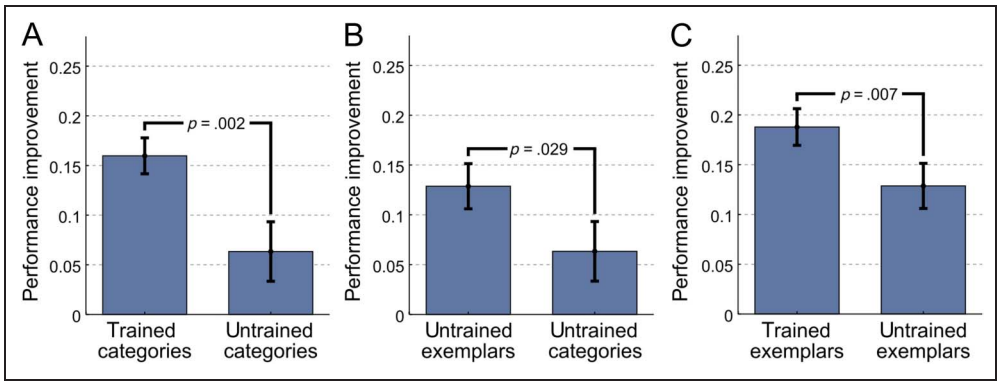


Figure 3. Specificity of perceptual learning in the 17-msec condition. Performance improvements were computed as proportion of correct responses posttraining minus pretraining. Error bars represent *SEM*. (A) Trained and untrained categories. (B) Untrained exemplars of trained categories and untrained categories. (C) Trained and untrained exemplars of trained categories.

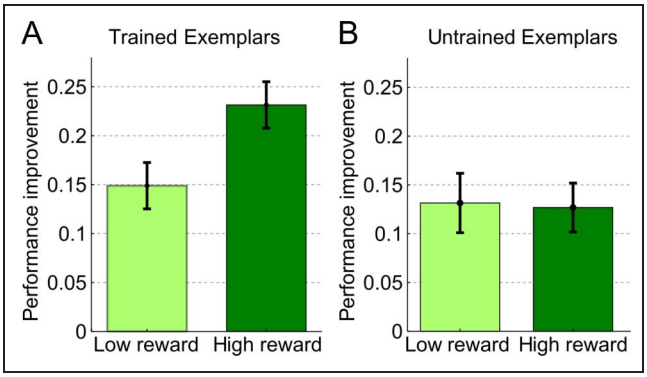


Figure 4. Effect of reward reinforcement on perceptual learning in the 17-msec condition. Error bars represent *SEM*. (A) Trained exemplars associated with low or high reward. (B) Untrained exemplars associated with low or high reward.

Beyond the within-category transfer, there was an additional advantage for trained compared with untrained exemplars ($t(17) = 3.07$, $p = .007$; Figure 3C), indicating a component of exemplar-specific perceptual learning.

Taken together, we found evidence for both category- and exemplar-specific effects of perceptual learning.

The Effect of Reward on Perceptual Learning

During training, one of the two trained category pairs was associated with a high monetary reward (€0.05) upon successful object discrimination, whereas the other respective category pair was associated with a low monetary reward (€0.01). Figure 4A and B shows the improvement for trained and untrained exemplars depending on the reward association. In a two-way repeated-measures ANOVA with factors Training (trained exemplars and untrained exemplars) and Reward (high reward and low reward), the main effect of Reward was not significant ($F(1, 17) = 2.96$, $p = .10$). However, there was a significant main effect of Training ($F(1, 17) = 10.1$, $p = .006$) and a significant Training × Reward interaction ($F(1, 17) = 4.75$, $p = .044$). A post hoc t test revealed a Reward effect specifically for trained exemplars ($t(17) = 2.78$, $p = .013$) but not for untrained exemplars ($t(17) = -0.15$, $p = .883$; see Figure 4A). Thus, only those exemplars of trained categories that were seen during the reinforcement phase (Sessions 2–4) showed an effect of reward reinforcement.

A Signature of Category-related Perceptual Learning in the Hippocampus

The object stimuli, although presented at the threshold of visibility, reliably engaged LOC ($p_{cFWE} < .001$, $t(51) = 13.27$; Figure 5), which is known to play a key role in object processing and object recognition (Grill-Spector, Kourtzi, & Kanwisher, 2001; Malach et al., 1995).

To assess neural correlates of category-specific training over and above unspecific learning effects, we examined the interaction of training and session. We found a significant effect of category-specific training in the inferior hippocampus (peak in the right hippocampus at [22, −28, −12], $t(51) = 5.17$, $p_{cFWE} = .008$; Figure 6A). At a more lenient threshold of $p_{unc.} < .001$, the left inferior hippocampus also showed an isolated cluster of activation (peak at [−26, −32, −12], $t(51) = 3.84$, $p_{unc.} = .0001$). When directly contrasting trained and untrained categories (trained > untrained) in the posttraining session, we likewise found significant effects in the right hippocampus ($t(17) = 6.20$, $p_{cFWE} = .005$, peak at [28, −28, −12]) and, at a more lenient threshold, left hippocampus ($t(17) = 4.17$, $p_{unc.} < .0001$, peak at [−26, −32, −12]), both in proximity to the peaks of the category-specific training effect. Thus, our results are clearly not because of baseline differences.

These bilateral clusters of activation showed considerable overlap with a specific subfield of hippocampus, the subiculum. At a liberal threshold of $p < .01$, most active voxels in both the left and right hippocampus were contained in an anatomical map of the subiculum (Eickhoff et al., 2005; left: 54% overlap; right: 59% overlap). When we considered only the most significant voxels ($p_{unc.} < 10^{-6}$), all 26 surviving voxels belonged to the map of the subiculum.

No other significant category-specific neural effects of training were observed, neither in the functionally localized LOC nor in any other brain region outside the hippocampus. To quantify changes for trained and untrained categories separately and to correlate individual category-specific training effects with behavior, we extracted the

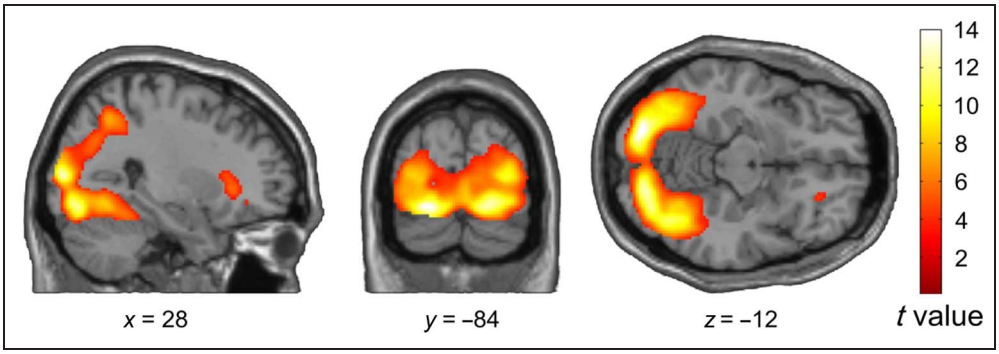


Figure 5. Stimulus-related brain activity in the object recognition task. The presentation of objects in the 17-msec condition reliably engaged the LOC. The depicted t maps represent the average response to the object stimuli across all conditions (against the implicit baseline) and are thresholded at $p_{unc.} < .001$.

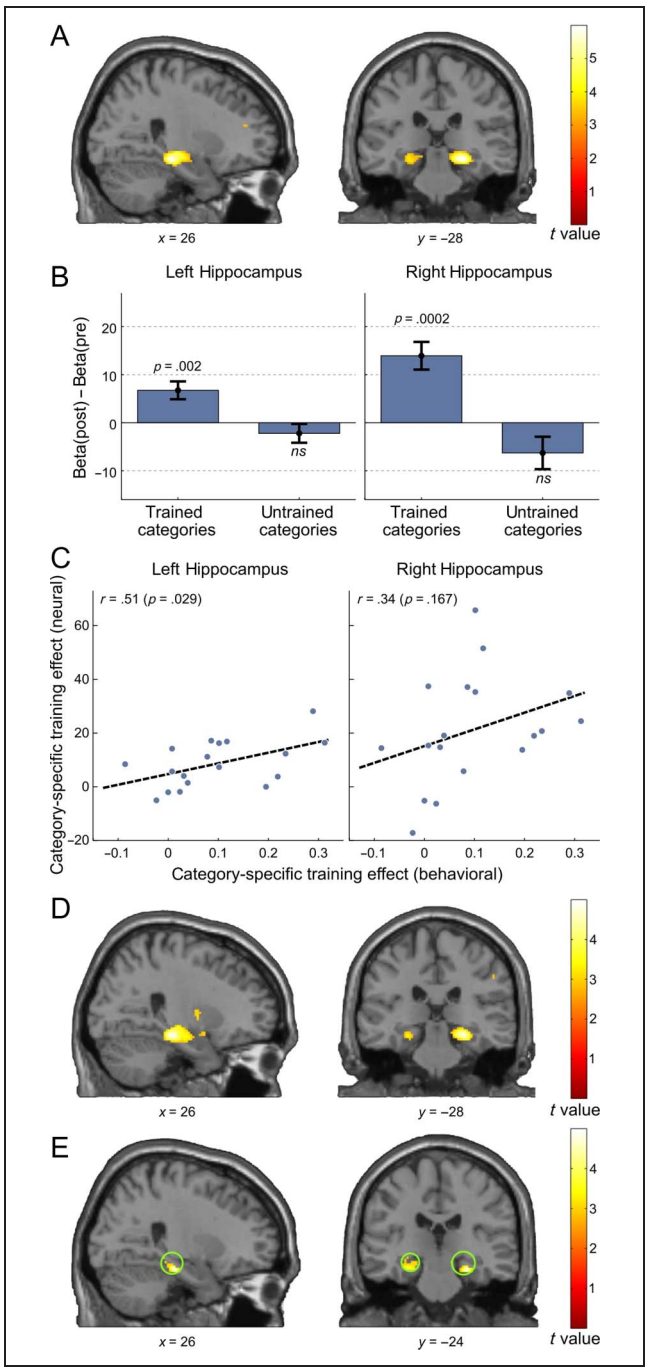


Figure 6. Neural correlates of perceptual training. Whole-brain *t* maps are thresholded at $p < .005$, uncorrected, for display. (A) Category-specific training effect. Voxels in the bilateral hippocampus showed an increase in activation for trained categories relative to untrained categories. (B) Session difference of hippocampal responses (peak voxels) for trained and untrained categories. Error bars represent *SEM*. (C) Correlation of the behavioral category-specific training effect (performance improvement for trained categories minus performance improvement of untrained categories) and the category-specific training effect at the neural level. (D) Within-category transfer. Bilateral hippocampal activation increased also for the subset of untrained exemplars (relative to untrained categories). (E) Neural correlate of reward reinforcement. *T* maps depict voxels that showed a significant correlation of the neural and behavioral reward effects within a spherical ROI (green circles) around the bilateral hippocampal peak voxels of the category-specific training effect.

contrast estimates at the group-level peak voxels in the left and right hippocampus. Figure 6B shows that the interaction effect in both the left and right hippocampus was largely driven by an increase of activity for trained categories, whereas the response to untrained categories remained unchanged.

To assess the functional relevance of hippocampal activation for successful perceptual learning, we probed the association between the category-specific training effect in the hippocampus and the behavioral training effect. We found a significant correlation in the left hippocampus ($r_{\text{Pearson}} = .51, p = .029$). The correlation in the right hippocampus, although in the same direction, was not significant ($r_{\text{Pearson}} = .34, p = .167$; Figure 6C).

As in the behavioral analysis, we next asked whether training effects were limited to those exemplars that were shown during training or whether the effects generalized to untrained exemplars of the same categories. Indeed, when only considering the subset of untrained exemplars, we again found a correlate of category-specific learning in the same location in the inferior hippocampus (peak at [26, −26, −14], $t(51) = 4.91$, $p_{\text{cFWE}} = .007$; Figure 6D). At a more liberal threshold of $p_{\text{unc.}} < .001$, we also found a cluster in the left inferior hippocampus (peak at [−26, −32, −12], $t(51) = 3.44$, $p_{\text{unc.}} = .0005$). Accordingly, a separate analysis directly comparing the training effect for trained and untrained exemplars within the trained category did not yield any significant activation in the hippocampus even at a liberal threshold of $p_{\text{unc.}} < .01$. Thus, also at the neural level, we found evidence for a within-category transfer of training-related changes.

Neural Correlates of Reward-related Perceptual Learning

Given the exemplar specificity of the reward effect at the behavioral level, we constrained the analyses on effects of reward reinforcement to trained exemplars. A whole-brain analysis did not show a significant interaction of reward reinforcement and session, that is, no brain region showed a significant BOLD increase for high-rewarded relative to low-rewarded exemplars. However, as the individual reward effects varied quite substantially at the behavioral level, we went on to investigate whether there were more subtle effects in terms of a correlation between the participants' behavioral and neural reward effects. We focused this post hoc analysis on the hippocampus, where we had found a strong category-specific training effect. To protect against type I error, we performed small-volume correction using a sphere four voxels (12 mm) in radius centered at the bilateral hippocampal peaks of the category-specific training effect. This yielded significant effects both in the left (peak at [−26, −20, −12], $t(16) = 4.35, p_{\text{SVC}} = .026$) and right (peak at [26, −24, −18], $t(16) = 4.6$, $p_{\text{SVC}} = .018$; Figure 6E) hippocampus. Thus, neural activity in the inferior hippocampus was modulated by monetary reward reinforcement during training.

DISCUSSION

We examined perceptual learning of object recognition under challenging perceptual conditions both at the behavioral and neural levels. Our design allowed us to isolate learning-related effects for specific object categories, to dissociate these effects from those related to particular object exemplars, and to assess the effects of reward on category-related perceptual learning.

Behaviorally, we found a marked improvement in object recognition performance over the course of 5 days. Crucially, this improvement was not limited to trained object exemplars but generalized to exemplars within the same object category that had not been presented during training. Thus, the training effect not only was because of the learning of specific low-level features of the particular objects shown during training but also reflected improved recognition based on category-related visual information. Furthermore, recognition performance was enhanced for stimuli associated with high reward during training, relative to stimuli linked to low reward. However, this effect of reward reinforcement did not generalize to those exemplars of the rewarded category that were not shown during training. A particular advantage of our stimulus presentation and task setup is that we can most likely exclude attention as an explanation for the observed reward effect. Participants did not know whether a low- or high-rewarded or unrewarded category was to be expected as an upcoming stimulus. Furthermore, the stimulus presentation itself was extremely brief (17 msec), wherefore it can be considered unlikely that attention or arousal processes could have influenced object recognition.

Analysis of fMRI data acquired before and after training revealed an increase of bilateral inferior hippocampus activation in response to trained categories, compared with untrained categories. Importantly, the inferior hippocampus showed an equally strong effect when the analysis was constricted to untrained exemplars—in analogy to the within-category transfer observed on the behavioral level.

Within-category Transfer of Perceptual Learning and Exemplar Specificity of Reward Reinforcement

Our finding of a category-specific training effect on recognition performance is in line with previous studies reporting enhanced recognition for trained compared with novel objects (Baeck & Op De Beeck, 2010; Furmanski & Engel, 2000; Grill-Spector et al., 2000). Our behavioral findings go beyond previous work by establishing within-category transfer for perceptual learning of object recognition, that is, generalization of the training-induced improvement in object recognition to untrained exemplars of the trained category. Of note, there was still an additional advantage for trained relative to untrained exemplars, suggesting that perceptual learning of object recognition not only is based on category-related features but also involves specific features of the trained object exemplars. Transfer of improved recognition performance to untrained exemplars within a trained category has been reported previously in expertise paradigms (Scott, Tanaka, Sheinberg, & Curran, 2006; Tanaka, Curran, & Sheinberg, 2005). Importantly, our research question was fundamentally different from studies investigating visual expertise in that we were interested in the recognition of known categories under conditions of reduced visibility, whereas the abovementioned expertise studies focused on learning of unfamiliar categories. Moreover, objects were presented only very briefly under conditions of backward masking, whereas expertise tasks typically involve longer and unobstructed stimulus presentations. Thus, the challenge in our task was to quickly extract informative category-related features under challenging viewing conditions rather than the learning of new category boundaries.

In contrast to the observed transfer of category-related learning to untrained exemplars, the additional impact of reward did not generalize to untrained exemplars. This suggests that the influence of reinforcement on perceptual learning is tightly linked to exemplar-specific features that involve early stages of visual processing rather than category-defining features that require higher-level processing. This is in line with earlier work showing that reward-related perceptual learning effects were limited to the eye at which the stimuli were presented, a hallmark of early visual processing (Seitz et al., 2009).

Neural Correlates of Perceptual Learning and Reward Reinforcement in the Hippocampus

The enhancement of activation in the inferior hippocampus in response to trained categories showed, in analogy to our observations at the behavioral level, within-category transfer to untrained exemplars. This clearly suggests that the hippocampus is involved in high-level perceptual learning that is based on complex stimulus features constitutive of object categories. The effect maps to the subiculum, a subfield of the hippocampus that seems to play a role in the retrieval of memories (Eldridge, Engel, Zeineh, Bookheimer, & Knowlton, 2005; Gabrieli, 1997). Recently, subicular activation has been proposed to reflect a "match-enhancement signal" (Dudukovic, Preston, Archie, Glover, & Wagner, 2011; Duncan, Curtis, & Davachi, 2009) caused by a neuronal firing rate increase after presentation of a target that matches an actively retained sample (Otto & Eichenbaum, 1992). It should be noted, however, that the spatial resolution of our fMRI protocol did not allow us to isolate activity to the subiculum exclusively. Other subfields, in particular, the neighboring subfields dentate gyrus and CA1, might as well be involved.

The functional impact of the match-enhancement signal has been interpreted either in terms of an attentional gain enhancement when a stimulus matches an internal representation (Muzzio, Kentros, & Kandel, 2009) or in terms of "pattern completion" whereby partial cues reinstate

information that was present during encoding (Dudukovic et al., 2011).

In the light of these previous findings, the training-related hippocampal activation observed in our study is likely to play a role in perceptual decision-making by means of the aforementioned match-enhancement signal. Applied to our object recognition paradigm, a match-enhancement signal could be generated when the current sensory information about an object matches the learned template of the trained object category. Improved object recognition would then be directly linked to an increased match-enhancement signal, which might serve to generate a more complete percept from the impoverished sensory information through pattern completion (Dudukovic et al., 2011).

Alternatively, hippocampal activation could be a correlate of recognition memory or familiarity. Participants were more frequently exposed to trained categories than to untrained categories (which were only shown in the pretraining and posttraining fMRI sessions). This could have led to a stronger memory imprint of trained categories, in line with previous studies linking the subiculum to the retrieval of memories (Eldridge et al., 2005; Gabrieli, 1997). The observed within-category transfer would then suggest that those stored memories consist of more general, category-related features rather than specific representations of individual exemplars. However, if the hippocampal activation would be based on recognition memory or familiarity, one would expect a stronger category-specific training effect in the 33-msec condition because of higher visibility. However, in fact, as we show in the supplementary Figure S2, the effect in the 33-msec condition, although stronger than the 0-msec effect, is significantly weaker than the effect in the 17-msec condition. This pattern of results in the hippocampus is in close and statistically highly significant correspondence with the behavioral training effects for the different ISIs (supplementary Figure S1) and therefore corroborates the link to training effects as opposed to the frequency of exposure. Furthermore, the supplementary analysis showed that the hippocampus did not display a baseline difference between the 17- and 33-msec conditions in the pretraining session. The stronger effect in the 17-msec condition was therefore mainly based on a stronger posttraining engagement of the hippocampus for trained categories in the 17-msec condition (relative to the 33-msec condition), which fits our interpretation of a supporting role of the hippocampus in object recognition in cases where viewing conditions are particularly challenging.

Evidence from patients with hippocampal damage regarding the role of the hippocampus in high-level perception remains controversial (for a review, see Lee et al., 2012) with an emphasis on scene (Graham et al., 2006; Lee, Buckley, et al., 2005; Lee, Bussey, et al., 2005) and context (Chun & Phelps, 1999) processing. A number of studies also investigated perceptual learning in patients with hippocampal damage (Mundy et al., 2013; Graham et al., 2006; Zaki, Nosofsky, Nenette, & Unverzagt, 2003; Manns & Squire, 2001; Chun & Phelps, 1999; Reed, Squire, Patalano, Smith, & Jonides, 1999; Knowlton & Squire, 1993). Most studies reported normal perceptual learning in patients (but see Mundy et al., 2013 and Zaki et al., 2003). These findings appear to be at odds with our results. However, it is important to note that the aforementioned patient studies focused on a narrow range of perceptual learning tasks, namely, visual search and category learning. Visual search is a special case, as it mainly imposes demands on visual attention rather than fine tuning of visual processing. Category learning, on the other hand, is related to expertise tasks in that participants have sufficient time to scrutinize the stimuli (in the order of several seconds), which, in addition, are presented clearly visible. This is an important difference to our study, in which we investigated object recognition under challenging viewing conditions both in space and time. Our favored interpretation for the functional role of hippocampus in high-level perception, pattern completion, is not probed in visual search or category learning tasks. Those crucial differences in perceptual task demands might explain why a number of studies did not find perceptual learning deficits in patients with hippocampal damage. In addition, a cautionary note regarding early category learning studies was brought up by Zaki et al. (2003), who provided evidence that patients indeed did show deficits in category learning in a more challenging task (involving two categories instead of only a single prototype category as in previous studies). Finally, a more general caveat with regard to lesion studies is the uncertainty to what degree brain function might have been reorganized to compensate for lost abilities after the damage of hippocampal brain tissue (e.g., Pascual-Leone, Bartres-Faz, & Keenan, 1999).

Interestingly, a recent patient study (Aly et al., 2013) found that the hippocampus was important for strength-based (related to the overall configuration of an image), but not state-based (related to local features), visual perception. This indicates that a hippocampal involvement in object recognition is likely based on the overall configural appearance of an object and not on local features, which is in line with the category-related (rather than exemplar-related) hippocampal training effect observed in our study.

With regard to the correlation of hippocampal activity and the behavioral reward effect, it is noteworthy that rodent studies have described the subiculum as an integrator of the raw sensory information and its past emotional connotation (Behr, Wozny, Fidzinski, & Schmitz, 2009; Naber, Witter, & Lopes da Silva, 2000). According to those studies, the subiculum is a recipient of both raw sensory information (from perirhinal and postrhinal cortices) and already processed or modulated version of the same information. Septal areas could modulate sensory input to the subiculum to signal the "emotional connotation in relation to the context of the situation where the organism was stimulated" (Behr et al., 2009). Thus,

the correlation of hippocampal activity and the strength of the reward effect could reflect the learned reward association with respect to the currently perceived stimulus.

Contrary to a previous study (Grill-Spector et al., 2000), we did not find a training-specific increase of the overall BOLD signal in LOC. As an important difference, our participants were trained on a set of specific categories (rather than hundreds of different objects). It is possible that the extensive training with a small set of object categories in our study led to more efficient representations in LOC and thus potentially counteracted gain effects in this brain region. Another possibility is long-term adaptation (van Turennout, Ellmore, & Martin, 2000).

Conclusion

Our study provides evidence for category-specific perceptual learning in object recognition at the behavioral level as well as the neural level. Hippocampal activation was consistent with both the training specificity and the within-category generalization of behavioral improvements. Our data thus suggest an involvement of hippocampal function in the enhancement of higher-level object recognition after perceptual learning. Although the hippocampus and, specifically, the subiculum have primarily been linked to the retrieval of memories, our data indicate a role beyond memory function in terms of a more direct involvement in object recognition. Hippocampal activation likely corresponds to a match-enhancement signal that serves to generate a more complete percept by matching the limited sensory information to an internal object template, thereby making object recognition under constrained viewing conditions more efficient. Furthermore, the observed modulation of hippocampal activity through reward reinforcement suggests that the hippocampus is involved in signaling the behavioral relevance of an object.

Acknowledgments

This study was supported by the DFG Research Training Group (GRK 1589/1) and partly by DFG grants STE 1430/2-1 and STE 1430/6-1.

Reprint requests should be sent to Matthias Guggenmos, Bernstein Center for Computational Neuroscience, Philippstraße 13, Haus 6, 10115 Berlin, Germany, or via e-mail: matthias.guggenmos@bccn-berlin.de.

REFERENCES

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature, 387,* 401–406.

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences, 8,* 457–464.

Aly, M., Ranganath, C., & Yonelinas, A. P. (2013). Detecting changes in scenes: The hippocampus is critical for strength-based perception. *Neuron, 78,* 1127–1137.

Baeck, A., & Op De Beeck, H. P. (2010). Transfer of object learning across distinct visual learning paradigms. *Journal of Vision, 10,* 1–9.

Baldassi, S., & Simoncini, C. (2011). Reward sharpens orientation coding independently of attention. *Frontiers in Neuroscience, 5,* 1–11.

Behr, J., Wozny, C., Fidzinski, P., & Schmitz, D. (2009). Synaptic plasticity in the subiculum. *Progress in Neurobiology, 89,* 334–342.

Chun, M. M., & Phelps, E. A. (1999). Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. *Nature Neuroscience, 2,* 844–847.

Dudukovic, N. M., Preston, A. R., Archie, J. J., Glover, G. H., & Wagner, A. D. (2011). High-resolution fMRI reveals match enhancement and attentional modulation in the human medial temporal lobe. *Journal of Cognitive Neuroscience, 23,* 670–682.

Duncan, K., Curtis, C., & Davachi, L. (2009). Distinct memory signatures in the hippocampus: Intentional states distinguish match and mismatch enhancement signals. *The Journal of Neuroscience, 29,* 131–139.

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., et al. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage, 25,* 1325–1335.

Eldridge, L. L., Engel, S. A., Zeineh, M. M., Bookheimer, S. Y., & Knowlton, B. J. (2005). A dissociation of encoding and retrieval processes in the human hippocampus. *The Journal of Neuroscience, 25,* 3280–3286.

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

Furmanski, C. S., & Engel, S. A. (2000). Perceptual learning in object recognition: Object specificity and size invariance. *Vision Research, 40,* 473–484.

Gabrieli, J. D. (1997). Separate neural bases of two fundamental memory processes in the human medial temporal lobe. *Science, 276,* 264–266.

Graham, K. S., Barense, M. D., & Lee, A. C. H. (2010). Going beyond LTM in the MTL: A synthesis of neuropsychological and neuroimaging findings on the role of the medial temporal lobe in memory and perception. *Neuropsychologia, 48,* 831–853.

Graham, K. S., Scahill, V. L., Hornberger, M., Barense, M. D., Lee, A. C. H., Bussey, T. J., et al. (2006). Abnormal categorization and perceptual learning in patients with hippocampal damage. *The Journal of Neuroscience, 26,* 7547–7554.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research, 41,* 1409–1422.

Grill-Spector, K., Kushnir, T., Hendler, T., & Malach, R. (2000). The dynamics of object-selective activation correlate with recognition performance in humans. *Nature Neuroscience, 3,* 837–843.

Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., & Turner, R. (2002). Image distortion correction in fMRI: A quantitative evaluation. *Neuroimage, 16,* 217–240.

Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science, 262,* 1747–1749.

Lee, A. C. H., Buckley, M. J., Pegman, S. J., Spiers, H., Scahill, V. L., Gaffan, D., et al. (2005). Specialization in the medial temporal lobe for processing of objects and scenes. *Hippocampus, 15,* 782–797.

Lee, A. C. H., Bussey, T. J., Murray, E. A., Saksida, L. M., Epstein, R. A., Kapur, N., et al. (2005). Perceptual deficits in amnesia: Challenging the medial temporal lobe "mnemonic" view. *Neuropsychologia, 43,* 1–11.

Lee, A. C. H., Yeung, L.-K., & Barense, M. D. (2012). The hippocampus and visual perception. *Frontiers in Human Neuroscience, 6,* 1–17.

Liu, J., Lu, Z.-L., & Dosher, B. A. (2012). Mixed training at high and low accuracy levels leads to perceptual learning without feedback. *Vision Research, 61,* 15–24.

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., et al. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences, U.S.A., 92,* 8135–8139.

Manns, J. R., & Squire, L. R. (2001). Perceptual learning, awareness, and the hippocampus. *Hippocampus, 11,* 776–782.

Mundy, M. E., Downing, P. E., Dwyer, D. M., Honey, R. C., & Graham, K. S. (2013). A critical role for the hippocampus and perirhinal cortex in perception learning of scenes and faces: Complementary findings from amnesia and fMRI. *Journal of Neuroscience, 33,* 10490–10502.

Muzzio, I. A., Kentros, C., & Kandel, E. (2009). What is remembered? Role of attention on the encoding and retrieval of hippocampal representations. *The Journal of Physiology, 587,* 2837–2854.

Naber, P. A., Witter, M. P., & Lopes da Silva, F. H. (2000). Networks of the hippocampal memory system of the rat. The pivotal role of the subiculum. *Annals of the New York Academy of Sciences, 911,* 392–403.

Otto, T., & Eichenbaum, H. (1992). Neuronal activity in the hippocampus during delayed non-match to sample performance in rats: Evidence for hippocampal processing in recognition memory. *Hippocampus, 2,* 323–334.

Pascual-Leone, A., Bartres-Faz, D., & Keenan, J. P. (1999). Transcranial magnetic stimulation: Studying the brain–behaviour relationship by induction of "virtual lesions." *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences, 354,* 1229–1238.

Pessoa, L., & De Weerd, P. (Eds.) (2003). *Filling-in: From perceptual completion to cortical reorganization*. New York: Oxford University Press.

Petrov, A. A., Dosher, B. A., & Lu, Z.-L. (2005). The dynamics of perceptual learning: An incremental reweighting model. *Psychological Review, 112,* 715–743.

Petrov, A. A., Dosher, B. A., & Lu, Z.-L. (2006). Perceptual learning without feedback in non-stationary contexts: Data and model. *Vision Research, 46,* 3177–3197.

Reed, J. M., Squire, L. R., Patalano, A. L., Smith, E. E., & Jonides, J. (1999). Learning about categories that are defined by object-like stimuli despite impaired declarative memory. *Behavioral Neuroscience, 113,* 411–419.

Roelfsema, P. R., van Ooyen, A., & Watanabe, T. (2010). Perceptual learning rules based on reinforcers and attention. *Trends in Cognitive Sciences, 14,* 64–71.

Scott, L. S., Tanaka, J. W., Sheinberg, D. L., & Curran, T. (2006). A reevaluation of the electrophysiological correlates of expert object processing. *Journal of Cognitive Neuroscience, 18,* 1453–1465.

Seitz, A. R., & Dinse, H. R. (2007). A common framework for perceptual learning. *Current Opinion in Neurobiology, 17,* 148–153.

Seitz, A. R., Kim, D., & Watanabe, T. (2009). Rewards evoke learning of unconsciously processed visual stimuli in adult humans. *Neuron, 61,* 700–707.

Seriès, P., & Seitz, A. R. (2013). Learning what to expect (in visual perception). *Frontiers in Human Neuroscience, 7,* 668.

Tanaka, J. W., Curran, T., & Sheinberg, D. L. (2005). The training and transfer of real-world perceptual expertise. *Psychological Science, 16,* 145–151.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage, 15,* 273–289.

van Turennout, M., Ellmore, T., & Martin, A. (2000). Long-lasting cortical plasticity in the object naming system. *Nature Neuroscience, 3,* 1329–1334.

Zaki, S. R., Nosofsky, R. M., Nenette, J. M., & Unverzagt, F. W. (2003). Categorization and recognition performance of a memory-impaired group: Evidence for single-system models. *Journal of the International Neuropsychological Society, 9,* 394–406.

SUPPLEMENTARY MATERIALS

Behavioral training effects for the 0ms and 33ms inter-stimulus intervals

The experimental paradigm included, apart from the learning condition with an inter-stimulus interval (ISI) of 17ms, trials with ISIs of 0ms and 33ms in order to foster perceptual learning through a mix of easy and difficult presentations (see methods). For completeness, we here report the data for the 0ms and 33ms condition alongside the 17ms condition.

Based on the learning curves in Fig. 3 of the main text, one would expect no category-specific training effect (CTE) for the 0ms condition (due to the absence of a main effect of session), and a small CTE in the 33ms condition (due to ceiling effects towards the final sessions). This prediction nicely matched the observed pattern of results in Fig. S1 (panel A). We found a marginally significant effect for 33ms ($F(1,17)=3.5$, $p=0.079$) and no effect for 0ms ($F(1,17) = 0.006$, $p=0.94$). In the full repeated measures ANOVA (rmANOVA) with factors session (pre, post), training (trained categories, untrained categories) and ISI (0ms, 17ms, 33ms) we still found a session-by-training interaction ($F(1,17) = 7.7$, $p=0.013$). In addition there was a session-by-training-by-ISI interaction ($F(2,34) = 4.1$, $p=0.024$), reflecting the described differences between the ISIs in regard to the CTE. Next we confined the above full rmANOVA to untrained exemplars corresponding to the within-category transfer (Fig. S1, panel B). As in the case of the CTE, the interaction of session and training remained significant ($F(1,17) = 5.0$, $p=0.039$) in an ANOVA across including all ISIs. In an rmANOVA with factors session, training (trained exemplars, untrained exemplars) and ISI there were trends for a session-by-training interaction ($F(1,17) = 2.7$, $p=0.11$) and a session-by-training-by-ISI interaction ($F(2,34) = 3.1$, $p=0.060$).

When tested separately, there was no significant difference between the performance improvement of trained and untrained exemplars (Fig. S1, panel C) both in the 0ms condition (T(17) = -0.86, p=0.40, two-tailed t-test) and the 33ms condition (T(17)= 0.99, p=0.34, two-tailed t-test).

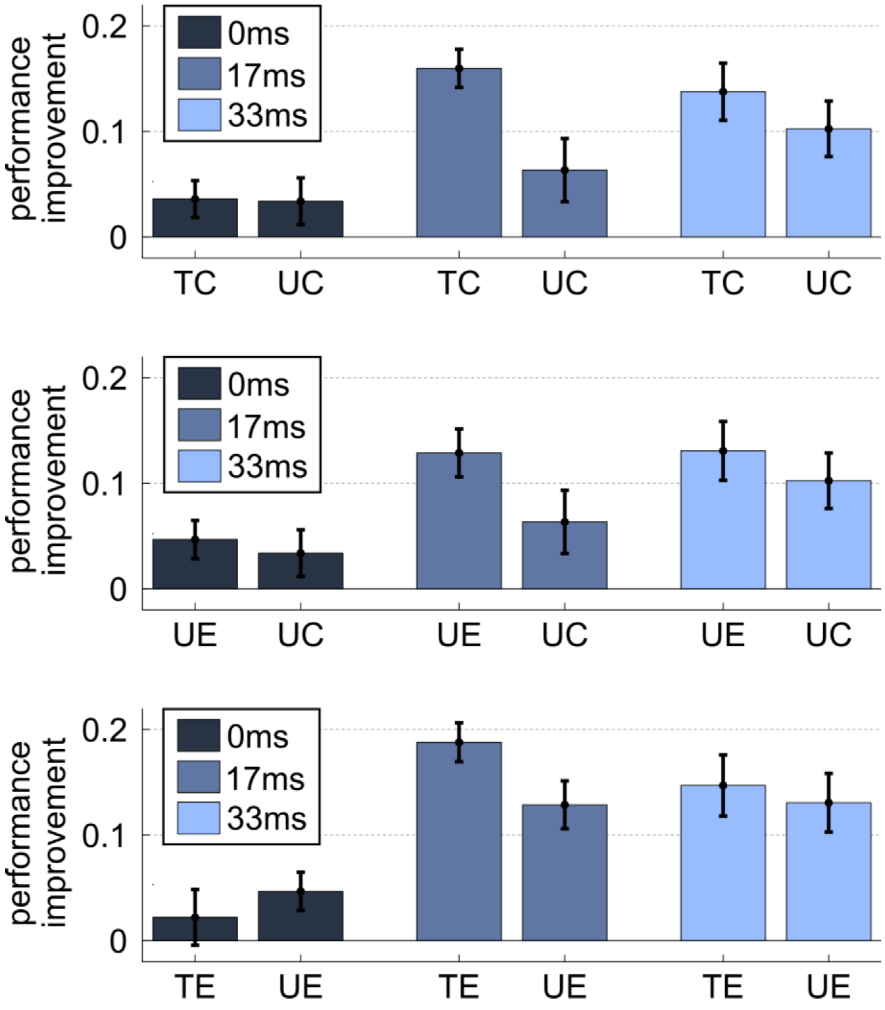


Figure S1. Specificity of perceptual learning for all inter-stimulus intervals. Performance improvements were computed as proportion correct responses post-training minus pre-training. Error bars represent SEM. (A) Trained and untrained categories. (B) Untrained exemplars of trained categories and untrained categories. (C) Trained and untrained exemplars of trained categories.

Neural category-specific training effect for the 0ms and 33ms inter-stimulus intervals

To ensure a fair comparison between ISIs regarding the category-specific training effect, we first computed a whole-brain rmANOVA with factors session (pre, post), training (trained categories, untrained categories) and ISI (0ms, 17ms, 33ms). The whole-brain T-maps in Fig. S2 (panel A) depict the interaction of training and session across all three ISIs. The results are similar, albeit weaker compared to the 17ms condition alone. The peaks in the right (x=20, y=-26, z=-10; T=3.59, p<0.001, uncorrected) and left (x=-28, y=-30, z=-12; T=2.37, p=0.009, uncorrected) hippocampus are in close proximity to the peaks of the 17ms condition alone. Panel B of Fig. S2 shows the contrast estimates for the CTE in those peak voxels. In direct comparison with the 0ms and the 33ms condition, the 17ms condition shows a stronger effect both in the left (0ms: T(17)=2.5, p=0.022; 33ms: T(17)=2.0, p=0.057) and the right (0ms: T(17)=3.3, p=0.005; 33ms: T(17)=2.6, p=0.018) hippocampus. Overall the pattern of results is in very good correspondence to the behavioral CTE with no effect for 0ms, a moderate effect for 33ms and the strongest effect for 17ms. We quantified this correspondence by calculating the correlation between the neural and the behavioral CTE across ISIs for each participant. This way we obtained one correlation coefficient per subject. The average correlation coefficient was positive and highly significant in both left (mean ± SEM, 0.58 ± 0.14, p<0.001, two-tailed t-test) and right hippocampus (mean ± SEM, 0.54 ± 0.14, p=0.003, two-tailed t-test), indicating that the strength of the neural effect of a particular ISI reflected the strength of the behavioral effect.

Notably, the fact that the hippocampal effect in the 33ms condition was significantly weaker than the 17ms condition is incompatible with a familiarity explanation for the

hippocampus results. Given the enhanced visibility of objects in the 33ms condition, a familiarity-based account would – if at all – predict a stronger CTE in the hippocampus. Note that the stronger effect in the 17ms relative to the 33ms condition is not due to a baseline difference between the two. We compared the hippocampal pre-training responses to trained categories between the 17ms and the 33ms condition (which is the relevant comparison for the CTE) and additionally assessed the pre-training main effect of ISI in a training(trained/untrained categories)-by-ISI(17ms/33ms) rmANOVA (since there should be no difference between trained and untrained categories in the pre-training session). Neither did the trained categories alone (peak in right hippocampus: p=0.61, t(17)=0.52; left: p=0.47, t(17)=-0.75; two-tailed t-test) differ between 17ms and 33ms in the pre-training session, nor was there a significant main effect of ISI in the statistically more powerful rmANOVA (right: p=0.25, F(1,17) = 1.41; left: F(1,17)=0.16, p=0.69).

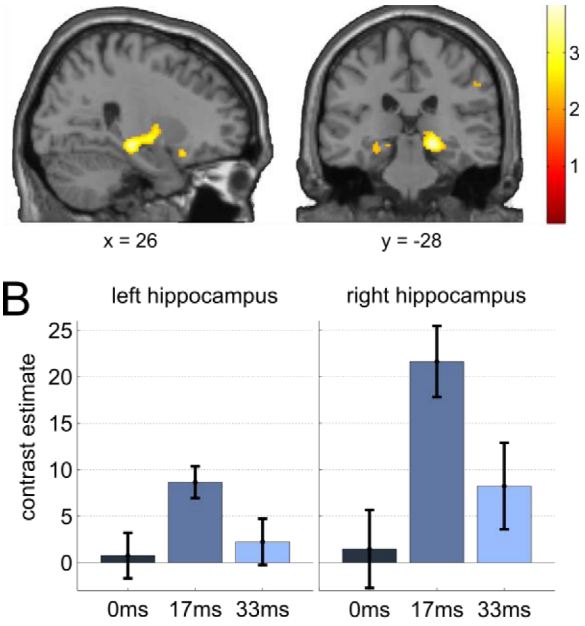


Figure S2. Neural category-specific training effect for all inter-stimulus intervals. (A) Whole-brain T-map for the interaction of session and training based on the full repeated-measures ANOVA with factors session, training and ISI. T-maps are thresholded at p<0.01, uncorrected, for display. (B) Contrast estimates for 0ms, 17ms and 33ms extracted from the peak voxels of left and right hippocampus.

6.2 Non-holistic coding of objects in lateral occipital complex with and without attention

Journal: NeuroImage

Acceptance date: 05 December 2014

URL: http://dx.doi.org/10.1016/j.neuroimage.2014.12.013

NeuroImage 107 (2015) 356–363

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Non-holistic coding of objects in lateral occipital complex with and without attention

Matthias Guggenmos [a,b,*], Volker Thoma [c], Radoslaw Martin Cichy [d], John-Dylan Haynes [a], Philipp Sterzer [a,b,1], Alan Richardson-Klavehn [e,f,1]

[a] Bernstein Center for Computational Neuroscience, Berlin, Germany
[b] Visual Perception Laboratory, Charité Universitätsmedizin, Berlin, Germany
[c] School of Psychology, University of East London, London, UK
[d] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA
[e] Department of Neurology, Otto von Guericke University, Magdeburg, Germany
[f] Leibniz Institute for Neurobiology, Magdeburg, Germany

ARTICLE INFO

Article history:
Accepted 5 December 2014
Available online 12 December 2014

Keywords:
Object recognition
Attention
Lateral occipital complex
Multivariate pattern analysis
fMRI

ABSTRACT

A fundamental issue in visual cognition is whether high-level visual areas code objects in a part-based or a view-based (holistic) format. Previous behavioral and neuroimaging studies that examined the viewpoint invariance of object recognition have yielded ambiguous results, providing evidence for either type of representational format. A critical factor distinguishing the two formats could be the availability of attentional resources, as a number of priming studies have found greater viewpoint invariance for attended compared to unattended objects. It has therefore been suggested that the activation of part-based representations requires attention, whereas the activation of holistic representations occurs automatically irrespective of attention. Using functional magnetic resonance imaging in combination with a novel multivariate pattern analysis approach, the present study probed the format of object representations in human lateral occipital complex and its dependence on attention. We presented human participants with intact and half-split versions of objects that were either attended or unattended. Cross-classifying between intact and split objects, we found that the object-related information coded in activation patterns of intact objects is fully preserved in the patterns of split objects and vice versa. Importantly, the generalization between intact and split objects did not depend on attention. We conclude that lateral occipital complex codes objects in a non-holistic format, both in the presence and absence of attention.

© 2014 Elsevier Inc. All rights reserved.

Introduction

A hallmark of human object perception is the recognition of objects despite variations in their exact appearance. In line with this characteristic, object representations in high-level visual brain areas have been found to generalize across changes in size, position or orientation (Eger et al., 2008; Grill-Spector et al., 1999). Yet, the specific representational code realizing such invariant representations is still largely unknown. In particular, one central question is whether an object is coded as a collection of *parts* (Biederman, 1987; Hummel and Biederman, 1992; Marr and Nishihara, 1978) or in a *view-based* format (Edelman and Bülthoff, 1992; Poggio and Edelman, 1990; Tarr and Pinker, 1989).

Part-based models propose that objects are encoded in terms of their constituent parts, the representations of which are independent of each other and dynamically bound together. Neurons that are tuned to a particular object part could therefore respond to the object part appearing in different configurations or views, allowing for robust object recognition across various manipulations, such as translation across the visual field, size changes and left–right reflection (Biederman, 1987; Hummel and Biederman, 1992; Hummel, 2001). By contrast, view-based models propose that objects are recognized by matching the incoming sensory information to stored views (Edelman and Bülthoff, 1992; Poggio and Edelman, 1990; Tarr and Pinker, 1989). View-based representations are holistic, as the parts of an object are not represented independent of each other and have fixed relative positions (static binding). Under a view-based scheme neurons respond most strongly if objects are presented in learned views or configurations. Nevertheless, recognition of objects in varying views is thought possible by storing many views of an object (Bülthoff and Edelman, 1992; Olshausen et al., 1993; Poggio and Edelman, 1990; Tarr and Gauthier, 1998; Tarr, 1995; Ullman, 1998), interpolating across these views (Logothetis et al., 1994; Poggio and Edelman, 1990; Ullman, 1989) or by a distributed neural representation across view-tuned neurons (Perrett et al., 1998).

Behavioral evidence on the format of object representations is equivocal, supporting both view-based (Edelman and Bülthoff, 1992;

* Corresponding author at: Bernstein Center for Computational Neuroscience, Philippstraße 13, Haus 6, 10115 Berlin, Germany.
E-mail address: matthias.guggenmos@bccn-berlin.de (M. Guggenmos).
[1] Contributed equally to this study.

http://dx.doi.org/10.1016/j.neuroimage.2014.12.013
1053-8119/© 2014 Elsevier Inc. All rights reserved.

Murray, 1999; Tarr and Pinker, 1989) and part-based representations (Biederman and Cooper, 1991; Biederman and Gerhardstein, 1993). Neuroimaging research, too, has sought to establish which format of representation underlies object recognition. Studies using functional magnetic resonance imaging (fMRI) show that blood oxygen level-dependent (BOLD) signals in various ventral visual stream regions, such as in lateral occipital and inferior temporal cortices, tend to decrease when an object is shown repeatedly, and found that this repetition suppression (Grill-Spector et al., 2006) is greatest when the repeated view of an object is identical to the original orientation, but decreases with the amount of view change (Andresen et al., 2009; Ewbank et al., 2005; Gauthier et al., 2002). However, in support for part-based representations, other fMRI studies have shown that the ventral stream is largely insensitive to the deletion of local image features or changes in image format (grayscale image vs. line drawing), as long as the individual object parts are present (Hayworth and Biederman, 2006; Kourtzi and Kanwisher, 2000).

Importantly, the representational format might be dependent on the absence or presence of attention. Attended visual objects exhibit robust repetition-priming effects even when their mirror-reflected (Stankiewicz et al., 1998) or half-split (Thoma and Henson, 2011; Thoma et al., 2004) versions are presented as prime stimuli, suggesting a part-based representation. However, when the same prime objects are unattended, visual priming is still found for objects presented in the same view, but completely abolished after view changes (see Thoma and Davidoff, 2007, for a brief review). Hummel (2001) therefore proposed a hybrid model, in which part-based representations are established with attention, whereas view-based representations are automatically activated irrespective of attention.

Inspired by these previous studies and theoretical considerations, the present fMRI study examined the representational format of objects in high-level visual cortex and its dependence on attention. Objects were presented in either an intact or a split configuration (Fig. 1B) and were either attended or unattended. The half-split manipulation, while preserving the constituent object parts, distorted the holistic image in a way that cannot be recovered by the aligning processes of view-based models (Hayward et al., 2010; Thoma et al., 2004). To prevent verbalization of the attended object as a confounding factor, we used a non-semantic attention task, in which participants detected brightness changes on either the object (attended condition) or a contralaterally presented noise stimulus (unattended condition). We reasoned that only if objects are coded as part-based, non-holistic representations, should activation patterns of split objects be informative about those of intact objects. Moreover, if attention was necessary for part-based representations, this configural invariance of object representations should only be observed for attended, but not for unattended objects.

To this end, we used a novel multivariate approach, in which we trained a support vector machine classifier to discriminate between activation patterns of intact objects and tested its predictive capacity for activation patterns of split objects and vice versa. The rationale was that successful generalization between activation patterns of intact and split objects is indicative of a non-holistic format of object representations. Our multivariate approach represents an important advance compared with previous repetition suppression studies, because of mounting evidence that high-level visual areas code objects in a distributed fashion across multiple neuronal populations (Haxby et al., 2001; Rice et al., 2014). In particular, different configurations of an object might activate identical neuronal populations and the difference between configurations only emerges at the pattern level as a distinct weighting of each population. Multivariate methods are able to pick up on these object- or view-specific multivoxel fingerprints, whereas repetition suppression—as a univariate technique—misses out on such pattern-related information. We focused our analyses on the LOC, given a large body of evidence supporting its pivotal role in object processing (Grill-Spector et al., 1998; Malach et al., 1995) and object recognition (Grill-Spector et al., 2000).

Materials and methods

Participants

Eighteen healthy participants (11 female, mean age ± SEM, 23.4 ± 0.8 years) participated in the experiment for payment after giving written informed consent. The study was conducted according to the declaration of Helsinki and approved by the local ethics committee.

Experimental design

The experimental design comprised the factors configuration (intact, split) and attention (attended, unattended) as factors of interest, as well as object (camera, watering can, chair) and side of presentation (left, right) as factors of no interest. Within each of eight experimental runs, an object appeared in four trials in each attention conditions (in two trials per side of presentation). The order of presentation was randomized across the 48 trials of each run.

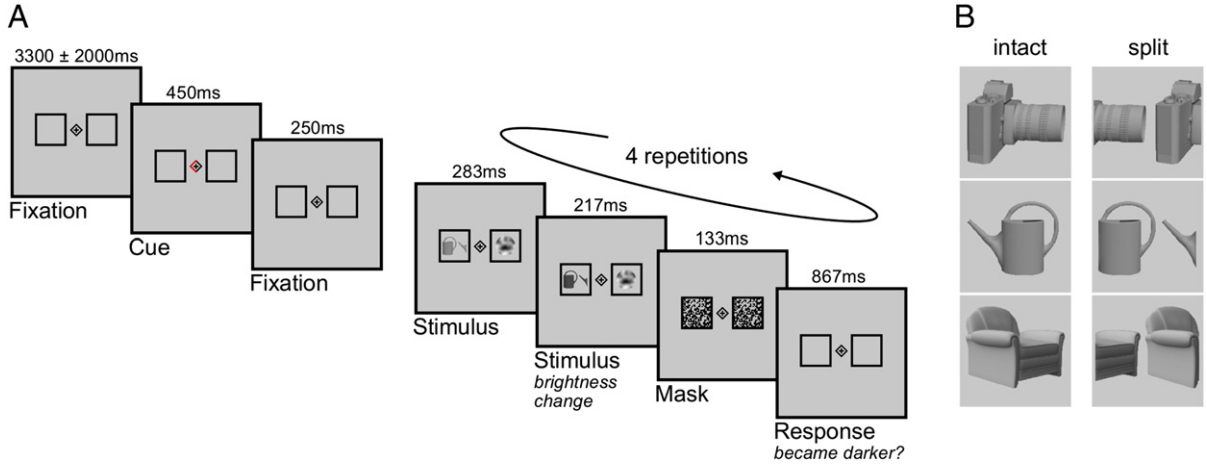


Fig. 1. Experimental procedures and stimuli. A. In each trial a cue indicated the side to which attention should be directed. Subsequently, four repetitions of the stimulus–response phase appeared, during each of which participants had to detect a decrease in brightness of either the object (attended condition) or the noise stimulus (unattended condition). B. The stimulus set consisted of three objects in an intact and half-split configuration.

Experimental procedures

A trial (Fig. 1A) started with a blank fixation screen for 3300 ms ± 2000 ms, after which one half of a central black fixation diamond turned red, indicating the side to which attention should be directed. After a fixed interval (250 ms), four repetitions of the stimulus–response phase appeared. Each stimulus–response phase lasted 1500 ms and comprised the presentation of the stimulus screen (500 ms), a pattern mask (133 ms) and the default screen (867 ms). An intact or split object appeared on one side of the fixation cross and a noise stimulus at the same offset on the other side of the stimulus screen. A brightness change occurred 283 ms after stimulus onset simultaneously on both the object and the noise stimulus, such that they became independently and randomly either darker or lighter. Participants were instructed to press a button on the response box when the stimulus on the cued side became darker. The cued stimulus could either be an intact or split image version of an object (*attended condition*) or the noise stimulus (*unattended condition*). Responses were counted as valid within a time window of 1000 ms after stimulus offset. In each repetition of the stimulus–response phase, the same object was shown at the same position. The noise stimulus, while also presented at the same position, was randomly generated for each repetition.

To independently identify object-responsive regions of lateral occipital complex (LOC) in each participant (Malach et al., 1995), we conducted a localizer run with five blocks of intact objects, five blocks of split objects and 10 blocks of grid-scrambled versions of the objects in randomized order. Blocks lasted for 15.8 s during which 20 images were presented for 600 ms each, followed by 200 ms blank screen. Pairs of identical objects were shown left and right of fixation, equaling the configuration of the main experiment in eccentricity and size. Participants performed a one-back task, in which they had to indicate via button press whenever the same stimulus display appeared twice in a row.

Stimuli

Stimuli were generated with Psychophysics Toolbox 3 (Brainard, 1997; Pelli, 1997) and projected with a Sanyo LCD projector at 60 Hz. The stimulus set consisted of three grayscale objects (camera, watering can, chair) based on realistic three-dimensional models presented either intact or half-split (Fig. 1B). The objects were selected for representing non-overlapping man-made categories to increase the discriminability of evoked neuronal activation patterns. Split versions were generated by relocating the two halves of an original image to the opposite side of the canvas. The noise stimuli matched the objects in terms of spatial extent and complexity to ensure that there would be no performance difference. They were randomly generated for each trial by sampling a 9 by 9 random binary matrix, scaling the matrix to 216 by 216 pixels, applying a low-pass filter with a cut-off frequency of 0.02/pixel and cropping pixels outside a circle of 216 pixels diameter. This procedure resulted in circular grayscale stimuli with randomly distributed smooth patches. Both the objects and the noise stimuli were scaled to grayscale RGB values between 50 and 205. To generate brightness changes, the underlying RGB histograms were shifted up or down by 50 (the image background remained constant with an RGB value of 200). Intact and split objects as well as the noise stimuli subtended 3.81 x 3.81 degrees of visual angle and were presented at a horizontal offset of 3.84 degrees of visual angle relative to the central fixation cross. The pattern masks were generated for each trial by sampling an 18 by 18 random binary matrix and scaling the matrix to 216 by 216 pixels.

Eyetracking

Eyetracking data were successfully collected in 16 of 18 subjects using an infrared video eyetracking system (iView XTM MRI 50Hz, SensoMotoric Instruments, Teltow, Germany). As a measure of fixation reliability, we computed the percentage of recorded eye gaze positions within a radius of 1.93 degree visual angle around the center of the fixation cross. This radius corresponded to the eccentricity of the inner edges of the two stimulus-containing boxes (see Fig. 1A).

FMRI data acquisition and preprocessing

FMRI data were acquired on a 3-T Siemens Trio (Erlangen, Germany) scanner using a gradient echo planar imaging (EPI) sequence and a 12-channel head-coil. We recorded eight experimental runs of 214 whole-brain volumes each (TR = 2 s, echo time (TE) 25 ms, flip angle 78°, 33 slices, 3 mm isotropic resolution, interslice gap 0.75 mm). The LOC localizer comprised 242 volumes. In addition, a high-resolution T1-weighted image was acquired (TR = 1.9 s, echo time (TE) 2.51 ms, flip angle 9°, 192 slices, resolution 1 mm isotropic). Preprocessing was performed using SPM8 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London) and included realignment and smoothing with an 8 mm Gaussian kernel. All main analyses were performed in native subject space.

Region of interest procedures

Our main region of interest (ROI) was LOC. To anatomically constrain LOC, which stretches from lateral occipital cortex to posterior fusiform gyrus (Grill-Spector et al., 1999), we generated a bilateral composite mask of the inferior occipital cortex, middle occipital cortex and the posterior half of the fusiform gyrus (derived from the AAL Atlas, Tzourio-Mazoyer et al., 2002). The LOC ROI was defined as the intersection of the anatomical mask and the functional localizer based on the group-level T-contrast *intact + split > scrambled* at a significance level of $p < 0.05$ (family-wise error (FWE) corrected at the whole-brain level). Additionally we created separate ROIs for two subregions of LOC, lateral occipital cortex (LO; corresponding to the inferior and middle occipital anatomical masks) and posterior fusiform gyrus (pFus; posterior fusiform gyrus mask), based on previous reports regarding a possible functional dissociation between the two (Cichy et al., 2013; Grill-Spector et al., 2001). The V1 ROI was defined as the intersection of Brodmann's area 17 (derived from the SPM Anatomy toolbox; Eickhoff et al., 2005) and the functional localizer based on the group-level t-contrast *intact + split + scrambled > implicit baseline* at a significance level of $p_{FWE} < 0.05$. The V1 ROI was based on a mask for Brodmann's area 17 derived from the SPM Anatomy toolbox (Eickhoff et al., 2005). All ROIs were reverse-normalized to native subject space.

FMRI data analysis

First-level general linear models (GLMs)

For each participant we estimated a GLM with separate experimental regressors for the factors configuration (split, intact), attention (attended, unattended), object (camera, watering can, chair) and side of presentation (left, right). Onsets of the experimental regressors were set to the beginning of the stimulus–response phase, and they were modeled as stick functions and convolved with a canonical hemodynamic response function. Further, six motion parameters from the realignment preprocessing step were included as regressors-of-no-interest.

The GLM for the functional localizer comprised regressors for intact objects, split objects and scrambled objects and six motion parameters. The experimental regressors were modeled as boxcar functions with durations equal to the block lengths (15.8 s) and convolved with a canonical hemodynamic response function.

Multivariate analyses

The estimated beta images of the GLM provided the basis for support vector machine (SVM) classification. SVM classification was performed

using *The Decoding Toolbox* (Hebart et al., 2014) with a linear C-SVM and a fixed cost parameter (c = 1).

We first performed a searchlight analysis (Kriegeskorte et al., 2006), in which a sphere with a radius of 4 voxels was centered at each voxel of the brain and decoding was based on the voxels within each sphere. A leave-one-run-out cross-validation procedure was used, such that in each fold the classifier was trained on the beta maps of seven runs and tested on the left out eighth run. The resulting decoding accuracies were assigned to the center voxel. We performed decoding separately between the three pairs of objects (camera-can, camera-chair, can-chair) in each of the four experimental conditions (intact attended, intact unattended, split attended, split unattended) and both sides of presentation (left, right). After averaging across object pairs and sides, we obtained information maps for each subject and experimental condition, which were subsequently normalized to a common template for group-level statistical inference.

The main analyses were based on ROI decoding (Haynes and Rees, 2005; Kamitani and Tong, 2005), in which the voxels of a given ROI in native space were used for classification. ROI decoding followed the same cross-validation procedure as detailed for the searchlight analysis. In addition, we used a nested feature selection procedure in order to select the most stimulus–responsive voxels. Thus, for each of the seven runs within a fold of the cross-validation procedure, voxels were ranked according to the magnitude (beta value) of the stimulus-related responses in the respective six other runs. Stimulus-related responses were derived from the t-contrast *all conditions* > *implicit baseline*.

We refer to decoding analyses in which training and testing was performed within the same configuration (e.g. training on intact objects, testing on intact objects) as *within-configuration* decoding. In the *cross-configuration* analysis we trained the classifier to discriminate between intact object categories and tested on split object categories and vice versa, and then averaged across train-test directions. In the *cross-attention* analysis the classifier was trained to discriminate between attended object categories and tested on unattended object categories and vice versa. We performed *cross-attention* classification both under the *within-* and the *cross-configuration* decoding scheme as described above.

For statistical inference we performed two-sided t-tests and repeated-measures ANOVAs. T-tests for decoding accuracies were tested against the null hypothesis of a chance level decoding performance of 50%.

Results

Behavioral results and fixation control

Participants detected and reported brightness changes of the objects and the noise stimuli highly accurately (performance > 98%), indicating that they focused their attention on the correct stimulus in each condition. On average, 98.3 ± 0.8% of recorded eye gaze positions were within the fixation area, demonstrating that the participants maintained fixation throughout the experiment.

Decoding of objects

Initially, we performed a searchlight analysis to identify brain areas that processed information about object categories (Fig. 2). We found above-chance classification for split and intact object in both the attended and the unattended condition in areas overlapping with the three a priori defined ROIs (LO, pFus,V1; peaks in the three ROIs were significant for all conditions at $p < 0.01$, FWE-corrected for small volumes). We did not observe significant above-chance decoding beyond the predefined ROIs. We therefore performed all subsequent analyses in those ROIs. Further, since we did not find any differences between pFus and LO in any of the following analyses, we present the results for a composite mask of pFus and LO (LOC ROI).

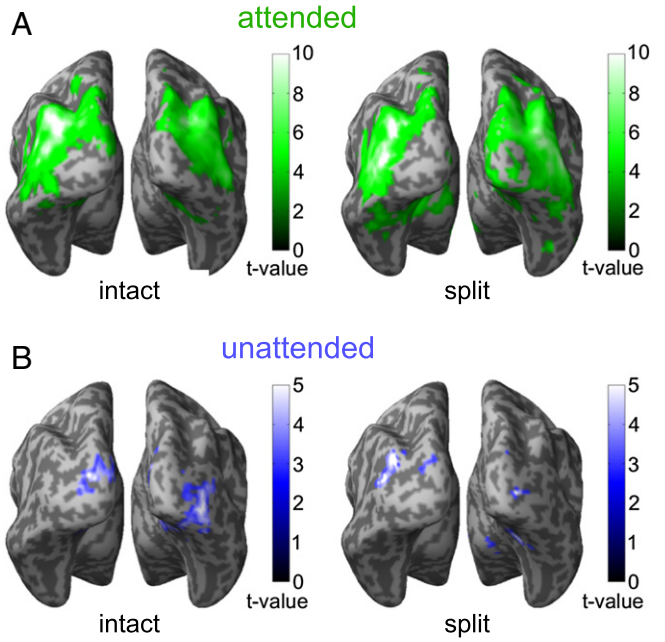


Fig. 2. Searchlight analysis results for intact and split objects based on a *within-configuration* decoding procedure. Whole-brain information maps are represented as T-maps indicating the statistical significance of voxel-wise decoding accuracies against the chance-level decoding accuracy of 50%. The T-maps are thresholded at $p < 0.005$, uncorrected, for illustration. A. Attended objects. B. Unattended objects.

ROI decoding accuracies were significantly above chance in all conditions in LOC and V1 (Table 1). Repeated-measures ANOVA (rmANOVA) with the factors attention and configuration showed a main effect of attention in LOC, ($F(1,17) = 39.4$, $p < 0.001$), but not in V1 ($F(1,17) = 2.0$, $p = 0.17$). The strong main effect of attention underscored the effectiveness of the attentional manipulation. No other main effects or interactions were significant in either ROI (all $p > 0.1$).

Generalization between intact and split object representations

The critical test for distinguishing formats of object representations in LOC was whether the classifier generalized between intact and split objects. For this purpose we conducted a *cross-configuration* analysis, training the classifier on intact objects and testing on split objects, and vice versa. We found that *cross-configuration* decoding of objects was significant in both the attended (67.3% accuracy, $t(17) = 8.9$, $p < 0.001$) and unattended condition (55.5% accuracy, $t(17) = 3.7$, $p = 0.002$; see Fig. 3A). This generalization clearly suggests a non-holistic format of object representations in LOC.

Table 1
Basic ROI decoding results in LOC and V1 for intact and split objects and for both the attended and unattended condition.

		Accuracy	t(17)	p
Intact attended	LOC	66.9%	9.8	.00000002
	V1	63.3%	8.9	.00000008
Split attended	LOC	67.9%	6.8	.000003
	V1	62.4%	8.0	.0000004
Intact unattended	LOC	54.1%	2.2	.043
	V1	59.3%	5.7	.00003
Split unattended	LOC	55.9%	3.6	0.002
	V1	61.6%	5.2	0.00007

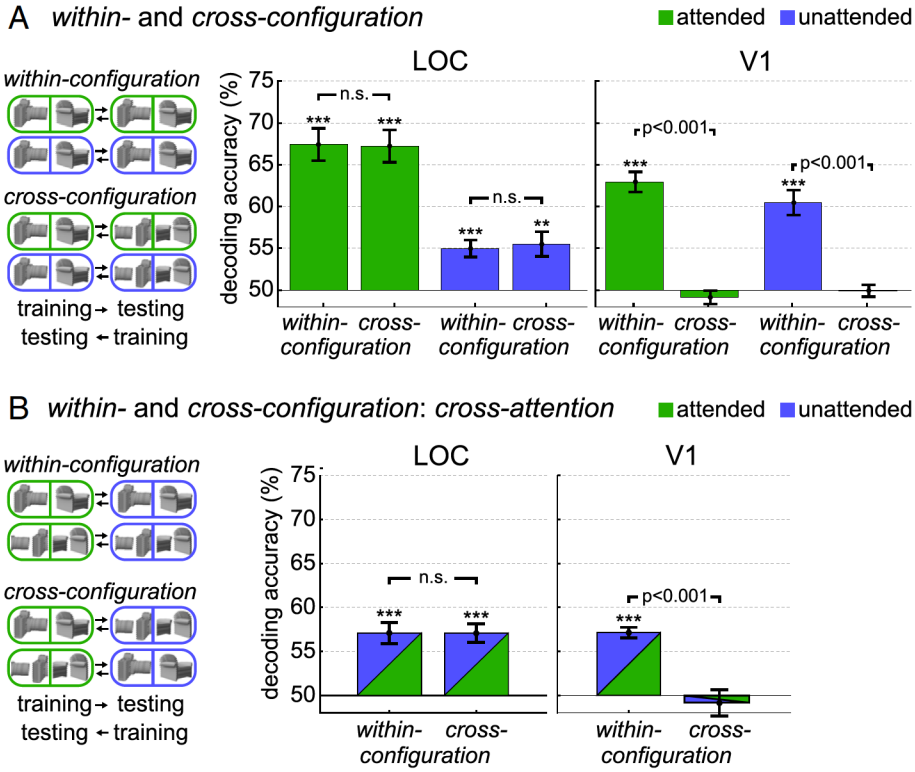


Fig. 3. *Within-* and *cross-configuration* decoding in LOC and V1. A. *Within-attention* decoding scheme. B. *Cross-attention* decoding scheme. Error bars represent SEM. P-values are based on two-tailed paired t-tests. Stars represent the significance of decoding accuracies based on two-tailed t-tests against the chance-level decoding accuracy of 50%: **p < 0.01 ***p < 0.001.

To assess how *cross-configuration* decoding compared with *within-configuration* decoding, we conducted a rmANOVA with factors decoding scheme and attention. There was a main effect of attention ($F(1,17) = 43.1$, $p < 0.001$), but neither a main effect of decoding scheme ($F(1,17) = 0.01$, $p = 0.91$) nor an attention-by-decoding scheme interaction ($F(1,17) = 0.1$, $p = 0.70$). Thus the classifier could equally well predict intact and split objects, irrespective of whether it was trained on intact or split objects (configural invariance).

In V1, by contrast, we found a main effect of decoding scheme ($F(1,17) = 71.8$, $p < 0.001$), such that *within-configuration* decoding was superior to *cross-configuration* decoding. There was neither a main effect of attention ($F(1,17) = 0.33$, $p = 0.57$) nor an attention-by-decoding scheme interaction ($F(1,17) = 3.6$, $p = 0.08$). A rmANOVA with the additional factor region revealed a region-by-decoding scheme interaction ($F(1, 17) = 32.0$, $p < 0.001$), showing that the observed configural invariance was present in LOC, but not V1 (Fig. 3A). The post-hoc probability to detect a main effect of decoding scheme in LOC of the same effect size as in V1 was 0.978.

To test whether the generalization between intact and split objects in LOC would also hold if they were presented in different hemifields, we repeated the above analyses in a cross-hemifield decoding scheme. The classifier was trained on stimuli in one hemifield and tested on stimuli in the other hemifield, both under a within- and cross-configuration scheme. As shown in Inline Supplementary Fig. S1A, cross-configuration decoding was significant for attended objects (54.8% accuracy, $t(17) = 5.1$, $p < 0.001$) and at the same level as within-configuration decoding (54.6% accuracy, $t(17) = 5.6$, $p < 0.001$). The analysis demonstrates that the finding of complete cross-configuration generalization persists for high-level neuronal populations with receptive fields encompassing an area of at least 5.7 degree visual angle left and right of fixation (11.4 degree in total). Due to insufficient power, no statement can be made about unattended objects (Inline Supplementary Fig. S1B).

Inline Supplementary Fig. S1 can be found online at http://dx.doi.org/10.1016/j.neuroimage.2014.12.013.

Generalization between attended and unattended object representations

Our finding that representations of both attended and unattended objects in LOC were insensitive to the split procedure does not preclude the possibility that LOC relies on different neural representations for attended and unattended objects. We therefore trained a classifier in a *cross-attention/within-configuration* analysis on attended objects and tested on unattended objects (and vice versa). The classifier was able to cross-classify activation patterns of attended and unattended objects (57.1% accuracy, $t(17) = 5.9$, $p < 0.001$) in LOC (Fig. 3B), strongly suggesting that attended and unattended objects share a common representational basis. Furthermore, *cross-attention* decoding was successful even under a *cross-configuration* decoding scheme (57.1% accuracy, $t(17) = 6.7$, $p < 0.001$; Fig. 3B), demonstrating that the information shared between attended and unattended object representations in LOC is coded non-retinotopically.

Cross-attention/within-configuration decoding was also successful in V1 (57.2% accuracy, $t(17) = 13.2$, $p < 0.001$; Fig. 3B), while *cross-attention/cross-configuration* decoding was not (48.8% accuracy, $t(17) = 1.0$, $p = 0.36$). This difference was significant ($t(17) = 7.2$, $p < 0.001$). Across regions (V1, LOC) there was a region-by-decoding scheme interaction ($F(1,17) = 31.9$, $p < 0.001$), consistent with the presence of configural invariance at the level of LOC, but not at the level of V1, found in the *within-attention* analysis.

Discussion

We investigated the representational format of objects in LOC and its relation to attention. We found that activation patterns of intact and split objects shared information that allowed mutual prediction of the

presented object (*cross-configuration* decoding). Remarkably, *cross-configuration* decoding was at the same level as *within-configuration* classification, that is, intact objects representations were predicted by activation patterns of split objects just as well as by activation patterns of intact objects and vice versa. Crucially, cross-configuration decoding did not depend on attention. This pattern of results strongly suggests that the representational code of objects in LOC is based on a non-holistic format irrespective of attention.

Previous studies showed that BOLD activity in the LOC is barely affected when objects are—as in our study—coarsely scrambled (half-split and two-fold splits: Lerner et al., 2001; 8-fold: Grill-Spector et al., 1998), whereas finer scrambling leads to a strong reduction (Grill-Spector et al., 1998; Lerner et al., 2001). However, whether the activation elicited by coarsely scrambled objects still contains meaningful object-related information remained unclear. We extended those findings by showing that the information coded in activation patterns of intact objects is preserved in the patterns of split objects. Our results in combination with the above reports therefore support a model in which LOC codes objects as part-based representations.

A part-based coding scheme is in line with behavioral priming studies reporting mirror (Biederman and Cooper, 1991), rotational (Biederman and Gerhardstein, 1993) and configural (Thoma et al., 2004) invariance of object representations. Part-based models, which posit independent encoding of object parts, correctly predict priming effects for the range of manipulations above, because the constituent object parts are preserved between prime and probe displays. At the neural level, a number of previous studies have probed the viewpoint dependence of object representations by examining repetition suppression (RS). Some of these studies found evidence for viewpoint-invariant representations in high-level visual cortex (Eger et al., 2004; James et al., 2002; Kourtzi et al., 2003), others found tolerance only in the left hemisphere (Vuilleumier et al., 2002) or not at all (Grill-Spector et al., 1999). While the present study cannot reconcile these reports, it introduces a new, multivariate perspective to the longstanding question of view dependence. Our cross-classification approach is sensitive to object-related information coded at the level of multivoxel activation patterns, which could not be assessed by previous imaging studies using univariate fMRI data analysis. At this pattern level we found a striking invariance of object representations with respect to the relative dislocation of object parts in LOC, indicative of a part-based code with all its theoretical advantages regarding robust and flexible coding of objects under various viewing conditions (Hummel and Biederman, 1992).

It should be noted that the results of our cross-configuration decoding analyses are not informative about the specific nature or complexity of object parts. For instance, since our split manipulation only distorted the overall holistic image but not individual parts, our results are open to the possibility that the part representations themselves are view-based, and not extracted structural descriptions, as proposed by Biederman (1987). However, our findings entail the constraint of relative position invariance of object parts on models of object recognition. In particular, the "chorus of fragments" model by Edelman and Intrator (2000), which poses "what + where" units coding the conjunction of part (fragment) information and retinotopic position, is at odds with this constraint. Our results, however, seem to fit a "bag of features" (Hayworth et al., 2011) model, as for instance proposed by Ullman and colleagues (Epshtein and Ullman, 2007; Ullman, 2007).

Our second key finding is that the format of object representations in LOC was independent of attention. We could cross-classify between intact and split objects, whether they were attended or unattended. This finding was further corroborated by the fact that we were able to predict the attended objects based on the activation patterns of unattended objects (and vice versa). Thus, not only appears the LOC to adhere to a part-based format irrespective of attention, but the underlying neural representations additionally seem to be shared between attended and unattended objects. Importantly, we found this *cross-attention* generalization also under a *cross-configuration* decoding scheme, demonstrating that the activation patterns of attended and unattended objects indeed shared non-retinotopic, high-level information. Taken together, our results do not provide evidence for a critical role of attention for part-based representations, as implicated by other empirical findings (Stankiewicz et al., 1998; Thoma and Henson, 2011; Thoma et al., 2004) and theoretical accounts (Hummel, 2001). The main difference between attended and unattended object representations in our study was of quantitative nature—superior decoding accuracy for attended objects, likely related to neural gain (see Pratte et al., 2013)—but not qualitative in terms of the underlying representational format. Although, given its focus on neural effects, our study is not in the position to challenge the findings from these previous behavioral priming studies, it has a number of important advantages. First, we ensured the effectiveness of our attentional manipulation both at the behavioral—by means of eyetracking—and at the neural level, based on superior decoding accuracies for attended relative to unattended objects. Second, we directly probed the representational format in LOC and its dependence on attention, whereas behavioral priming studies rely on indirect inference from reaction times. And third, our brightness discrimination task attenuated the semantic aspect of object recognition, which is arguably more pronounced in the object naming tasks employed by many priming studies. Semantic top-down feedback from higher areas is a potential confounding factor in these studies, since feedback might be responsible for view-invariant priming, but might itself be dependent on attention. Future studies could investigate whether our finding of configurational invariance for both attended and unattended object representations is conditional on using a non-semantic object perception task, or whether it holds for tasks requiring object identification.

Our results also differ from a previous neuroimaging study that found evidence for a part-based format in LOC only for attended, but not for unattended, objects (Thoma and Henson, 2011). However, the paradigm and the analyses of Thoma and Henson deviate in a number of important aspects from those of the present study. The presentation times in Thoma and Henson were considerably shorter than in our study (135 ms, compared to 2 s in our study), which opens the possibility that the instantiation of part-based representations requires longer presentations times in the absence of attention. This assertion would be consistent the notion that attention boosts neuronal processing of stimuli by increasing the signal-to-noise ratio of neuronal responses (Bisley, 2011). Another noteworthy difference is the fact that in our study objects were repeated multiple times throughout the experiment, whereas each object in Thoma and Henson appeared in exactly one trial. Under the assumption of a view-based object format, the multiple repetitions in our study could have therefore led to the instantiation of view-based representations of (previously unfamiliar) split objects. However, the fact that we found perfect generalization between intact and split objects, despite the large number of repetitions, argues against a build-up (or prior existence) of view-based representations for split objects. On the methodological side, Thoma and Henson assessed the effects of RS, whereas we employed a cross-classification approach that utilized the full pattern information. Given that unattended objects evoke a weaker BOLD response (Murray and Wojciulik, 2004; O'Craven et al., 1999; Serences et al., 2004), it seems possible that RS was not sensitive enough to detect representational commonalities between intact and split unattended objects. Further, the analysis of repetition suppression misses out on object-related information stored in activation pattern, which likewise could explain the observed discrepancies for unattended objects.

As noted above, an important consideration with respect to our support vector machine approach is that successful decoding between objects cannot conclusively inform about the particular stimulus features underlying classification. A possible concern in this regard is that between-object decoding might have entirely been based on low-level visual features. For instance, the surface of an object might have a

certain characteristic texture, and further, the same kind of texture might even be present at a similar retinotopic location after the splitting procedure—hence explaining the observed cross-configuration generalization. However, for several reasons we consider a pure low-level account of our results unlikely. First, if retinotopic low-level features were an important source of discriminative information between objects, cross-configuration decoding should have worked in V1 as well, which it did not. Second, the intact and split versions of our object images had very little low-level commonalities in retinotopic coordinates as an image analysis with a biologically plausible model of visual cortex confirmed (Inline Supplementary Fig. S2). Third, the voxels entering the multivariate analysis were precisely selected for preferring complex features over low-level features. And fourth, the results of the cross-hemifield analysis show that the cross-configuration generalization holds also for high-level representations with near-complete location invariance (Inline Supplementary Fig. S1). Therefore, while acknowledging the possibility that the classifier could have only picked up on low-level information, we consider such an account highly unlikely for the reasons outlined.

Inline Supplementary Fig. S2 can be found online at http://dx.doi.org/10.1016/j.neuroimage.2014.12.013.

In summary, our study provides novel evidence indicating that neural representations of both attended and unattended objects in LOC rely on a non-holistic rather than view-based format. Moreover, our data strongly suggest that attended and unattended objects rely on a common representational format.

Acknowledgments

This research was supported by the German Research Foundation (DFG) through the Research Training Group GRK1589/1 (to M.G. and P.S.), and Grants STE1430/6-1 (to P.S.), and RI1847/1-1 and SFB779TPA10N (to A.R.-K). R.C. was funded by a Feodor Lynen Grant of the Alexander von Humboldt Foundation. We thank Guy Middleton for assistance with rendering the object images from 3D models.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2014.12.013.

References

Andresen, D.R., Vinberg, J., Grill-Spector, K., 2009. The representation of object viewpoint in human visual cortex. NeuroImage 45, 522–536.
Biederman, I., 1987. Recognition-by-components: a theory of human image understanding. Psychological Review 94, 115–147.
Biederman, I., Cooper, E.E., 1991. Evidence for complete translational and reflectional invariance in visual object priming. Perception 20, 585–593.
Biederman, I., Gerhardstein, P.C., 1993. Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint in variance. J. Exp. Psychol. Hum. Percept. Perform. 19, 1162–1182.
Bisley, J.W., 2011. The neural basis of visual attention. J. Physiol. 589, 49–57.
Brainard, D.H., 1997. The psychophysics toolbox. Spat. Vis. 10, 433–436.
Bülthoff, H.H., Edelman, S., 1992. Psychophysical support for a two-dimensional view interpolation theory of object recognition. Proc. Natl. Acad. Sci. U. S. A. 89, 60–64.
Cichy, R.M., Sterzer, P., Heinzle, J., Elliott, L.T., Ramirez, F., Haynes, J.-D., 2013. Probing principles of large-scale object representation: category preference and location encoding. Hum. Brain Mapp. 34, 1636–1651.
Edelman, S., Bülthoff, H.H., 1992. Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. Vision Res. 32, 2385–2400.
Edelman, S., Intrator, N., 2000. (Coarse coding of shape fragments) Representation of structure. Spat. Vis. 13, 255–264.
Eger, E., Henson, R.N.A., Driver, J., Dolan, R.J., 2004. BOLD repetition decreases in object-responsive ventral visual areas depend on spatial attention. J. Neurophysiol. 92, 1241–1247.
Eger, E., Ashburner, J., Haynes, J.-D., Dolan, R.J., Rees, G., 2008. fMRI activity patterns in human LOC carry information about object exemplars within category. J. Cogn. Neurosci. 20, 356–370.
Eickhoff, S.B., Stephan, K.E., Mohlberg, H., Grefkes, C., Fink, G.R., Amunts, K., Zilles, K., 2005. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. NeuroImage 25, 1325–1335.
Epshtein, B., Ullman, S., 2007. Semantic Hierarchies for Recognizing Objects and Parts. IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07, pp. 1–8.
Ewbank, M.P., Schluppeck, D., Andrews, T.J., 2005. fMR-adaptation reveals a distributed representation of inanimate objects and places in human visual cortex. NeuroImage 28, 268–279.
Gauthier, I., Hayward, W.G., Tarr, M.J., Anderson, A.W., Skudlarski, P., Gore, J.C., 2002. BOLD activity during mental rotation and viewpoint-dependent object recognition. Neuron 34, 161–171.
Grill-Spector, K., Kushnir, T., Hendler, T., Edelman, S., Itzchak, Y., Malach, R., 1998. A sequence of object-processing stages revealed by fMRI in the human occipital lobe. Hum. Brain Mapp. 6, 316–328.
Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., Malach, R., 1999. Differential processing of objects under various viewing conditions in the human lateral occipital complex. Neuron 24, 187–203.
Grill-Spector, K., Kushnir, T., Hendler, T., Malach, R., 2000. The dynamics of object-selective activation correlate with recognition performance in humans. Nat. Neurosci. 3, 837–843.
Grill-Spector, K., Kourtzi, Z., Kanwisher, N., 2001. The lateral occipital complex and its role in object recognition. Vision Res. 41, 1409–1422.
Grill-Spector, K., Henson, R., Martin, A., Grillspector, K., 2006. Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn. Sci. 10, 14–23.
Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.
Haynes, J.-D., Rees, G., 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat. Neurosci. 8, 686–691.
Hayward, W.G., Zhou, G., Man, W.-F., Harris, I.M., 2010. Repetition blindness for rotated objects. J. Exp. Psychol. Hum. Percept. Perform. 36, 57–73.
Hayworth, K.J., Biederman, I., 2006. Neural evidence for intermediate representations in object recognition. Vision Res. 46, 4024–4031.
Hayworth, K.J., Lescroart, M.D., Biederman, I., 2011. Neural encoding of relative position. J. Exp. Psychol. Hum. Percept. Perform. 37, 1032–1050.
Hebart, M.N., Görgen, K., Haynes, J.-D., 2014. The Decoding Toolbox (TDT): A versatile software package for multivariate analyses of functional imaging data. Frontiers in Neuroinformatics 8.
Hummel, J.E., 2001. Complementary solutions to the binding problem in vision: implications for shape perception and object recognition. Vis. Cogn. 8, 489–517.
Hummel, J.E., Biederman, I., 1992. Dynamic binding in a neural network for shape recognition. Psychol. Rev. 99, 480–517.
James, T.W., Humphrey, G.K., Gati, J.S., Menon, R.S., Goodale, M.A., 2002. Differential effects of viewpoint on object-driven activation in dorsal and ventral streams. Neuron 35, 793–801.
Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8, 679–685.
Kourtzi, Z., Kanwisher, N., 2000. Cortical regions involved in perceiving object shape. J. Neurosci. 20, 3310–3318.
Kourtzi, Z., Erb, M., Grodd, W., Bülthoff, H.H., 2003. Representation of the perceived 3-D object shape in the human lateral occipital complex. Cereb. Cortex 13, 911–920.
Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. 103, 3863–3868.
Lerner, Y., Hendler, T., Ben-Bashat, D., Harel, M., Malach, R., 2001. A hierarchical axis of object processing stages in the human visual cortex. Cereb. Cortex 11, 287–297.
Logothetis, N.K., Pauls, J., Bülthoff, H.H., Poggio, T., 1994. View-dependent object recognition by monkeys. Curr. Biol. 4, 401–414.
Malach, R., Reppas, J.B., Benson, R.R., Kwong, K.K., Jiang, H., Kennedy, W.A., Ledden, P.J., Brady, T.J., Rosen, B.R., Tootell, R.B., 1995. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proc. Natl. Acad. Sci. 92, 8135–8139.
Marr, D., Nishihara, H.K., 1978. Representation and recognition of the spatial organization of three-dimensional shapes. Proc. R. Soc. Lond. Ser. B Biol. Sci. 200, 269–294.
Murray, J.E., 1999. Orientation-specific effects in picture matching and naming. Mem. Cogn. 27, 878–889.
Murray, S.O., Wojciulik, E., 2004. Attention increases neural selectivity in the human lateral occipital complex. Nat. Neurosci. 7, 70–74.
O'Craven, K.M., Downing, P.E., Kanwisher, N., 1999. fMRI evidence for objects as the units of attentional selection. Nature 401, 584–587.
Olshausen, B.A., Anderson, C.H., Essen, D.C. Van, 1993. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J. Neurosci. 13, 4700–4719.
Pelli, D.G., 1997. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat. Vis. 10, 437–442.
Perrett, D.I., Oram, M.W., Ashbridge, E., 1998. Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. Cognition 67, 111–145.
Poggio, T., Edelman, S., 1990. A network that learns to recognize 3D objects. Nature 343, 263–266.
Pratte, M.S., Ling, S., Swisher, J.D., Tong, F., 2013. How attention extracts objects from noise. J. Neurophysiol. 110, 1346–1356.
Rice, G.E., Watson, D.M., Hartley, T., Andrews, T.J., 2014. Low-level image properties of visual objects predict patterns of neural response across category-selective regions of the ventral visual pathway. J. Neurosci. 34, 8837–8844.
Serences, J.T., Schwarzbach, J., Courtney, S.M., Golay, X., Yantis, S., 2004. Control of object-based attention in human cortex. Cereb. Cortex 14, 1346–1357.

Stankiewicz, B.J., Hummel, J.E., Cooper, E.E., 1998. The role of attention in priming for left–right reflections of object images: evidence for a dual representation of object shape. J. Exp. Psychol. Hum. Percept. Perform. 24, 732–744.
Tarr, M.J., 1995. Rotating objects to recognize them: a case study on the role of viewpoint dependency in the recognition of three-dimensional objects. Psychon. Bull. Rev. 2, 55–82.
Tarr, M.J., Gauthier, I., 1998. Do viewpoint-dependent mechanisms generalize across members of a class? Cognition 67, 73–110.
Tarr, M.J., Pinker, S., 1989. Mental rotation and orientation-dependence of shape recognition. Cogn. Psychol. 21, 233–282.
Thoma, V., Davidoff, J., 2007. Object recognition: attention and dual routes. In: Osaka, N., Rentschler, I., Biederman, I. (Eds.), Object Recognition, Attention, and Action. Springer, Tokyo, pp. 141–158.
Thoma, V., Henson, R.N., 2011. Object representations in ventral and dorsal visual streams: fMRI repetition effects depend on attention and part-whole configuration. NeuroImage 57, 513–525.
Thoma, V., Hummel, J.E., Davidoff, J., 2004. Evidence for holistic representations of ignored images and analytic representations of attended images. J. Exp. Psychol. Hum. Percept. Perform. 30, 257–267.
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15, 273–289.
Ullman, S., 1989. Aligning pictorial descriptions: an approach to object recognition. Cognition 32, 193–254.
Ullman, S., 1998. Three-dimensional object recognition based on the combination of views. Cognition 67, 21–44.
Ullman, S., 2007. Object recognition and segmentation by a fragment-based hierarchy. Trends Cogn. Sci. 11, 58–64.
Vuilleumier, P., Henson, R.N., Driver, J., Dolan, R.J., 2002. Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. Nat. Neurosci. 5, 491–499.

Supplementary Materials

1. Cross-hemifield decoding

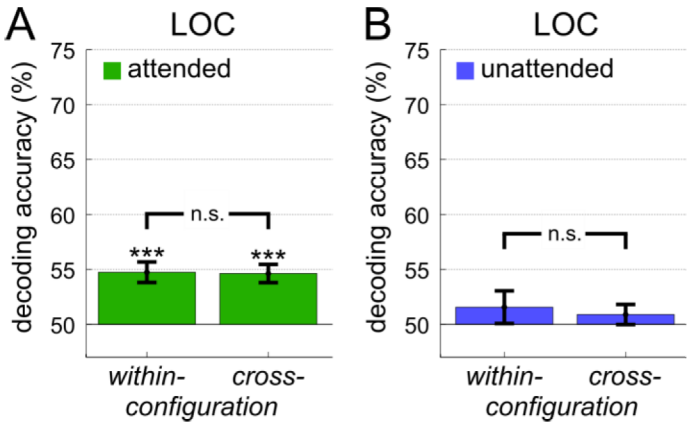


Fig. S1. Within- and cross-configuration decoding in LOC under a cross-hemifield decoding scheme. Stars represent the significance of decoding accuracies based on two-tailed – t-tests against the chance-level decoding accuracy of 50%: *** $p<0.001$.

2. Image analysis with HMAX

In this control analysis we tested the possibility that the object stimuli contained general image properties, which could be picked up by retinotopic feature detectors in V1 or LOC even after a half-split of the image (an example would be a common orientation structure across the whole object, e.g. due to a certain texture or shading). To this aim we trained a state-of-the-art neural network model of primate visual cortex (a variant of HMAX; Serre et al., 2005) on our stimulus set. The model contains 4 layers (S1, C1, S2, C2), where the first two layers correspond to primary visual cortex (V1), S2 corresponds to an intermediate area and C2 to a high-level visual area with position invariance. We passed our stimulus set to the network and computed the representational dissimilarity matrices (Kriegeskorte et al., 2008) of our object stimuli based on the activity of output nodes in each layer (Fig. S2). We found

that in layers S1-S2, the dissimilarity between intact and split versions of the same object (S1: 0.87±0.05; C1: 0.82±0.06; S2: 0.85±0.012) and between different objects (S1: 0.82±0.01; C1: 0.74±0.02; S2: 1.00±0.04) was at a comparable level. It was only in layer C2 that the within-object (intact versus split) dissimilarity sharply dropped (0.13±0.02), while the between-object dissimilarity remained at a high level (1.04±0.03). Thus the representational *similarity* of intact and split objects emerged only at the level of C2, in which position invariance is established. The possibility that common image characteristics were present at similar retinotopic locations in both the intact and the half-split version of the objects is therefore unlikely.

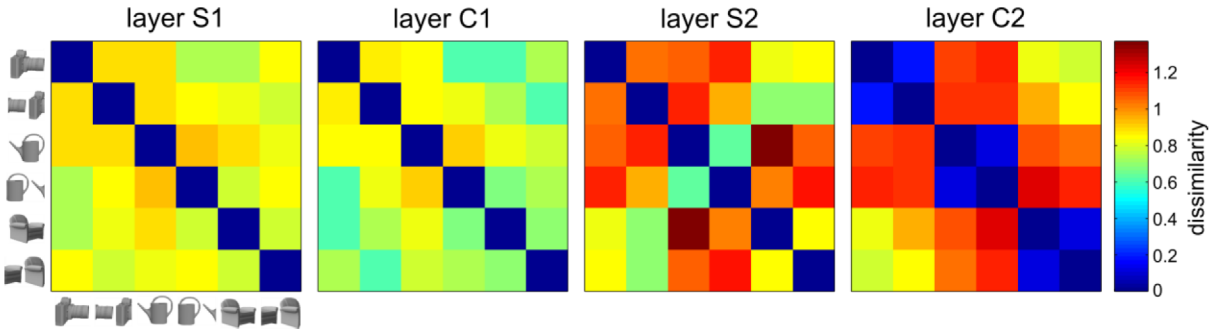


Fig. S2. Image analysis using HMAX. Color depicts the pairwise dissimilarity (1–Pearson's correlation coefficient) between activation patterns of stimuli at each layer.

Supplementary references

Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2, 1–28.

Serre, T., Wolf, L., Poggio, T., 2005. Object Recognition with Features Inspired by Visual Cortex. IEEE Int. Conf. Comput. Vis. pattern Recognit. 2, 994–1000.

6.3 Spatial attention enhances object coding in local and distributed representations of the lateral occipital complex

Journal: NeuroImage

Acceptance date: 01 April 2015

URL: http://dx.doi.org/10.1016/j.neuroimage.2015.04.004

NeuroImage 116 (2015) 149–157

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Spatial attention enhances object coding in local and distributed representations of the lateral occipital complex

Matthias Guggenmos [a,b,*], Volker Thoma [c], John-Dylan Haynes [a], Alan Richardson-Klavehn [d], Radoslaw Martin Cichy [e,1], Philipp Sterzer [a,b,1]

[a] Bernstein Center for Computational Neuroscience, Berlin, Germany
[b] Visual Perception Laboratory, Charité Universitätsmedizin, Berlin, Germany
[c] School of Psychology, University of East London, London, UK
[d] Department of Neurology, Otto von Guericke University, Magdeburg, Germany
[e] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA

ARTICLE INFO

Article history:
Received 30 December 2014
Accepted 1 April 2015
Available online 10 April 2015

Keywords:
Attention
Objects
Lateral occipital complex
Multivariate pattern analysis
Mutual information
fMRI

ABSTRACT

The modulation of neural activity in visual cortex is thought to be a key mechanism of visual attention. The investigation of attentional modulation in high-level visual areas, however, is hampered by the lack of clear tuning or contrast response functions. In the present functional magnetic resonance imaging study we therefore systematically assessed how small voxel-wise biases in object preference across hundreds of voxels in the lateral occipital complex were affected when attention was directed to objects. We found that the strength of attentional modulation depended on a voxel's object preference in the absence of attention, a pattern indicative of an amplificatory mechanism. Our results show that such attentional modulation effectively increased the mutual information between voxel responses and object identity. Further, these local modulatory effects led to improved information-based object readout at the level of multi-voxel activation patterns and to an increased reproducibility of these patterns across repeated presentations. We conclude that attentional modulation enhances object coding in local and distributed object representations of the lateral occipital complex.

© 2015 Elsevier Inc. All rights reserved.

Introduction

Attention is a cognitive process that enables us to focus on certain aspects of the environment for the benefit of improved performance (Bashinski and Bacharach, 1980; Cameron et al., 2002; Carrasco et al., 2000; Hawkins et al., 1990). One way in which attention has been found to impact neural processing is through an amplification of neural responses to attended spatial locations, objects, or features (for review, see Treue, 2003). In the visual domain, attentional amplification has been found throughout the visual processing hierarchy, from the earliest stage of visual neural processing in the lateral geniculate nucleus (O'Connor et al., 2002), primary visual cortex (Gandhi et al., 1999; Martínez et al., 1999; Somers et al., 1999), up to high-level visual cortices (Murray and Wojciulik, 2004; O'Craven et al., 1999; Serences et al., 2004) and the frontal lobes (Gitelman et al., 1999). However, the nature of attentional modulation remains a topic of debate. A number of studies have reported that attention leads to a multiplicative scaling of neuronal responses (Di Russo et al., 2001; McAdams and Maunsell, 1999; Treue and Martínez Trujillo, 1999; Treue and Maunsell, 1999), which results in an increase of a neuron's signal to noise ratio. In contrast, other studies reported results that violated the predictions of the multiplication hypothesis, by showing that spatial attention leads to increased neural responses in visual areas in the absence of any visual stimulation (Kastner et al., 1999; Luck et al., 1997; Ress et al., 2000; Silver et al., 2007). According to these studies, attentional modulation involves an unspecific baseline shift of activity.

A common approach to investigate the effects of visual attention is the recording of neural responses across a range of a stimulus parameter (e.g., orientation or motion direction) both in the presence and absence of attention. In this way, previous studies have examined the attentional modulation of single-neuron (McAdams and Maunsell, 1999; Motter, 1993; Treue and Martínez Trujillo, 1999) or voxel (Saproo and Serences, 2010, 2014) tuning profiles. However, a complicating factor for the investigation of attentional modulation in high-level object-coding areas like the human lateral occipital complex (LOC) is the lack of analogous neuronal tuning functions. Similarly, the analysis of contrast response functions – a technique that has been used to study the nature of attentional modulation for low-level visual stimuli (Reynolds et al., 2000; Williford and Maunsell, 2006) – is problematic, because object-related neuronal responses become increasingly invariant to contrast along the visual hierarchy (Avidan et al., 2002; Rolls and

* Corresponding author at: Bernstein Center for Computational Neuroscience, Philippstraße 13, Haus 6, 10115 Berlin, Germany.
E-mail address: matthias.guggenmos@bccn-berlin.de (M. Guggenmos).
[1] Contributed equally.

http://dx.doi.org/10.1016/j.neuroimage.2015.04.004
1053-8119/© 2015 Elsevier Inc. All rights reserved.

Baylis, 1986) and this invariance may itself depend on attention (Murray and He, 2006). In the present work we therefore used a different approach by exploiting the fact that the LOC represents objects in a distributed fashion across ensembles of neural populations (Haxby et al., 2001; Rice et al., 2014). At the spatial resolution of fMRI this distributed code is expressed in a differential preference of voxels for a given stimulus, likely representing the cumulative stimulus preference of neurons within these voxels. Thus, if attention causes an amplification of neural activity as opposed to a mere baseline shift, these preferences should be augmented with attention, and as a consequence single- and multi-voxel responses should become more informative about the stimuli encoded in these voxels.

In the present study we presented human participants with objects under conditions of spatial attention and inattention in a functional magnetic resonance imaging (fMRI) experiment. We had two aims. First, we sought to probe the nature of attentional modulation of visual object responses in the LOC as described above, by examining whether attentional modulation increased with a voxel's preference for a given object in the absence of attention, or whether the modulation was independent of object preference. In a second step we investigated whether these local modulatory effects of attention resulted in a more informative and reliable object code. To this end we used a mutual information metric (Saproo and Serences, 2010; Serences et al., 2009) to assess whether single-voxel responses became more informative about object identity with attention. At the multi-voxel pattern level we examined how these local changes affected the quality of object representations through pattern similarity and classification-based analyses.

Materials and methods

Disclosure

A previous article (Guggenmos et al., 2015) was based on the same fMRI dataset, but pursued a different research question and orthogonal analyses.

Participants

Eighteen healthy participants (11 female, mean age ± SEM, 23.4 ± 0.8 years) took part in the experiment for payment after giving written informed consent. The study was conducted according to the declaration of Helsinki, and approved by the local ethics committee.

Experimental design

Our key experimental manipulation was to direct participants' spatial attention to either an object (attended condition) or a noise stimulus (unattended condition). Overall the experimental design comprised the factors attention (attended, unattended) as a factor of interest, as well as object (camera, watering can, chair), configuration (intact, split) and side of presentation (left, right) as factors of no interest. Configuration was manipulated by minimally scrambling (half-splitting) the objects, but note that the analyses in this article were based on intact objects only. Within each of 8 experimental runs, an object appeared in 4 trials in each attention condition (in 2 trials per side of presentation). The order of presentation was randomized across the 48 trials of each run.

Experimental procedures

In each trial (Fig. 1A), participants viewed a stimulus display that contained an object and a noise stimulus on either side of a central fixation cross. Spatial attention was manipulated by means of a brightness discrimination task that was performed either on the object (attended condition) or the contralateral noise stimulus (unattended condition). A trial (Fig. 1A) started with a blank fixation screen for 3300 ms ± 2000 ms, after which one half of a central black fixation diamond turned red, indicating the side to which attention should be directed. Following this cue and a short fixed interval (250 ms), four repetitions of the stimulus–response phase appeared. Each stimulus–response phase lasted 1500 ms and comprised the presentation of the stimulus screen (500 ms), a pattern mask (133 ms) and a response screen (867 ms). The object appeared on one side of the fixation cross (offset 3.84° of visual angle) and a noise stimulus at the same offset on the other side of the stimulus screen. All visual stimuli subtended 3.81 by 3.81° of visual angle. A brightness change occurred 283 ms after stimulus onset simultaneously on both the object and the noise stimulus, such that they became independently and randomly either darker or lighter. Participants were instructed to press a button on the response box when the stimulus on the cued side became darker. Responses were counted as valid within a time window of 1000 ms after stimulus offset. In each repetition of the stimulus–response phase, the same object was shown at the same position. The noise stimulus, while also presented at the same position, was randomly generated for each repetition.

To independently identify object-responsive regions of the lateral occipital complex (LOC) in each participant (Malach et al., 1995), we

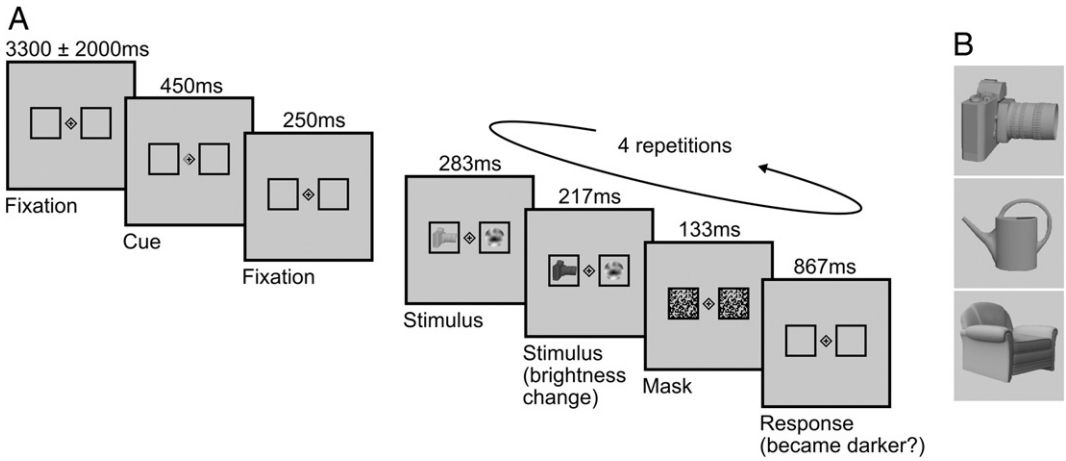


Fig. 1. Experimental procedures and stimuli. A. In each trial a cue indicated the side to which attention should be directed. Subsequently, four repetitions of the stimulus–response phase appeared, during each of which participants had to detect a decrease in brightness of either the object (attended condition) or the noise stimulus (unattended condition). B. The stimulus set consisted of three objects in an intact and half-split configuration.

conducted a localizer run with 5 blocks of intact objects, 5 blocks of split objects and 10 blocks of grid-scrambled versions of the objects in randomized order. Blocks lasted for 15.8 s during which 20 images were presented for 600 ms each, followed by 200 ms blank screen. Pairs of identical objects were shown left and right of fixation, equaling the configuration of the main experiment in eccentricity and size. Participants performed a one-back task on the object pairs, in which they had to indicate via button press whenever the same stimulus display appeared twice in a row.

Stimuli

Stimuli were generated with Psychophysics Toolbox 3 (http://psychtoolbox.org) and projected with a Sanyo LCD projector at 60 Hz. The stimulus set consisted of three grayscale objects (camera, watering can, chair) based on realistic three-dimensional models presented either intact or half-split (Fig. 1B). The objects were selected for representing non-overlapping man-made categories to increase the discriminability of evoked neuronal activation patterns. The noise stimuli matched the objects in terms of spatial extent and complexity to ensure that there would be no performance difference. They were randomly generated for each trial by sampling a 9×9 random binary matrix, scaling the matrix to 216×216 pixels, applying a low-pass filter with a cut-off frequency of 0.02/pixel and cropping pixels outside a circle of 216 pixels diameter. This procedure resulted in circular grayscale stimuli with randomly distributed smooth patches that approximately matched the objects in terms of spatial extent. Both the objects and the noise stimuli were scaled to grayscale RGB values between 50 and 205. To generate these brightness changes, the underlying RGB histograms were shifted up or down by 50 (the image background remained constant with an RGB value of 200). The pattern masks were generated for each trial by sampling an 18×18 random binary matrix and scaling the matrix to 216×216 pixels.

Eyetracking

Eyetracking data were successfully collected in 16 of 18 subjects using an infrared video eyetracking system (iView XTM MRI 50Hz, SensoMotoric Instruments, Teltow, Germany). For each run, the horizontal eye movement data were low-pass filtered and drift corrections were performed. As a measure of fixation reliability, we computed the percentage of recorded eye gaze positions during stimulus presentation within a 1.93° visual angle circle around the center of the fixation cross. This radius corresponded to the eccentricity of the inner edges of the two stimulus-containing boxes (see Fig. 1A). In addition, we computed the number of saccades to the intact objects and the noise stimuli, separately for the attended and the unattended condition. Saccades were defined as events of at least three consecutive data points in velocity space exceeding a velocity criterion of 30°/s. Saccades were qualified as object-directed or noise-directed saccades, when their endpoint was located within the object-containing box, or the noise-containing box, respectively.

FMRI data acquisition and preprocessing

FMRI data were acquired on a 3-Tesla Siemens Trio (Erlangen, Germany) scanner using a gradient echo planar imaging (EPI) sequence and a 12-channel head-coil. We recorded 8 experimental runs of 214 whole-brain volumes each, and one LOC localizer run of 242 volumes (TR = 2 s, echo time (TE) 25 ms, flip angle 78°, 33 slices, 3 mm isotropic resolution, interslice gap 0.75 mm). In addition, a high-resolution T1-weighted image was acquired (TR = 1.9 s, echo time (TE) 2.51 ms, flip angle 9°, 192 slices, resolution 1 mm isotropic). The data of the experimental runs were realigned using SPM8 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London). Data analyses for the main experiment were generally performed in native subject space. An exception was an illustrative display of the whole-brain group-level T-maps for the main effect of attention, for which we generated spatially normalized (MNI) and smoothed (8 mm Gaussian kernel) volumes. Preprocessing of the localizer data included realignment, spatial normalization to an MNI template and smoothing (8 mm Gaussian kernel).

FMRI data analysis

First-level general linear models (GLMs)

For each participant we estimated a GLM including the stimulus-onset regressors, accounting for the factors attention (attended, unattended), object (camera, can, chair) and configuration (intact, split). The onsets of each experimental regressor were set to the beginning of the stimulus–response phase. In addition, six motion parameters were included as regressors-of-no-interest. All experimental regressors were modeled as stick functions and convolved with a canonical hemodynamic response function.

The GLM for the functional localizer comprised regressors for objects and scrambled objects and six motion parameters. The experimental regressors were modeled as boxcar functions with durations equal to the block lengths (15.8 s) and convolved with a canonical hemodynamic response function as implemented in SPM8.

Region of interest procedures

Our region of interest (ROI) was the LOC, a functionally defined region that responds more to images of objects than their counterparts and stretches from the lateral occipital cortex to posterior fusiform gyrus (Grill-Spector et al., 1999). We anatomically constrained the LOC by a bilateral composite mask of the inferior occipital cortex, middle occipital cortex and the posterior half of the fusiform gyrus (derived from the AAL Atlas, Tzourio-Mazoyer et al., 2002). Then the LOC ROI was defined as the intersection of the anatomical mask and the functional localizer based on the t-contrast *intact objects > scrambled objects* of the functional localizer at a significance level of $p < 0.05$ (family-wise error (FWE) corrected at the whole-brain level). Note that this t-contrast was a group-level t-contrast, because the statistical power in the first-level localizer contrasts was not sufficient to define individual ROIs in all participants at the $p_{FWE} < 0.05$ level. To ensure a homogenous generation of the LOC ROI for all participants we thus first defined the LOC ROI in group-level (MNI) space and subsequently reverse-normalized the generated ROI to each participant's native space.

Quantifying changes in mean BOLD activity

To estimate neural activity in the LOC ROI and its dependence on attention, we extracted the voxel-wise beta values for attended and unattended objects separately and averaged across objects and voxels. This procedure resulted in single values representing the average BOLD response to attended and unattended objects.

To visualize the spatial extent of the attentional modulation at a whole-brain level, we performed a group-level repeated-measures ANOVA with factors attention and object and computed the post-hoc contrast *attended > unattended*. This analysis was based on normalized and smoothed data. Voxels were considered statistically significant at a level of $p < 0.05$, FWE-corrected at the whole-brain level.

Analyzing attentional modulation as a function of object preference

We next analyzed whether the attentional modulation depended on the preference of a voxel for a given object. We reasoned that if attention leads to an amplification of neural responses, the difference between a voxel's attentional modulation for the preferred object and the modulation for the non-preferred objects should increase as a function of object preference. By contrast, if attention led to an unspecific baseline shift irrespective of a voxel's preference for the presented object, the attentional modulation should not differ between the

presentations of the voxel's preferred and non-preferred object. We therefore defined a preference index PI(i) for each object i and each voxel based on the data of the unattended condition:

$$\mathrm{PI}(i) = \beta_{\mathrm{unatt}}(i) - \langle \beta_{\mathrm{unatt}}(\backslash i) \rangle \ ,$$

where $\beta_{\mathrm{unatt}}(i)$ and $\beta_{\mathrm{unatt}}(\backslash i)$ are the voxel-wise beta values in the unattended condition for object i and all objects except i (denoted as "not i": $\backslash i$) respectively; the symbol <> denotes the average operation (here across objects). PI was based on the unattended condition to circumvent the potential issue that the object preference of a voxel in the attended condition might not be independent of the magnitude of the attention effect. To compute the strength of the attentional modulation for an object i relative to the other objects $\backslash i$, we defined a relative attentional modulation index RAI(i) as follows:

$$\mathrm{RAI}(i) = \beta_{\mathrm{att}}(i) - \beta_{\mathrm{unatt}}(i) - \langle \beta_{\mathrm{att}}(\backslash i) - \beta_{\mathrm{unatt}}(\backslash i) \rangle \ ,$$

where $\beta_{\mathrm{att}}(i)$ and $\beta_{\mathrm{att}}(\backslash i)$ are the voxel-wise beta values in the attended condition for object i and all objects except i respectively.

Finally, we quantified the RAI as a function of PI. To preclude a selection bias we used a leave-one-run-out procedure, such that PI and RAI were computed on independent data. The leave-one-run-out procedure was performed for each object i separately as follows. In each fold, we sorted the pooled voxels from the LOC ROI according to PI(i) based on data from all but one experimental runs. We then divided the sorted voxels into 10 equinumerous bins (deciles) according to PI(i) and computed the average RAI(i) for the voxels in each bin based on the data of the held-out run. Subsequently, we computed an average RAI across objects for each bin, resulting in a single RAI for each bin.

Computing the mutual information between BOLD response and presented objects

To investigate whether attention increased object information encoded in the activity of individual voxels, we used a mutual information (MI) metric. MI estimates the extent to which the uncertainty about one variable Y (here: BOLD response to the object being presented) is reduced by measuring another variable X (here: the object being presented) (cf. Saproo and Serences, 2010; Serences et al., 2009). The mutual information (MI) measure is defined as the difference of the total entropy $H(X)$ and the noise entropy $H(X|Y)$:

$$\begin{aligned} \mathrm{MI}(X;Y) &= H(X) - H(X|Y) \\ &= -\sum_{x \in X} p(x) \log_2 p(x) - \left(-\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 p(x|y) \right). \end{aligned}$$

Thus we subtract from the total entropy $H(X)$, which corresponds to the overall dynamic range of responses, the noise entropy, which is a measure for the noise in the data conditional on each presented object. The remainder quantifies to what degree the variation in the BOLD signal is informative about the presented object. To compute the total and noise entropies, estimated BOLD responses were transformed into a discrete variable (X) by dividing the entire range of responses into a set of 10 equinumerous bins (deciles). This discretization was based on the pooled range of responses from all voxels in either the attended or the unattended condition after subtracting out the respective mean activation levels of the attended and the unattended condition. This subtraction was done to avoid errors in the binning process due to additive shifts attributed to attention (Saproo and Serences, 2010). In the above formulation, $p(x)$ corresponds to the frequency with which a response in a given voxel falls into bin x. The noise entropy term $H(X|Y)$ additionally required the computation of the probability $p(y)$ of each object y – 1/3 in our case, given that the experiment consisted of three equally often appearing objects – and $p(x|y)$, which corresponds to the frequency with which a response in a given voxel falls into bin x, given object y was presented. We normalized the mutual information for each participant to a range between 0 and 1 by dividing MI(X;Y) by the total entropy $H(X)$ (Kojadinovic, 2005). A normalized MI value of 0 indicates that BOLD response X and object label Y are completely independent, whereas a normalized MI value of 1 indicates that response X gives complete information about the object label Y. The MI metric was applied to the responses of attended and unattended objects separately.

Analyzing the effects of attention at the multi-voxel pattern level

To assess the effect of attention at the multi-voxel pattern level, we examined object-related activation patterns with and without attention by means of a pattern similarity measure and support vector machine (SVM) classification. The two methods are complementary in the sense that the similarity measure provided a transparent quantification of the reproducibility (*within-object pattern similarity*) across runs, whereas the SVM classification assessed the amount of information that can be read out from these activation patterns.

Support vector machine classification

Support vector machine classification (SVM) was performed using *The Decoding Toolbox* (Hebart et al., 2014) with a linear C-SVM and a fixed cost parameter ($c = 1$). We quantified object information in the LOC for attended and unattended objects. We trained the classifier to discriminate between objects based on multi-voxel activation patterns in the LOC ROI in native subject space (Haynes and Rees, 2005; Kamitani and Tong, 2005). A leave-one-run-out cross-validation procedure was used, such that in each of 8 folds the classifier was trained on the beta maps of seven runs and tested on the left out eighth run. We performed pair-wise decoding between the three pairs of objects (camera–can, camera–chair, can–chair) separately for the attended and the unattended condition. Subsequently the decoding accuracies were averaged across folds and object pairs.

Pattern similarity analysis

The pattern similarity analysis was based on z-transformed correlations between activation patterns in the LOC ROI. The *within-object pattern similarity* (WPS) measured the correlation between the patterns evoked by the same object across the 8 runs (separately for attended and unattended objects). For each object this led to $8 \cdot (8-1)/2 = 28$ correlation coefficients for the pairwise combinations of runs, which were z-transformed and averaged across permutations and objects. This procedure yielded a single within-object pattern similarity value for both the attended and the unattended condition. As a control analysis, we also computed the *between-object pattern similarity* (BPS). BPS was assessed analogously to WPS, except that the correlation coefficients were computed between patterns evoked by different objects, resulting in three between-object comparisons (camera–can, camera–chair, can–chair). To avoid an overestimation of pattern similarity due to within-run autocorrelations, we excluded all within-run comparisons (Mumford et al., 2014).

Results

Behavioral results and fixation control

Participants detected and reported brightness changes of the objects and the noise stimuli highly accurately (performance > 98%), indicating focused attention on the correct stimulus. On average, 98.3 ± 0.8% (mean ± SEM) of recorded eye gaze positions during stimulus presentation were within the fixation area, demonstrating that the participants maintained fixation throughout the experiment. There was no difference in the overall number of saccades between the attended and the unattended condition (attended: 3.1 ± 1.6 saccades in the experiment; unattended: 3.6 ± 2.2; $p = 0.43$, $t(15) = -0.80$, two-tailed t-test), and neither was there a difference with respect to the number of object-directed (attended: 2.9 ± 1.6; unattended: 0.9 ± 0.6; $t(15) = 1.36$,

$p = 0.19$) or noise-directed saccades (attended: 0.2 ± 0.2; unattended: 2.8 ± 2.0; $t(15) = -1.26$, $p = 0.22$). The interaction of saccade direction (object-directed, noise-directed) and attention (attended, unattended) was not significant ($F(1,15) = 1.71$, $p = 0.21$, repeated-measures ANOVA). These results, as well as the low absolute number of object- or noise-directed saccades, indicate that differences between the neural correlates of the attended and the unattended condition are unlikely to ensue from effects of eye movements.

Attention amplifies responses to objects in the lateral occipital complex

To examine the influence of covert attention on neural activity, we compared the overall average BOLD response for attended and unattended objects within the LOC averaged over objects and sides of presentation. Attended objects led to a significant increase of neural activation ($t(17) = 5.00$, $p < 0.001$, Cohen's $d = 1.17$; Fig. 2A).

In order to test whether the effect of the attention manipulation was confined to object-selective cortex, we quantified the overlap between the thresholded ($p_{FWE} < 0.05$) whole-brain T-maps of the contrasts *attended* > *unattended* (main experiment) and *intact* > *scrambled* (functional localizer). We found that 94.8% of the voxels showing an effect in the attention contrast overlapped with voxels classified as object-selective (Fig. 2B). Thus our focus on the LOC was justified by the spatial extent of the attentional modulation. It should be noted, however, that the attended and the unattended conditions differed only with respect to the attended stimulus type (object vs. noise pattern), but neither systematically with respect to low-level features (likely canceling out effects of attention in earlier visual areas in the contrast attended > unattended) nor with respect to task (likely canceling out effects of attention in executive cortices). The spatial restriction of attentional modulation to LOC therefore reflects a deliberate property of our design and the specific contrast, rather than the absence of attentional modulation in other brain areas.

Attention modulates neural activity as a function of object preference

We reasoned that if attention led to an amplification of neural activity (as opposed to a mere baseline shift), the attentional modulation should be greater for a voxel's preferred object relative to its non-preferred objects. To quantify the difference between the attentional modulation for preferred and non-preferred objects, we computed a relative attentional modulation index (RAI). Further, we determined a preference index (PI) for each voxel based on the mean response to a given object relative to the response of the other objects in the unattended condition. We hypothesized that RAI should increase as a function of PI.

To this end, we used a leave-one-run-out procedure, in which we sorted the voxels according to their PI, divided the voxels into 10 equinumerous bins (deciles) and computed the average RAI for each bin. We found that the RAI increased as a function of PI (linear slope [mean ± SEM]: 0.066 ± 0.022, $t(17) = 3.00$, $p = 0.009$, two-tailed t-test against the null hypothesis of a slope of zero; Fig. 3). A comparison between the collapsed data of preferred (i.e. $PI > 0$) and non-preferred objects (i.e. $PI < 0$) confirmed a higher average RAI for preferred objects ($t(17) = 2.86$, Cohen's $d = 0.67$). These results show that the modulation of neural activity through spatial attention comprises an amplificatory component and is not due to a baseline shift only.

Attention increases the mutual information between BOLD responses and object identity

To test whether the increase of neural activity increased a voxel's information about the presented objects, we applied a mutual information metric separately to the BOLD responses of attended and unattended objects. We found that attention increased the mutual information of a voxel's responses about the objects presented ($t(17) = 5.72$, $p < 0.001$, Cohen's $d = 1.35$). The percentage of voxels showing higher mutual information in the attended relative to the unattended condition was $55.1\% \pm 0.9\%$ (mean ± SEM across participants), which was significantly different from the chance level of 50% ($t(17) = 5.81$, $p < 0.001$). Thus, attention reduced the uncertainty of BOLD responses about object identity, implying enhanced object coding at the level of single voxels.

Attention enhances object representations at the pattern level

A growing body of evidence suggests that the LOC codes object not by means of individual neurons or neuronal populations, but across multiple distributed neuronal populations (Haxby et al., 2001; Rice et al., 2014). Thus, if attention has an enhancing effect on sensory representations, the above finding of object-specific local modulation by attention should improve the quality of multi-voxel activation patterns.

In a first step we assessed the effect of attention on the reproducibility of activation patterns by computing the within-object pattern similarity (WPS) of activation patterns across repeated presentations of the same object, separately for attended and unattended object presentations. We found that attention significantly increased WPS ($t(17) = 10.51$, $p < 0.001$, two-tailed t-test; Fig. 4A), indicating that attention improved the reproducibility of responses at the pattern level. However, in a control analysis we found that attention also led to a considerable increase of the between-object pattern similarity (BPS; $t(17) = 9.73$, $p < 0.001$). If the increase in reproducibility for the same object (WPS) was outweighed by a simultaneous increase of the ambiguity between different objects (BPS), nothing is gained. We therefore directly compared WPS and BPS and found that the increase in WPS was greater than the increase in BPS ($t(17) = 4.22$, $p < 0.001$), indicating that attention led to a functionally relevant improvement of the reproducibility of multi-voxel activation patterns.

In a second step we directly assessed how attention affected the readout of object information from the LOC by performing support

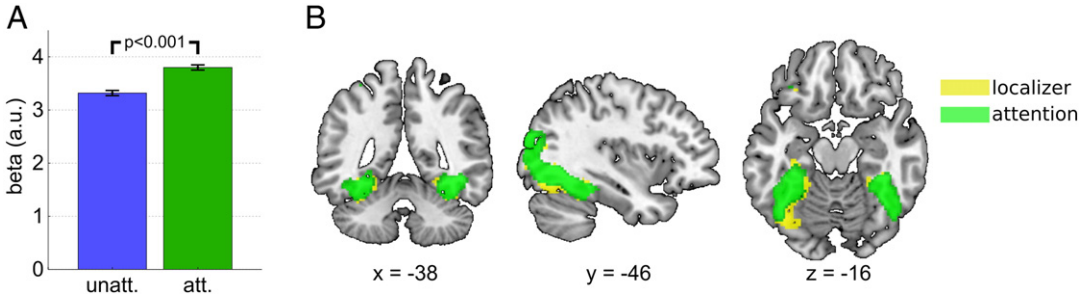


Fig. 2. Modulatory effect of attention. A. LOC ROI. The bars represent average beta values of the LOC ROI for the attended and unattended condition, averaged across objects and sides of presentation. Error bars denote SEM corrected for between-subject variance (Cousineau, 2005). Statistical comparison was based on a two-tailed t-test. B. Whole-brain analysis. Overlay of object-selective voxels (based on the independent functional localizer, intact > scrambled, thresholded at $p_{FWE} < 0.05$, colored in yellow) and voxels showing a significant effect of attention (attended > unattended, thresholded at $p_{FWE} < 0.05$, colored in green).

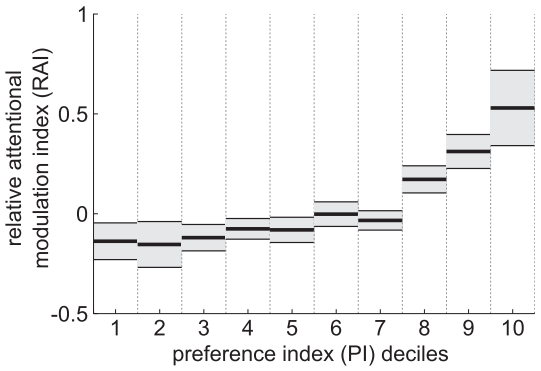


Fig. 3. Relative attentional modulation as a function of object preference. The relative attentional modulation index (RAI) quantifies the attentional modulation for a given object relative to the average modulation of the other objects. For each participant voxels were binned into deciles according to their object preference index (PI). The plot shows the averaged RAI for each preference bin. Error bars denote SEM corrected for between-subject variance (Cousineau, 2005).

vector machine classification between objects. Decoding accuracies were significantly above chance in both the attended (66.6% ± 2.0%; $t(17) = 8.36$, $p < 0.001$, two-tailed t-test against the chance decoding accuracy of 50%) and the unattended condition (54.8% ± 1.8%; $t(17) = 2.66$, $p = 0.017$; Fig. 4B). Importantly, classification performance was significantly and markedly greater in the attended compared to the unattended condition ($t(17) = 4.74$, $p < 0.001$, Cohen's $d = 1.12$). Thus, the attentional modulation of neuronal responses increased object information in the LOC at the multi-voxel pattern level.

Enhanced readout at the pattern level is linked to the local increase in mutual information, but not mean activation

Finally, we assessed whether the attentional modulation at the single-voxel level was related to the enhancement of object representations at the pattern level. In the single-voxel-level analyses we found that attention led to an increase of (1) BOLD signal, and (2) mutual information between BOLD responses and object identity. We therefore correlated – across participants – both effects with the increase in decoding accuracy. We found that the increase in decoding accuracy correlated with the increase in mutual information (Pearson's $r = 0.59$, $p = 0.009$), but not with the increase in BOLD activation ($r = 0.01$, $p = 0.96$). A direct comparison confirmed that the increase in mutual information explained significantly more variance than the BOLD increase (z-score = 2.11, $p = 0.034$, Steiger's z-test; Steiger, 1980). Although the absence of a significant contribution of the BOLD increase is surprising (possibly caused by ceiling effects of the BOLD increase), the relationship between mutual information and decoding accuracy suggests that the local attentional modulation of neuronal responses increases the information content of object representations at the pattern level.

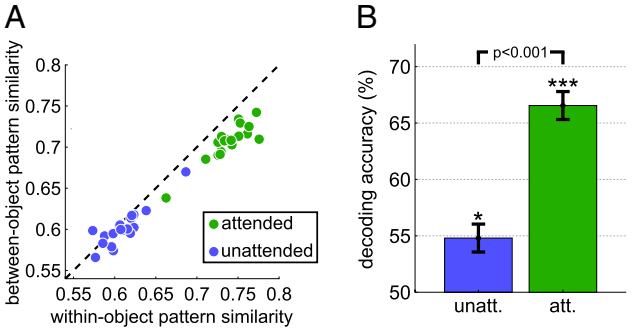


Fig. 4. Pattern level. A, between-object and within-object pattern similarity within the LOC ROI. Each dot represents one participant. The dashed diagonal line indicates identical within- and between-object similarity of activation patterns. Attention leads to a shift of data points below the diagonal line, indicating higher pattern similarity for repeated presentations of the same object compared to the pattern similarity of different objects. Between-subject variance was removed for illustration. B, SVM decoding results based on percent correct classification (decoding accuracy). Error bars denote SEM corrected for between-subject variance (Cousineau, 2005). The statistical comparison was based on a two-tailed t-test. Stars represent the significance of decoding accuracies based on two-tailed t-tests against the chance-level decoding accuracy of 50%: $^{*}p < 0.05$ $^{***}p < 0.001$.

Discussion

We examined tuning-dependent attentional modulation of object representations in the LOC and the resulting enhancement of object representations at the single-voxel level and the multi-voxel pattern level. At the single-voxel level we found that (1) responses in the LOC were considerably stronger when an object was attended relative to when a noise stimulus was attended; (2) the relative attentional modulation (the attentional modulation for a given object relative to the average modulation of other objects) increased as a function of a voxel's preference for the given object; and (3) mutual information between a voxel's responses and object identity increased, demonstrating that responses became more informative about a presented object when the object was attended compared to when it was unattended. All three results provide evidence against a mere baseline-shift effect of attention. Further analyses showed that these local changes resulted in increased object information at the level of multi-voxel patterns and increased similarity of these patterns across multiple presentations, indicating increased reproducibility of distributed neuronal responses.

Effects of attention at the level of individual voxels

A key goal of this study was to investigate whether the observed increase in activity involved amplificatory attentional modulation, or merely an unspecific baseline shift. Previous neuroimaging studies reported that neural activity increased with attention in high-level visual cortex (Baldauf and Desimone, 2014; Connor et al., 1997; Murray and Wojciulik, 2004; O'Craven et al., 1999; Serences et al., 2004), and showed that the effects of attention were specific to coarse functional modules, such as parahippocampal place area (PPA) or fusiform face area (FFA). However, given that objects are known to be coded across distributed neuronal ensembles in visual cortex (Haxby et al., 2001; Rice et al., 2014), it is desirable to analyze attentional modulation at a more fine-grained level, thereby accounting for the differential tuning of neuronal populations within these areas. Here we provide evidence for voxel-wise object-specific attentional modulation of responses in the LOC by identifying a relationship between attentional modulation and object preference. The consistent increase of the relative attentional modulation across preference levels suggests that subtle difference in preference measured in the absence of attention became amplified as attention was directed to the objects. Our additional information-theoretic analyses indicated that such attentional modulation effectively increased the information of voxel-wise responses about object identity, in line with previous work on orientation coding in V1, which likewise found an increase in mutual information with attention (Saproo and Serences, 2010).

How do these results relate to the multiplicative gain hypothesis of attention derived from neurophysiological recordings in monkeys (McAdams and Maunsell, 1999; Treue and Martínez Trujillo, 1999)? It should be noted that a direct comparison between the BOLD responses in our study and spiking activity in these previous studies is difficult for two reasons: first, BOLD responses are more closely related to the local field potentials and hence synaptic activity than to spiking neuronal activity (Ekstrom, 2010; Logothetis, 2003; Logothetis et al., 2001); and second, efficient event-related fMRI designs such as ours do not permit inferences about the absolute level of stimulus-related BOLD activity, which would be necessary to quantify the ratio between attended and

unattended responses analogous to the ratio of firing rates in these previous neurophysiological studies. Nevertheless, our results do provide indirect evidence for the multiplicative gain as opposed to a mere baseline shift hypothesis. Consider the result of increased attentional modulation with object preference. A voxel's preference for a given object may indicate that, for a fixed number of neurons tuned to different objects, the tuning curves of neurons are biased more towards the given object than to the other objects. Alternatively, it may indicate that for a fixed bias towards the given object an overall greater number of neurons prefer the given object. Importantly, in both cases an unspecific baseline shift would lead to an equal increase of neural activity for preferred and non-preferred objects, which is at odds with our results. To illustrate why the increase in MI provides evidence for a multiplicative gain mechanism as opposed to a pure baseline shift explanation, it is helpful to consider two objects A and B and a hypothetical voxel consisting of neurons with a preference for, e.g., object A. In case of a pure baseline shift the voxel would show increased responses to both objects and neural responses would therefore not become more informative about whether object A or B was presented. In contrast, in case of multiplicative scaling, attention will lead to greater response amplification for object A compared to object B, increasing the dynamic range of responses and resulting in increased mutual information between neural responses and presented objects. Thus, the increase in mutual information by attention provides a second line of evidence in favor of a multiplicative gain mechanism and against a pure baseline shift explanation.

Effects of attention at the multi-voxel pattern level

At the level of multi-voxel activation patterns we found improved decodability of attended relative to unattended objects, which is in accordance with similar reports for early (Jehee et al., 2011; Kamitani and Tong, 2005) and high-level visual areas (Pratte et al., 2013; Reddy and Kanwisher, 2007). This result demonstrates that the attentional modulation increased the information content of distributed object representations in the LOC, potentially benefitting information readout from the LOC by high-level executive cortices. An analysis of pattern similarity showed that attention increased the reproducibility of activation patterns of the same object. Such an increase in reproducibility would be expected on the assumption of a multiplicative attentional scaling mechanism, where neuronal responses become amplified without an equivalent increase of the noise (which increases as the square-root of the signal). Another possibility is that the increase in reproducibility is the result of more discrete neural processing with attention, as proposed for conscious relative to non-conscious percepts (Sackur and Dehaene, 2009; Schurger et al., 2010). When discrete decisions are reached at each (object) processing stage, before they are dispatched to the next stage, the resulting activation patterns might become more stereotypical and reproducible.

A number of previous fMRI studies have used MVPA to study the effects of attention on neural responses (Esterman et al., 2009; Jiang et al., 2013; Pratte et al., 2013; Reddy and Kanwisher, 2007; Reddy et al., 2009; Tamber-Rosenau et al., 2011). In particular, Reddy and Kanwisher (2007) and Reddy et al. (2009) investigated the decodability of complex stimuli in high-level visual cortex when they were presented alongside a second object and were either attended or unattended. Reddy and Kanwisher (2007) found that information about object categories encoded in multi-voxel activation patterns was strongly reduced to the point of being abolished when attention was diverted. In the present study we showed that multi-voxel responses were reduced, but still informative about object categories even when attention was diverted. This difference may be explained by the fact that participants in the study by Reddy and Kanwisher (2007) directed their attention to complex distractor stimuli (which, in addition, were the relevant stimuli in other trials), whereas participants in our study viewed noise stimuli in the unattended condition. It is conceivable that the absence of high-level visual cortex information for unattended objects in Reddy and Kanwisher (2007) was caused by distractor-related neural responses interfering with the activation pattern of the unattended target object. Along similar lines, Reddy et al. (2009) interpreted the informational gain for attended objects (or loss for unattended objects) in the biased competition framework. According to this view, attention serves to disambiguate the overlapping multi-voxel patterns of different objects through a shift towards the pattern of the currently attended object. Aside from investigating the effect of attention in sensory cortices, other studies have successfully used MVPA to study the initiation and control of attentional shifts. For instance, Esterman et al. (2009) and Tamber-Rosenau et al. (2011) showed that spatial patterns of brain activity within the medial superior parietal lobule reliably differentiated between several domains of cognitive attentional control at a given moment. Thus, in our and previous studies, MVPA presented a powerful technique to probe distributed neural underpinnings of different attentional phenomena, from the initiation of attentional shifts to the modulation of sensory representations.

Linking the single-voxel and the multi-voxel pattern level

Finally, we linked the effects of attention at the single-voxel level with the effects at the pattern levels by correlating the increase in decoding accuracy of multi-voxel activation patterns to the increase in either BOLD signal or mutual information. Unexpectedly, we found that the increase in mean activation was not related to the increase in decoding accuracy. This negative finding could indicate that the attentional manipulation in our paradigm operated in a range, in which effects at the pattern level were insensitive to the overall magnitude (e.g., because the BOLD increase was at maximum). Alternatively, as the overall effect of attention may involve both a multiplicative component and a baseline shift, the unspecific baseline shift component might have masked the effect of the relevant multiplicative component. In contrast, we found that the increase in mutual information explained a considerable amount of variance of improvements in pattern-based decoding. This result demonstrates that the increase of object information at the single-voxel level substantially translated to an enhanced object code at the pattern level. This link is informative, as the information content encoded in the linear combination of voxels can show strong gains, while information encoded in the individual voxel may show only small changes (for examples of such scenarios see Haynes and Rees, 2006). It is currently not clear whether the distributed object code in LOC represents the immediate neural correlate of perception, or whether it reflects object processing prior to perception. In either case our data indicate that the enhancement of sensory representations through attention – which may directly or indirectly underlie perceptual improvements – is not a phenomenon that solely emerges at the level of distributed object fingerprints. Instead, the improvement in pattern decoding likely represents the cumulative result of informational gains in multiple local units of LOC.

Implications for mechanisms of visual attention

The results of the present study corroborate the notion that behavioral benefits of attention are based on an enhanced stimulus processing in sensory brain areas (Bisley, 2011). Our finding that the magnitude of attentional modulation increased with object preference suggests a response gain mechanism that magnifies stimulus-driven responses as a function of response strength without attention. Importantly, our information-theoretic analyses demonstrate that the attentional modulation effectively increases object information encoded in high-level visual cortex, which may facilitate the readout in executive cortices and thus benefit perceptual decision making. A unifying theoretical framework for such attentional modulation of neural activity is provided by the normalization model of attention (Reynolds and Heeger, 2009). The model describes the modulation of attention by two processes: a

multiplication of neuronal responses by an attention field and a division (normalization) by a suppressive drive. Thus, our observed differences between neural responses to attended and unattended objects may not only be caused by a boost of neural processes tuned to the attended object, but also by a suppression of activity related to the unattended object. Another key aspect of the model is that it makes specific predictions regarding the effect of different attentional strategies on neural activity. According to the model, a purely spatial attention strategy causes a scaling of the entire tuning curves (because the attention field is then assumed to be constant across feature dimensions), whereas a purely feature-based attention strategy causes a sharpening of tuning curves. The fact that our brightness discrimination task emphasized spatial attention strategies over feature-based strategies may thus explain the strong amplitude modulation of the BOLD response in our study. Future neuroimaging studies could test whether our findings of tuning-dependent attentional modulation and information-theoretic gains through endogenous visual spatial attention generalize to other forms of attention, e.g. to involuntary (exogenous) shifts of attention or to other sensory modalities.

In conclusion, our results show that visual spatial attention modulates neural activity as a function of voxel-based object preferences. Through these modulatory processes, attention enhances object coding both at the single-voxel and pattern level, which may give rise to improved perception and perceptual decisions.

Acknowledgments

This research was supported by the German Research Foundation (DFG) through the Research Training Group GRK1589/1 (to M.G. and P.S.), and Grants STE1430/6-1 (to P.S.), and RI1847/1-1 and SFB779TPA10N (to A.R.-K). R.C. was funded by a Feodor Lynen Grant of the Alexander von Humboldt Foundation. We thank Guy Middleton for assistance with rendering the object images from 3D models.

References

Avidan, G., Harel, M., Hendler, T., Ben-Bashat, D., Zohary, E., Malach, R., 2002. Contrast sensitivity in human visual areas and its relationship to object recognition. J. Neurophysiol. 87, 3102–3116.
Baldauf, D., Desimone, R., 2014. Neural mechanisms of object-based attention. Science 344, 424–428.
Bashinski, H.S., Bacharach, V.R., 1980. Enhancement of perceptual sensitivity as the result of selectively attending to spatial locations. Percept. Psychophys. 28, 241–248.
Bisley, J.W., 2011. The neural basis of visual attention. J. Physiol. 589, 49–57.
Cameron, E.L., Tai, J.C., Carrasco, M., 2002. Covert attention affects the psychometric function of contrast sensitivity. Vis. Res. 42, 949–967.
Carrasco, M., Penpeci-Talgar, C., Eckstein, M., 2000. Spatial covert attention increases contrast sensitivity across the CSF: support for signal enhancement. Vis. Res. 40, 1203–1215.
Connor, C.E., Preddie, D.C., Gallant, J.L., Van Essen, D.C., 1997. Spatial attention effects in macaque area V4. J. Neurosci. 17, 3201–3214.
Cousineau, D., 2005. Confidence intervals in within-subject designs: a simpler solution to Loftus and Masson's method. Tutor. Quant. Methods Psychol. 1, 42–45.
Di Russo, F., Spinelli, D., Morrone, M.C., 2001. Automatic gain control contrast mechanisms are modulated by attention in humans: evidence from visual evoked potentials. Vis. Res. 41, 2435–2447.
Ekstrom, A., 2010. How and when the fMRI BOLD signal relates to underlying neural activity: the danger in dissociation. Brain Res. Rev. 62, 233–244.
Esterman, M., Chiu, Y., Tamber-rosenau, B.J., Yantis, S., 2009. Decoding cognitive control in human parietal cortex. Proc. Natl. Acad. Sci. U. S. A. 106, 17974–17979.
Gandhi, S.P., Heeger, D.J., Boynton, G.M., 1999. Spatial attention affects brain activity in human primary. Proc. Natl. Acad. Sci. U. S. A. 96, 3314–3319.
Gitelman, D.R., Nobre, A.C., Parrish, T.B., LaBar, K.S., Kim, Y.H., Meyer, J.R., Mesulam, M., 1999. A large-scale distributed network for covert spatial attention: further anatomical delineation based on stringent behavioural and cognitive controls. Brain 122, 1093–1106.
Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., Malach, R., 1999. Differential processing of objects under various viewing conditions in the human lateral occipital complex. Neuron 24, 187–203.
Guggenmos, M., Thoma, V., Cichy, R.M., Haynes, J.-D., Sterzer, P., Richardson-Klavehn, A., 2015. Non-holistic coding of objects in lateral occipital complex with and without attention. Neuroimage 107, 356–363.
Hawkins, H.L., Hillyard, S.A., Luck, S.J., Mouloua, M., Downing, C.J., Woodward, D.P., 1990. Visual attention modulates signal detectability. J. Exp. Psychol. Hum. Percept. Perform. 16, 802–811.
Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.
Haynes, J.-D., Rees, G., 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat. Neurosci. 8, 686–691.
Haynes, J.-D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7, 523–534.
Hebart, M.N., Görgen, K., Haynes, J.-D., 2014. The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Front. Neuroinform. 8.
Jehee, J.F.M., Brady, D.K., Tong, F., 2011. Attention improves encoding of task-relevant features in the human visual cortex. J. Neurosci. 31, 8210–8219.
Jiang, J., Summerfield, C., Egner, T., 2013. Attention sharpens the distinction between expected and unexpected percepts in the visual brain. J. Neurosci. 33, 18438–18447.
Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8, 679–685.
Kastner, S., Pinsk, M.A., De Weerd, P., Desimone, R., Ungerleider, L.G., 1999. Increased activity in human visual cortex during directed attention in the absence of visual stimulation. Neuron 22, 751–761.
Kojadinovic, I., 2005. On the use of mutual information in data analysis: an overview. Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis (ASMDA '05), pp. 738–747.
Logothetis, N.K., 2003. The underpinnings of the BOLD functional magnetic resonance imaging signal. J. Neurosci. 23, 3963–3971.
Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., Oeltermann, A., 2001. Neurophysiological investigation of the basis of the fMRI signal. Nature 412, 150–157.
Luck, S.J., Chelazzi, L., Hillyard, S.A., Desimone, R., 1997. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. J. Neurophysiol. 77, 24–42.
Malach, R., Reppas, J.B., Benson, R.R., Kwong, K.K., Jiang, H., Kennedy, W.A., Ledden, P.J., Brady, T.J., Rosen, B.R., Tootell, R.B., 1995. Object-related activity revealed by functional magnetic resonance imaging in humanoccipital cortex. Proc. Natl. Acad. Sci. U. S. A. 92, 8135–8139.
Martínez, A., Anllo-Vento, L., Sereno, M.I., Frank, L.R., Buxton, R.B., Dubowitz, D.J., Wong, E.C., Hinrichs, H., Heinze, H.J., Hillyard, S.A., 1999. Involvement of striate and extrastriate visual cortical areas in spatial attention. Nat. Neurosci. 2, 364–369.
McAdams, C.J., Maunsell, J.H.R., 1999. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. J. Neurosci. 19, 431–441.
Motter, B.C., 1993. Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. J. Neurophysiol. 70, 909–919.
Mumford, J.A., Davis, T., Poldrack, R.A., 2014. The impact of study design on pattern estimation for single-trial multivariate pattern analysis. Neuroimage 103, 130–138.
Murray, S.O., He, S., 2006. Contrast invariance in the human lateral occipital complex depends on attention. Curr. Biol. 16, 606–611.
Murray, S.O., Wojciulik, E., 2004. Attention increases neural selectivity in the human lateral occipital complex. Nat. Neurosci. 7, 70–74.
O'Connor, D.H., Fukui, M.M., Pinsk, M.A., Kastner, S., 2002. Attention modulates responses in the human lateral geniculate nucleus. Nat. Neurosci. 5, 1203–1209.
O'Craven, K.M., Downing, P.E., Kanwisher, N., 1999. FMRI evidence for objects as the units of attentional selection. Nature 401, 584–587.
Pratte, M.S., Ling, S., Swisher, J.D., Tong, F., 2013. How attention extracts objects from noise. J. Neurophysiol. 110, 1346–1356.
Reddy, L., Kanwisher, N., 2007. Category selectivity in the ventral visual pathway confers robustness to clutter and diverted attention. Curr. Biol. 17, 2067–2072.
Reddy, L., Kanwisher, N.G., VanRullen, R., 2009. Attention and biased competition in multi-voxel object representations. Proc. Natl. Acad. Sci. U. S. A. 106, 21447–21452.
Ress, D., Backus, B.T., Heeger, D.J., 2000. Activity in primary visual cortex predicts performance in a visual detection task. Nat. Neurosci. 3, 940–945.
Reynolds, J.H., Heeger, D.J., 2009. The normalization model of attention. Neuron 61, 168–185.
Reynolds, J.H., Pasternak, T., Desimone, R., 2000. Attention increases sensitivity of V4 neurons. Neuron 26, 703–714.
Rice, G.E., Watson, D.M., Hartley, T., Andrews, T.J., 2014. Low-level image properties of visual objects predict patterns of neural response across category-selective regions of the ventral visual pathway. J. Neurosci. 34, 8837–8844.
Rolls, E.T., Baylis, G.C., 1986. Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. Exp. Brain Res. 65, 38–48.
Sackur, J., Dehaene, S., 2009. The cognitive architecture for chaining of two mental operations. Cognition 111, 187–211.
Saproo, S., Serences, J.T., 2010. Spatial attention improves the quality of population codes in human visual cortex. J. Neurophysiol. 104, 885–895.
Saproo, S., Serences, J.T., 2014. Attention improves transfer of motion information between V1 and MT. J. Neurosci. 34, 3586–3596.
Schurger, A., Pereira, F., Treisman, A., Cohen, J.D., 2010. Reproducibility distinguishes conscious from nonconscious neural representations. Science 327, 97–99.
Serences, J.T., Schwarzbach, J., Courtney, S.M., Golay, X., Yantis, S., 2004. Control of object-based attention in human cortex. Cereb. Cortex 14, 1346–1357.
Serences, J.T., Saproo, S., Scolari, M., Ho, T., Muftuler, L.T., 2009. Estimating the influence of attention on population codes in human visual cortex using voxel-based tuning functions. Neuroimage 44, 223–231.
Silver, M.A., Ress, D., Heeger, D.J., 2007. Neural correlates of sustained spatial attention in human early visual cortex. J. Neurophysiol. 97, 229–237.
Somers, D.C., Dale, A.M., Seiffert, A.E., Tootell, R.B.H., 1999. Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. Proc. Natl. Acad. Sci. U. S. A. 96, 1663–1668.

Steiger, J.H., 1980. Test for comparing elements of a correlation matrix. Psychol. Bull. 87, 245–251.
Tamber-Rosenau, B.J., Esterman, M., Chiu, Y.-C., Yantis, S., 2011. Cortical mechanisms of cognitive control for shifting attention in vision and working memory. J. Cogn. Neurosci. 23, 2905–2919.
Treue, S., 2003. Visual attention: the where, what, how and why of saliency. Curr. Opin. Neurobiol. 13, 428–432.
Treue, S., Martínez Trujillo, J.C., 1999. Feature-based attention influences motion processing gain in macaque visual cortex. Nature 399, 575–579.
Treue, S., Maunsell, J.H.R., 1999. Effects of attention on the processing of motion in macaque middle temporal and medial superior temporal visual cortical areas. J. Neurosci. 19, 7591–7602.
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation ofthe MNI MRI single-subject brain. NeuroImage 15, 273–289.
Williford, T., Maunsell, J.H.R., 2006. Effects of spatial attention on contrast response functions in macaque area V4. J. Neurophysiol. 96, 40–54.

6.4 Evidence for neural encoding of Bayesian surprise in human somatosensation

Journal: NeuroImage

Acceptance date: 26 April 2012

URL: http://dx.doi.org/10.1016/j.neuroimage.2012.04.05

NeuroImage 62 (2012) 177–188

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Evidence for neural encoding of Bayesian surprise in human somatosensation

Dirk Ostwald [a,b,*], Bernhard Spitzer [a], Matthias Guggenmos [a], Timo T. Schmidt [a], Stefan J. Kiebel [a,c], Felix Blankenburg [d]

[a] Department of Neurology and Bernstein Center for Computational Neuroscience, Charité Universitätsmedizin Berlin, Germany
[b] School of Psychology, University of Birmingham, UK
[c] Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[d] Dahlem Institute for Neuroimaging of Emotion, Freie Universität Berlin, Berlin, Germany

ARTICLE INFO

Article history:
Accepted 26 April 2012
Available online 3 May 2012

Keywords:
Bayesian brain hypothesis
Somatosensory mismatch response
EEG single trial modeling
Computational neuroimaging

ABSTRACT

Accumulating empirical evidence suggests a role of Bayesian inference and learning for shaping neural responses in auditory and visual perception. However, its relevance for somatosensory processing is unclear. In the present study we test the hypothesis that cortical somatosensory processing exhibits dynamics that are consistent with Bayesian accounts of brain function. Specifically, we investigate the cortical encoding of Bayesian surprise, a recently proposed marker of Bayesian perceptual learning, using EEG data recorded from 15 subjects. Capitalizing on a somatosensory mismatch roving paradigm, we performed computational single-trial modeling of evoked somatosensory potentials for the entire peri-stimulus time period in source space. By means of Bayesian model selection, we find that, at 140 ms post-stimulus onset, secondary somatosensory cortex represents Bayesian surprise rather than stimulus change, which is the conventional marker of EEG mismatch responses. In contrast, at 250 ms, right inferior frontal cortex indexes stimulus change. Finally, at 360 ms, our analyses indicate additional perceptual learning attributable to medial cingulate cortex. In summary, the present study provides novel evidence for anatomical-temporal/functional segregation in human somatosensory processing that is consistent with the Bayesian brain hypothesis.

© 2012 Elsevier Inc. All rights reserved.

Introduction

The Bayesian brain hypothesis postulates that the brain uses probabilistic inference for perception and perceptual learning (Doya et al., 2007). These mechanisms can be implemented using Bayesian inference based on an internal generative model, which comprises a distribution over sensory data given an external cause (the sensory data likelihood) and a prior distribution over different causes (Friston, 2010; Knill and Pouget, 2004). Perception is modeled as the process of computing a posterior distribution over causes using the generative model and sensory input, while perceptual learning is explained as the updating of the brain's representation of the prior distribution based on the inferred posterior distribution over causes (Friston, 2003; Kersten et al., 2004).

It has been suggested that these Bayesian mechanisms are encoded by neuronal populations whose responses to novel sensory input are interpreted as dynamics induced by the violation of prior expectations (Mumford, 1992; Rao and Ballard, 1999; Strange et al., 2005). Typical neurobiological markers of this violation are EEG novelty responses such as the auditory mismatch negativity (aMMN) (Näätänen et al., 2011), the P300 (Polich, 2007), or the dishabituation of laser evoked-potentials (Mouraux and Iannetti, 2008; Wang et al., 2010). In the framework of the Bayesian brain hypothesis, one way to formally quantify the novelty of sensory input is Bayesian surprise, a recently proposed information theoretic quantity (Baldi and Itti, 2010; Itti and Baldi, 2009). Bayesian surprise quantifies the effect sensory input has on the internal generative model as the divergence between the encoded prior and posterior distribution over causes. Representing Bayesian surprise may enable an observer like the brain to efficiently and dynamically encode the statistical (ir)regularities of its environment.

While empirical evidence suggests a role of Bayesian perceptual learning for shaping neural responses in auditory and visual perception (Garrido et al., 2009a, 2009b, 2009c; Harrison et al., 2007; Rao and Ballard, 1999), its relevance for somatosensory processing is unclear. Here, we address the hypothesis that somatosensory processing, as assessed with somatosensory mismatch responses (sMMRs) in EEG, exhibits dynamics that are consistent with Bayesian theories of perceptual learning and specifically, the encoding of Bayesian surprise.

Although less studied than the aMMN, a number of investigations have previously described novelty or mismatch responses for somatosensory evoked potentials (SEPs) (Näätänen, 2009). Consistently, a fronto-parietal negative shift between 100 and 200 ms contralateral to the side of stimulation has been observed for unexpected stimuli (Akatsuka et al., 2007b; Kekoni et al., 1997; Kida et al., 2004;

* Corresponding author at: Bernstein Center for Computational Neuroscience, Charité Universitätsmedizin Berlin, Philippstr. 13 Haus 6, 10115 Berlin, Germany. Fax: +49 30 2093 6771.
E-mail addresses: dirk.ostwald@bccn-berlin.de, dirk.ostwald@mpib-berlin.mpg.de (D. Ostwald).

1053-8119/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2012.04.050

Restuccia et al., 2007; Shinozaki et al., 1998; Spackman et al., 2007, 2010), while some studies additionally reported earlier mismatch responses 60 to 90 ms after stimulus onset (Akatsuka et al., 2007a, 2007b; Götz et al., 2011). Further, in analogy to other sensory modalities, somatosensory oddball stimuli that capture the observer's attention elicit a parietal positive response at about 300 ms post-stimulus, commonly referred to as P300 (Restuccia et al., 2009; Tarkka et al., 1996). While the precise neural generation mechanism of the sMMR remains to be established, it provides a unique experimental tool to investigate perceptual learning in the somatosensory system from a Bayesian perspective.

We investigate our hypothesis by capitalizing on recently developed model-based analyses of single-trial EEG or fMRI activity (Friston and Dolan, 2010; Mars et al., 2008, 2010). To derive single-trial estimates of Bayesian surprise we employ a sequential Bayesian stimulus probability learning algorithm and, to account for the assumption that the brain uses finite time-windows to dynamically update its generative model, we employ a forgetting mechanism in the learning scheme (Harrison et al., 2011; Kiebel et al., 2008a). Using EEG data from 15 subjects, we perform computational modeling of evoked somatosensory potentials on the source level and for the entire peri-stimulus time period. By means of Bayesian model selection, we identify both cortical substrates and critical time windows of ongoing Bayesian surprise encoding in the cortical-temporal hierarchy of somatosensory processing.

Materials and methods

Participants

Fifteen healthy volunteers (21–31 years, six females) participated in the experiment after providing written informed consent. The study was approved by the Ethical Committee of the Charité University Hospital Berlin and corresponded to the Human Subjects Guidelines of the Declaration of Helsinki.

Stimuli

Electrical stimuli of 0.2 ms duration were delivered to the left median nerve via adhesive electrodes attached to the wrist. Two intensity levels (low/high stimulus amplitude) were adjusted on an individual subject basis to account for subject specific sensory thresholds. The low stimulus intensity (mean 4.0 ± 1.6 STD mA) was determined to be close to detection threshold but clearly noticeable for every stimulus replication. The high stimulus intensity (6.0 ± 2.2 STD mA) was chosen to be markedly distinguishable from the low stimulus intensity, but not painful and below the motor threshold.

Experimental procedure

Upon familiarization with the experimental stimulation, participants underwent nine to ten experimental runs of an oddball-like roving paradigm (Baldeweg et al., 2004): stimuli were delivered in consecutive trains of alternating stimulus intensity with a constant inter-stimulus interval of 650 ms (Fig. 1). In contrast to classical oddball paradigms, which comprise the repeated presentation of standard stimuli occasionally interrupted by the presentation of physically different deviant stimuli (Näätänen et al., 1978), in roving paradigms stimuli with different physical properties can take on the role of both deviant (oddball) and standard stimulus. By averaging over deviant and standard potentials evoked by physically different stimuli, this allows to discount differential responses to the physical stimulus per se in observed mismatch effects.

The length of each stimulus train was chosen at random from the set {2,4,8,16}, using equal probabilities. The participants were instructed to count the number of stimulus trains per experimental run, i.e., to attend to the changes from low to high and high to low amplitude. Thereby, the applied paradigm differed from classical mismatch tasks in so far that participants directed their attention to the stimuli. To render the counting task nontrivial, the number of stimulus trains in each run was sampled at random from a normal distribution with expectation 72 and standard deviation 5. Consequently, ca. 72 stimulus trains corresponded to roughly 500 stimuli delivered per run, and to about 5000 stimuli per subject in total. As the first stimulus in a stimulus train (high or low stimulus amplitude) is, by definition, a deviant, approximately 720 deviant responses were recorded per subject. After each experimental run, the subjects reported the number of experimental trains and were informed about the correct outcome.

EEG recording and pre-processing

EEG data were recorded using a 64-channel active electrode system at a sampling rate of 2048 Hz (ActiveTwo, BioSemi), with electrodes placed in an elastic cap according to the extended 10–20 system. Individual electrode locations were registered with respect to three fiducial markers (left and right preauricular points and nasion) using an electrode positioning system (Zebris Medical) to improve subsequent source space analyses. All further data preprocessing steps were performed using Statistical Parametric Mapping (SPM8) (Litvak et al., 2011). Specifically, the data were downsampled to a sampling rate of 512 Hz, referenced against average reference, band-pass filtered (1 to 40 Hz) and corrected for eye-movements using a topological confound approach originally developed by Berg and Scherg (1994) and implemented in SPM8 (Litvak et al., 2007). The data were epoched using a peri-stimulus time interval of −100 ms to 600 ms. Trials containing amplitudes larger than 150 μV were excluded from further analysis. SEPs for experimental conditions of interest were computed using standard averaging and were averaged across subjects to yield grand mean SEPs. The experimental conditions of interest were 1) the somatosensory evoked potential averaged over all stimuli, abbreviated 'SEP', 2) the response to deviant stimuli, averaged across high and low intensity stimuli, abbreviated 'Deviant', 3) the response to stimuli immediately preceding deviant stimuli, averaged across high and low intensity stimuli, abbreviated 'Standard'. It should be noted that this nomenclature differs

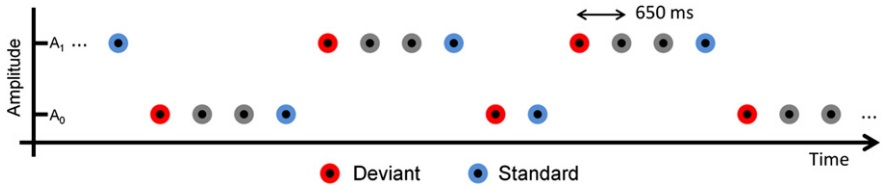


Fig. 1. Experimental paradigm. Electrical stimuli of two amplitudes, high (A_1) and low (A_0), were delivered to the median nerve with an inter-stimulus interval of 650 ms. Trains of identical stimuli, i.e. of either high or low amplitude, comprised 2, 4, 8 or 16 stimuli. For clarity only the cases of trains of 2 and 4 stimuli are shown in the figure. The first stimulus in each train of identical stimuli was labeled a deviant. To compare deviant (red) and standard (blue) responses based on the same number of trials, only those stimuli immediately preceding a deviant stimulus were labeled standard. This experimental paradigm is an adaptation of a previously established roving paradigm for the somatosensory domain (Baldeweg et al., 2004).

from standard mismatch negativity paradigms, which usually define as 'Standard' stimuli all non-deviant stimuli. However, defining the 'Standard' condition in this manner allows estimating both 'Deviant' and 'Standard' conditions from the same number of trials. Further, by collapsing the 'Deviant' and 'Standard' conditions over both stimulus amplitudes allows discounting differential responses to the physical properties of the stimulus per se.

Anatomical network identification

As we aimed to perform the model-based trial-by-trial analyses based on anatomically specific substrates (rather than on the electrode level), a combination of previous findings on the generators of the SEP as well as source reconstructions of the present data set was used to identify a set of six brain regions involved in the generation of the trial-by-trial EEG response. The anatomical localization of these sources enabled us to project the trial-by-trial electrode space EEG data onto a set of oriented equivalent current dipoles (ECDs) placed at corresponding locations as described below. Previous research established that the cortical SEP comprises the succession of (at least) three EEG components reflecting different processing stages of somatosensory information (Niedermeyer and Silva, 2004; Thees et al., 2003): 1) the parietal N20 reflecting a dipolar generator in primary somatosensory cortex (S1) situated in the posterior bank of the rolandic fissure, 2) the fronto-central P45/N60 complex of unclear origin, and 3) the parieto-temporal N140 peaking in the 60–160 ms latency range, presumably reflecting additional stimulus processing in secondary somatosensory cortex (S2). The localization of contralateral S1 and bilateral S2 for the current data set is hence obligatory. Moreover, experimental evidence suggests the involvement of frontal regions in the generation of mismatch responses, e.g., for the aMMN (Näätänen et al., 2011; Rinne et al., 2000, 2005, 2006; Tse and Penney, 2008). We thus were also interested in monitoring trial-by-trial activity in frontal cortex, the exact anatomical location being derived from source reconstruction of the deviant response of the present data set. Finally, cingulate cortex has been repeatedly implicated in the generation of oddball as well as P300 responses across a wide range of experimental paradigms (Linden, 2005; Thees et al., 2003), motivating the inclusion of a cingulate source.

In summary, strong anatomical–functional evidence suggests the inclusion of contralateral primary and bilateral secondary somatosensory, bilateral inferior frontal, and cingulate cortex in a six-dipole model of single-trial EEG responses for the current paradigm. To identify the precise anatomical substrates in the current study in a data-informed manner, we employed a two-stage procedure to obtain MNI coordinates and moment vectors for a multiple dipole model: 1) Using grand mean event-related potentials (ERPs) a distributed source localization of the regions suggested by the literature was performed to obtain MNI coordinates, and 2) the moments of ECDs located at these MNI coordinates were fitted, again based on grand mean ERPs. The distributed source localization and dipole orientation fitting procedures are detailed below. Both the condition-specific grand mean ERPs and time windows used in these analyses were chosen to be consistent with previous findings on the neural generators of somatosensory and mismatch ERPs as discussed in the section above and are documented in detail in Tables 1 and 2. The final set of six location and six moment vectors was subsequently used to project the data of each trial, for each subject, to anatomical brain space.

Statistical distributed source localization

At the first stage, the sources of the evoked EEG activity were determined using the distributed source reconstruction algorithm as implemented in SPM8. A forward model was constructed for each subject using a 8196 vertex template cortical mesh co-registered at the individual electrode positions via three fiducial markers. The lead field for the forward model was generated using the three-shell BEM EEG head model as provided by SPM8. Source estimates were computed on the canonical cortical mesh using multiple sparse priors (Friston et al., 2008) under group constraints (Litvak and Friston, 2008). Source power increases were statistically analyzed at the group level using one-sample t-tests. For display purposes statistical parametric maps were thresholded at $p<0.001$ (uncorrected), and random field theory was used to control for family-wise error in source space (Kiebel and Friston, 2004; Worsley, 1994). Finally, the SPM Anatomy toolbox was employed to establish cytoarchitectonic references (Eickhoff et al., 2005). This procedure enabled the reliable identification of four of the six sources (Table 1). The exact procedure used for obtaining these four sources and the remaining two sources to form a six-dipole model is described below under results.

Dipole fitting and timecourse extraction

At the second stage, the six obtained MNI source locations were used to fit ECDs to project the evoked electrode data into source space. To this end, grand mean group evoked potentials were subjected to the Variational Bayes — Equivalent Current Dipole (VB-ECD) algorithm implemented in SPM8 (Kiebel et al., 2008b). For each source, either a single dipole or a symmetric dipole pair was chosen with tight location priors centered on the sets of coordinates obtained from the distributed source analysis, unless otherwise noted. The moment parameters of the respective dipolar sources were then optimized using VB-ECD, and the posterior moment expectations normalized to a Euclidean norm of 1. The dipole specific SEPs, prior locations and moment expectations, as well as the peri-stimulus time-points analyzed are listed in Table 2. Finally, the trial-by-trial event-related electrode space data were projected onto the set of fixed and oriented ECDs using SPM8's spm_eeg_extract_waveforms.m function (Litvak et al., 2011).

Functional model of evoked source activity

To relate single-trial source activity to Bayesian principles of perceptual learning, we computed Bayesian surprise for each single trial using a sequential (online) Bayesian learning algorithm of stimulus probabilities (Bishop, 2007) (pp. 68–78). Briefly, the model assumes that the brain implements a trial-by-trial Bayesian parameter learning scheme starting from an uninformative prior and computes Bayesian surprise as the divergence between the parameter prior and posterior probability density functions (PDF) at the single-trial level. Moreover, variants of this model assume that the brain only incorporates observed trials which lie in a variable time-window of the close past into its estimation of the parameter PDF, where the length of the time-window is governed by an exponential forgetting (i.e. relative down-weighting) of stimulus observations in the distant past.

Formally, the model employed here assumes that the probability of observing a low ($S=0$) or high ($S=1$) intensity stimulus on the n-th trial is described by a Bernoulli distribution with expectation $\mu\in[0,1]$

$$p(S)=\mu^{S}(1-\mu)^{1-S}. \tag{1}$$

Here μ is the probability of observing a stimulus of high intensity on the n-th trial. To model the initial uncertainty about the parameter μ the model assumes a uniform beta prior distribution over μ, which, on each trial, is sequentially updated according to the observed data likelihood to form a posterior distribution over μ. The posterior distribution over μ after observing l stimuli of low intensity and m stimuli of high intensity, i.e. after $l+m$ trials in total, is given by

$$p(\mu|m,l)=\frac{\Gamma(m+l)}{\Gamma(m)\Gamma(l)}\mu^{m}(1-\mu)^{l} \tag{2}$$

Table 1
Distributed source reconstruction statistics.

Label	ERP	PST window	p-Cluster (FWE)	Z-value	p-Peak (FWE)	MNI coordinates			Cytotechtonic reference
S1	SEP	18–25 ms	<0.001	5.87	<0.001	48	−30	50	Right postcentral gyrus
				5.78	<0.001	36	−30	64	Area 3b (90% [40–100%])
				5.76	<0.001	44	−32	60	Area 1 (50% [40–80%])
						42	**−31**	**58**	Area 4a (20% [0–20%])
rS2	SEP	30–160 ms	0.001	4.14	0.074	**62**	**34**	**12**	Right superior temporal gyrus
									IPC (PFcm): 30% [10–40%]
									IPC (PF): 20% [0–50%]
									(Activation extending into OP1 (60% [20–60%]))
lS2			0.117	3.92	0.151	**−60**	**44**	**12**	Left superior temporal gyrus
rIFG	Deviant	340–360 ms	0.001	3.91	0.165	48	16	12	Right inferior frontal gyrus
				3.82	0.214	36	24	8	(pars triangularis)
				3.81	0.220	40	30	2	Area 45 40% [20–50%]
						41	**23**	**7**	

Column 1: Based on previous findings in the literature, the group distributed source reconstruction method implemented in SPM8 (Litvak and Friston, 2008) was used to determine the MNI coordinates of four sources of interest (S1, rS2, lS2, rIFG). Column 2: The statistical evaluation was based on reconstructing the source activity of the 'SEP' and the 'Deviant' waveform. Column 3: Peri-stimulus times of interest. These were derived from the grand mean, see Fig. 3 and text. Column 4: Corrected p-values obtained by using one-sample t-tests and family-wise error (FWE) correction at the cluster level (Worsley, 1994). Columns 5–7: Up to three peak MNI coordinates more than 8 mm apart, their corresponding Z and p-values at the voxel level. For multiple peak clusters, the arithmetic mean of the peak coordinates was used as the source MNI coordinate, set in bold in column 7. These coordinates were entered into SPM8's Anatomy toolbox (Eickhoff et al., 2005) to obtain an anatomical and probabilistic cytotechtonic reference, as reported in column 8.

where $\Gamma:\mathbb{R}\to\mathbb{R}$ denotes the gamma function. Corresponding to the sequential Bayesian learning approach, this distribution then acts as the prior distribution for inference on the $(l+m+1)$-th trial.

In order to implement a forgetting kinetic, instead of using the accumulative stimulus counts l_n and m_n on the $l_n+m_n=n$-th trial, the model employed here weights the stimulus counts with an exponential function over trials, such that

$$l_{w_n} = \sum_{i=0}^{n} \exp\left(-\frac{1}{\tau}(n-i)\right) l_i \tag{3}$$

and

$$m_{w_n} = \sum_{i=0}^{n} \exp\left(-\frac{1}{\tau}(n-i)\right) m_i \tag{4}$$

yield the weighted stimulus counts m_{w_n}, l_{w_n} at the n-th trial and $\tau>0$ denotes the time constant of the forgetting kinetic. This results in a modulated posterior distribution on the n-th trial/prior distribution on the $(n+1)$-th trial given by

$$p\left(\mu|m_{w_n}, l_{w_n}\right) = \frac{\Gamma\left(m_{w_n}+l_{w_n}\right)}{\Gamma\left(m_{w_n}\right)\Gamma\left(l_{w_n}\right)} \mu^{m_{w_n}} (1-\mu)^{l_{w_n}}. \tag{5}$$

With respect to the Bayesian brain hypothesis, the product of the sensory data likelihood (1) and prior distribution (5) form the internal generative model over the external cause μ at the $(n+1)$-th trial, the formation of the posterior distribution (5) corresponds to perception on the n-th trial, and the iterative exchange of prior and posterior distributions based on weighted stimulus counts corresponds to perceptual learning with forgetting.

Finally, the model quantifies its degree of perceptual learning on the n-th trial as Bayesian surprise, i.e. the Kullback–Leibler divergence

Table 2
Equivalent current dipoles.

Label	ERP	PST point	Location prior exp.			Moment prior exp.			Location posterior exp.			Moment posterior exp.		
S1	SEP	21 ms	42	−31	58	0	0	0	42	−31	58	−0.0037	0.5743	0.8186
			(Informative)			(Uninformative)								
rS2	SEP	140 ms	62	−34	12	0	0	0	62	−34	12	−0.8890	0.4335	0.1766
			(Informative symmetric pair)			(Uninformative)								
lS2			−62	−34	12	0	0	0	−62	−34	12	0.8279	0.5174	0.2162
			(Informative symmetric pair)			(Uninformative)								
rIFG	Deviant	351 ms	41	23	7	0	0	0	41	23	7	0.0832	0.0184	0.9964
			(Informative symmetric pair)			(Uninformative)								
lIFG			−41	23	7	0	0	0	−41	23	7	0.0886	0.0184	0.9959
			(Informative symmetric pair)			(Uninformative)								
MC	Deviant–standard	351 ms	0	0	0	0	0	0	−1	−21	36	0.0253	0.3975	0.9172
			(Uninformative)			(Uninformative)								

After MNI coordinates for the anatomical regions of interest had been determined (Table 1), the normalized ECD moments were obtained by using SPM8's VB-ECD method with informative prior location expectation and uninformative prior moment expectation (Kiebel et al., 2008b). Column 1: Sources of interest with medial cingulate cortex (MC). Columns 2 and 3: The event-related potentials (ERP) and peri-stimulus time points (PST point) analyzed for each source. For the bilateral S2 and the IFG sources, a coupled symmetric ECD pair was fitted, where the location prior expectation was set to the MNI coordinates of the right hemisphere sources and their homologue coordinates. Columns 6 and 7: The posterior means of the location and moments (normalized). For the MC source, a single ECD model with uninformative location and moment prior was used, i.e. the location of the MC source is based solely on the VB-ECD solution.

between the prior and posterior distribution over μ at trial n (Baldi and Itti, 2010; Cover and Thomas, 1991; Itti and Baldi, 2009).

$$\begin{aligned}\text{Bayesian Surprise} &:= \mathrm{KL}\left(p\left(\mu|m_{w_{n-1}}, l_{w_{n-1}}\right)||p\left(\mu|m_{w_n}, l_{w_n}\right)\right)\\ &= \int p\left(\mu|m_{w_{n-1}}, l_{w_{n-1}}\right)\ln\left(\frac{p\left(\mu|m_{w_{n-1}}, l_{w_{n-1}}\right)}{p\left(\mu|m_{w_n}, l_{w_n}\right)}\right)d\mu\end{aligned} \tag{6}$$

Due to the use of conjugate priors, the Kullback–Leibler divergence can be evaluated analytically as a function of the parameters $m_{w_{n-1}}, l_{w_{n-1}}$ and m_{w_n}, l_{w_n} which significantly simplifies the integration over μ (Penny, 2001).

We used this approach to generate subject- and session-specific trial-by-trial Bayesian surprise sequences using the output of Eq. (6) for each single trial. These sequences were used as regressor/predictor variables in the trial-by-trial EEG data analysis, see next section. To implement varying degrees of forgetting, we varied the time-window of temporal stimulus integration from long to short, (Eqs. (3) and (4)), yielding a set of five Bayesian surprise models with different time constants τ, abbreviated BS0, BS1, BS2, BS3 and BS4, where we chose $\tau_0 = \infty, \tau_1 = 8, \tau_2 = 4, \tau_3 = 2.\bar{6}$ and $\tau_4 = 2$. This choice of time constants was motivated by sampling a wide range of possible temporal integration windows, while simultaneously maximizing the differences between the resulting regressors. Note that larger time-constants τ correspond to longer time windows. We show an example for one of the Bayesian surprise regressors ($\tau_2 = 4$) in Fig. 2. In Table 3, we translate the five time-constants to the weighting of the trial history.

Additionally, we generated three control regression models that implement less complex but widely used conventional hypotheses about the functional origin of trial-by-trial source amplitude variations: 1) a regressor indexing deviant stimuli with 1's and standard stimuli with 0's ('stimulus change model', model SC), 2) a parametric model implementing a linear relationship between the expression of the evoked source activity and the number of standards preceding a deviant stimulus ('linearly modulated stimulus change model', model LIN), and 3) a constant null regressor comprising a vector of 1's (model M0). Model M0 was used as a baseline model to compare all other models against. Importantly, note that model SC is the standard model for the analysis of (auditory) mismatch negativity studies. For the roving paradigm, model LIN has also been used in modified form in Baldeweg et al. (2004).

Functional model evaluation using parametric empirical Bayes

Each of the functional models provides a single, stimulus sequence-specific regressor. We used a parametric empirical Bayes (PEB) approach as implemented in SPM8's spm_PEB.m function (Friston et al., 2002a, 2002b, 2007) to fit the models and to compute the corresponding Bayesian model evidences for subsequent model selection (Gelman et al., 1995; Penny, 2012). The Bayesian model evidence framework enables the formal statistical comparison of computational models by accounting for the accuracy-complexity trade-off in explaining experimental data and constitutes a well-established approach in statistics (Hoeting et al., 1999), computational psychology (Pitt and Myung, 2002), and neuroimaging (Woolrich, in press).

To this end, the single subject, single session, single peri-stimulus time bin data for $n \in \mathbb{N}$ trials was modeled according to the two-level linear model

$$p(\bar{y}|\lambda) = N(\bar{y}; \bar{X}\bar{\theta}, C(\lambda) + C_\theta) \tag{7}$$

i.e., the probability distribution of the augmented data $\bar{y} \in \mathbb{R}^{n+2}$ was assumed to be multivariate normal with expectation $\mu = \bar{X}\bar{\theta} \in \mathbb{R}^{n+2}$ and parameterized covariance $\bar{\Sigma} = (C(\lambda) + C_\theta) \in \mathbb{R}^{n+2\times n+2}, (C(\lambda) + C_\theta) \succ 0$, where $\lambda \in \mathbb{R}^2$ refers to the covariance constraint weighting coefficients, the free parameters of the hierarchical linear model. Following the notation in Friston et al. (2002a, 2002b, 2007), the augmented data is given by

$$\bar{y} = \begin{pmatrix} y \\ 0 \\ 0 \end{pmatrix} \in \mathbb{R}^{n+2} \tag{8}$$

where $y \in \mathbb{R}^n$ refers to the single subject, single session (comprising $n \in \mathbb{N}$ trials), single peri-stimulus time bin data, which, following standard approaches to multiple linear regression, was normalized to a mean of zero and a variance of 1 (z-score normalization). The two additional zeros encode the expectation of the second level error and the prior of the second level parameters. The augmented design matrix is given by

$$\bar{X} = \begin{pmatrix} X^{(1)} & X^{(1)}X^{(2)} \\ & I_2 \end{pmatrix} \in \mathbb{R}^{n\times 2} \tag{9}$$

where $X^{(1)} \in \mathbb{R}^{n\times 1}$ denotes the BS0–BS4/SC/LIN/M0-model specific regressor normalized to a mean of zero and a l_2-norm of 1 (except for the regressor of model M0, which was not normalized), $X^{(2)} = 0 \in \mathbb{R}$ is the second level design matrix allowing single level Bayesian inference with priors on the parameters, and $I_2 \in \mathbb{R}^{2\times 2}$ is the identity matrix.

Further,

$$\bar{\theta} = \begin{pmatrix} \varepsilon^{(2)} \\ \theta^{(2)} \end{pmatrix} \in \mathbb{R}^2 \tag{10}$$

is a vector of latent variables corresponding to the second level error and linear parameter obtained by substitution of the hierarchical form of the model

$$y = X^{(1)}\theta^{(1)} + \varepsilon^{(1)} \tag{11}$$

$$\theta^{(1)} = X^{(2)}\theta^{(2)} + \varepsilon^{(2)}. \tag{12}$$

The parameterized covariance of the model $C(\lambda)$ is given by

$$C(\lambda) = \textstyle\sum_{i=1}^{2} \lambda_i Q_i \tag{13}$$

where

$$Q_1 = \begin{pmatrix} I_n & 0 \\ 0 & 0_2 \end{pmatrix} \in \mathbb{R}^{n+2\times n+2} \tag{14}$$

with the identity matrix $I_n \in \mathbb{R}^{n\times n}$ and the zero matrix $0_2 \in \mathbb{R}^{2\times 2}$ embeds the independence assumption over trials (justified by the separation of neighboring trials by 650 ms) and

$$Q_2 = \begin{pmatrix} 0_{nn} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{n+2\times n+2} \tag{15}$$

with the zero matrix $0_n \in \mathbb{R}^{n\times n}$ embeds the second level covariance constraint. Finally, an uninformative prior for the second level parameters was specified by setting

$$C_\theta = \begin{pmatrix} 0_{n+1} & 0 \\ 0 & \exp(32) \end{pmatrix} \in \mathbb{R}^{n+2\times n+2}. \tag{16}$$

Upon specification of the hierarchical linear model for each single-trial regressor, the model parameters $\lambda \in \mathbb{R}^2$ were estimated using an EM algorithm for maximum likelihood estimation and the model log-evidence was approximated using the variational free energy (Friston et al., 2007; Garrido et al., 2007, 2009a). For single subjects, the model

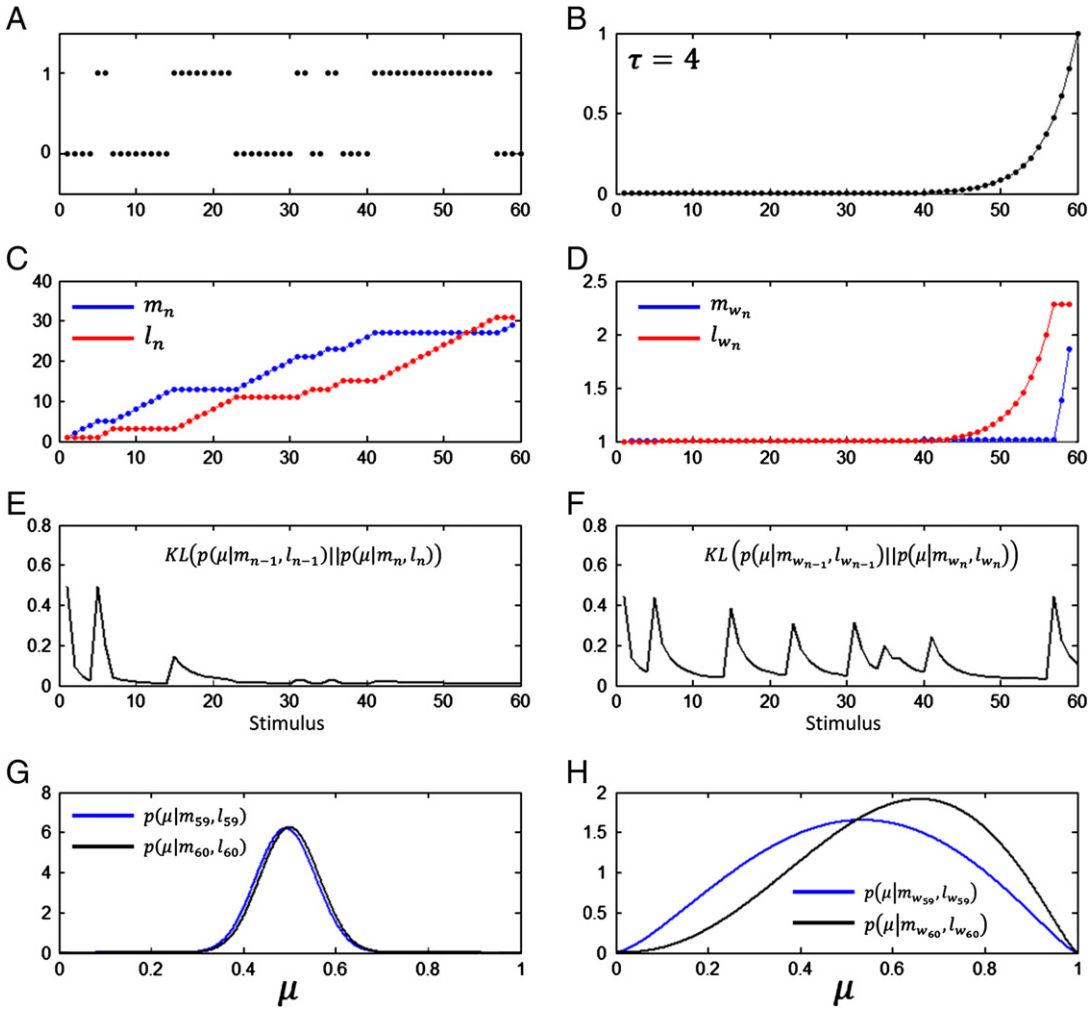


Fig. 2. Computational model. A) Typical stimulus sequence of 60 stimuli alternating between two stimulus amplitudes $A_0 = 0$ and $A_1 = 1$. B) Stimulus-specific weights implemented a forgetting kinetic with a time constant $\tau = 4$ for the 60th stimulus. C) Increase of the beta distribution parameters m_n and l_n over trials, implementing Bayesian learning up to stimulus 60 for no trial weighting. D) Result of applying the weighting function of B) to the temporal evolution of m_n and l_n, resulting in the weighted parameters m_{w_n} and l_{w_n}. E) Illustration of the Bayesian surprise regressor without forgetting or infinite time of integration constant $\tau = 0$ for the stimulus sequence depicted in A). The predicted surprise for this stimulus sequence is large at the beginning and for the first switch of stimulus amplitudes, but close to zero for the remaining amplitude switches. F) Bayesian surprise predictor obtained for the stimulus sequence depicted in A), but under application of the forgetting kinetic shown in B). Note that the amount of Bayesian surprise decreases with the number of stimuli in an identical train of stimuli and increases with the number of preceding stimuli of the opposite amplitude. G) Prior and posterior probability density functions over the parameter μ for the 60th trial of the stimulus sequence shown in A) without forgetting. H) Prior and posterior probability density functions over the parameter μ for the 60th trial of the stimulus sequence shown in A) under the forgetting kinetic depicted in B).

log-evidences across experimental runs were averaged to obtain an estimate of the subject-specific model log-evidence for each model. Following Garrido et al. (2007, 2009a), the single-subject model log-evidences were summed over subjects to yield the group log-evidence for each model, as a function of dipole and peri-stimulus time-bin. Additionally, for a subset of time windows, the group log-evidences were averaged over time and the pairwise differences (i.e. log Bayes factors) between models were plotted. The ensuing difference maps are thresholded at a group log-evidence difference of 3, indicating strong evidence of a particular model, compared to another model (Penny et al., 2004).

Table 3
Interpretation of the Bayesian surprise model time constants τ_0 to τ_4.

BS model	τ	63% (s/stim)	99% (s/stim)
BS0	∞	∞	∞
BS1	8	5.2/8	26.0/40
BS2	4	2.6/4	13.0/20
BS3	2.6	1.7/6	8.6/13.3
BS4	2	1.3/2	6.5/10

Column 1: The five Bayesian surprise models BS0–BS4 were derived from the same set of governing Eqs. (1)–(6), but with varying values of the time constant τ in Eqs. (3) and (4) listed in Column 2. Columns 3 and 4 list the time in seconds and the number of stimuli (ISI 650 ms) corresponding to a 63% and 99% down-weighting of past observations for each of the models/time constants.

Results

Event-related potentials

Grand mean event-related potentials and electrode space results for the somatosensory mismatch response are shown in Fig. 3. Inspection of the grand mean SEP at the channel level, obtained by averaging over all experimental trials, confirmed the presence of well-established SEP components (N20, N45/P60, N140, Fig. 3A). Fig. 3B depicts the 'Deviant–Standard' difference waveform, i.e. the difference between averages evoked by deviant and their immediately preceding standard stimuli (averaged over both stimulus intensities). Fig. 3C depicts the grand mean SEP and the 'Deviant–Standard' waveforms averaged over electrodes over contralateral (C4, C6, CP4, CP5) and ipsilateral (C3, C5, CP3, CP5) somatosensory cortices, showing the expected pattern of stronger contralateral responses. Across all

electrodes (Fig. 3B), early sMMR effects were observed around 40 and 85 ms post-stimulus, whereas more pronounced effects were observed at 140, 200, and 350 ms post-stimulus. The corresponding topographies of these differences across electrodes are shown in the upper row of Fig. 3D. The two early effects display fronto-central and parieto-central negativities. The sMMR at 140 ms exhibits a bilateral centro-parietal topography with a stronger contralateral negativity, the sMMR at 200 ms a fronto-contralateral negativity and the sMMR at 350 ms a fronto-central positivity. To establish the reliability of these effects across subjects, we performed paired two-sample t-tests between 'Deviant' and 'Standard' responses in the time windows indicated by the gray bars in Fig. 3B. For the electrodes exhibiting the most pronounced effects as shown in the lower row of Fig. 3D, these differences were significant ($p_{FWE}<0.05$, Bonferroni corrected for the number of electrodes) for the effects observed at 40, 140, and 350 ms post-stimulus (Table 4). The T-value plots across post-stimulus time (Fig. 3D, middle row) indicate a relatively low between-subject variability in the expression of the observed effects. In summary, a reliable sMMR was recorded using the current experimental paradigm.

Distributed source localization and ECD orientation fitting

Figs. 4A–C summarizes the statistical analysis of the distributed source reconstruction results and Fig. 4D shows the set of oriented ECD sources used as basis for the trial-by-trial analyses (cf. Tables 1 and 2 for details).

To localize sources, we selected time windows of interest based on the peak times of the most prominent deflections in both the 'SEP' and the 'Deviant–Standard' difference response (Figs. 3A, B). To localize S1, the SEP was reconstructed in a time-window of 18–25 ms (i.e. around the N20 effect, Fig. 3A) after stimulus onset, resulting in the expected activation pattern of contralateral S1 (Fig. 4A, Table 1). Likewise, reconstruction of the SEP in a time-window of 130–160 ms (i.e., around the N140 effect, Fig. 3A) after stimulus onset resulted in bilateral activation of posterior S2 (Fig. 4B, Table 1). Right inferior frontal gyrus activity is typically implicated in the response to the deviant and was therefore located by reconstructing the 'Deviant' response in a time window 340–360 ms (as identified from the difference response, Fig. 3B) after stimulus onset (Fig. 4C, Table 1). In accordance with previous studies on the aMMN this source was mirrored for the left hemisphere to derive a symmetric frontal source pair (Rinne et al., 2000, 2005). The distributed source localization analysis did not reveal a significant activation of cingulate cortex for the 'Deviant' condition. Hence, the VB-ECD method with uninformative prior location was used to spatially localize this source at the time point of maximal expression identified from the 'Deviant–Standard' difference response, i.e., at 351 ms, Fig. 3B. The deficiency of the distributed source localization analysis to detect activation of cingulate cortex

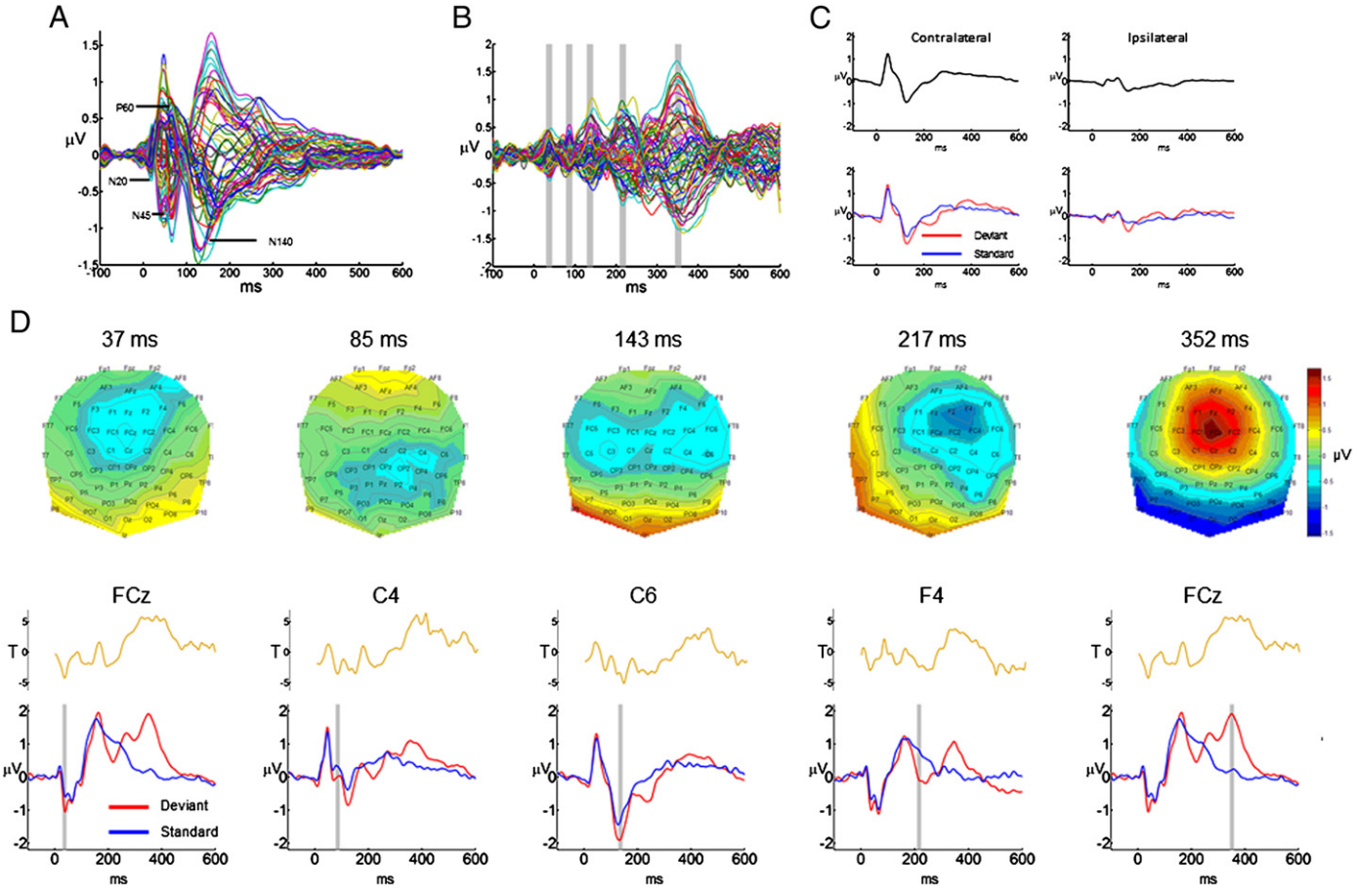


Fig. 3. Event related potentials. A) Grand mean SEP across all stimuli and subjects, for all electrodes. The classical SEP peaks (N20, N45/P60 and N140) are labeled. B) Grand mean 'Deviant–Standard' difference waveform, for all electrodes. The largest differences between deviant and standard SEPs are observed in time windows around 140, 200 and 350 ms post-stimulus, while smaller differences are found around 40 and 85 ms post-stimulus. In conjunction, the difference waveforms up to 200 ms are here referred to as sMMR, while the difference around 350 ms reflects the P300. C) Grand mean SEP (upper row) and grand mean 'Deviant' and 'Standard' waveforms (lower row) averaged over contralateral (C4, C6, CP4, CP6, left panels) and ipsilateral (C3, C5, CP3, CP5, right panels) electrodes above somatosensory cortices. D) Upper row: topographies for the 'Deviant–Standard' difference waveform for all electrodes at the time points indicated by the gray bars in panel B. Middle row: post-stimulus T-values across time for the contrast 'Deviant–Standard' for the electrodes which express the largest 'Deviant–Standard' difference effects for time points 40, 85, 140, 200, and 350 ms. Lower row: the peri-stimulus waveforms for the 'Deviant' (red) and 'Standard' (blue) conditions are shown for the selected electrodes.

may be most likely due to the known bias of source imaging methods to superficial sources (Fuchs et al., 1999; Michel et al., 2004) which is alleviated, but not abolished, by the multiple sparse prior approach used in the current study (Friston et al., 2008). As discussed above, previous findings in the literature as well as the observed potential topography for deviant responses in the current study speak for the inclusion of a cingulate source in the ECD model. As a caveat, based on the approach taken here, we cannot rule out the possibility that activity in cortical areas other than cingulate cortex is captured by the mediate cingulate source. Hence, the topographical interpretation of our modeling results for this source exhibits some uncertainty. On the other hand, the interpretation of the temporal–functional specificity of the modeling results for this source is unaffected by any spatial misattribution of scalp EEG activity.

After MNI coordinates were derived for each of the sources, the VB-ECD method with informative location priors was used to determine the normalized moments for ECDs located at the respective MNI coordinates as documented in Table 2.

Plausibility of the anatomical source model

To investigate whether the data reduction furnished by projecting the electrode space data onto the six ECDs of fixed orientation is sensible, the activity time-courses of the six ECDs were computed for the grand mean SEP (Fig. 5A). This plot indicates a neurobiologically plausible spatiotemporal activity pattern: The S1 dipole captures most of the early N20 activity, while for the N140 activity the S2 and IFG sources contribute most. Late components >200 ms post-stimulus are captured by a mixture of sources, but in particular by the MC source.

To investigate to which degree the data reduction onto a limited set of basis vectors was able to capture the grand mean SEP activity, the channel percent variance explained (PVE) was computed over time (Fig. 5B). Performing the analysis for hierarchical subversions of the six dipole model comprising 1) only S1, 2) S1 and bilateral S2, 3) S1, bilateral S2, bilateral IFG and 4) the complete six ECD model revealed that the PVE over time was largest for the complete model, in particular between 120 and 300 ms.

In summary, based on both the results of previous studies and the present analysis, we concluded that the six ECD model forms an appropriate anatomical basis for the evaluation of trial-by-trial EEG data.

Model-based trial-by-trial analyses

Having established a reliable recording of the sMMR in electrode space and a plausible anatomical source reconstruction, we studied the functional specialization of each of the six sources. To this end, the peri-stimulus trial-by-trial electrode data were projected onto the identified set of ECDs, each peri-stimulus time-point's trial-by-trial ECD activity was modeled using seven different functional models (BS0-BS4, SC and LIN), and for each model, the group model log-evidence was determined using parametric empirical Bayes.

In Fig. 6A, each ECD-specific panel depicts the group model log-evidences for the seven models, relative to the constant model M0, over peri-stimulus time. As indicated by the prominent peaks in the log model-evidence maps in three post-stimulus time-windows (around 140, 250, and 360 ms, dotted rectangles) for three different anatomical source ECDs (rS2, rIFG, and MC), Bayesian surprise encoding exhibits a high degree of anatomical–temporal specificity. In the following, we provide a detailed account of this key result by considering the model comparisons for each time-window and ECD source in turn.

In the time window around 140 ms post-stimulus, single-trial variability of contra-lateral secondary somatosensory cortex is best explained by Bayesian surprise models with forgetting (Fig. 6A, rS2, models BS1–BS4). Pair-wise comparison of the time-averaged model log-evidences (Fig. 6B, rS2, 109–171 ms) shows that for this ECD source and time-window, these models explain the data better than the Bayesian surprise model without forgetting (BS0) and both conventional models (stimulus change/linearly modulated stimulus change, SC and LIN, respectively). Next, in the time window around 250 ms post-stimulus, the highest model log-evidences are observed for right inferior frontal cortex (Fig. 6A, rIFG). Here, the Bayesian surprise models with forgetting (BS2–BS4) and the conventional stimulus change (SC) model explain the data best. Specifically, the SC model performs better than all models except the BS4 model for the time-integrated model log-evidence (Fig. 6B, rIFG, 210–291 ms) and at 254 ms supervenes the BS4 model in its ability to explain the data (model log-evidence difference SC − BS4: 3.5). Finally, in a time window around 360 ms post-stimulus, the largest model log-evidences are observed for mediate cingulate cortex (Fig. 6A, MC). Here, Bayesian surprise models with forgetting (BS2–BS4) perform best over an extended period of time, with model BS4 explaining the observed data better than all other models except BS3 (Fig. 6B, panel MC, 310–416 ms). In contrast to the results for secondary somatosensory cortex at 140 ms, the superiority of the Bayesian surprise models over the stimulus change model is less pronounced (model log-evidence difference for rS2 at 140 ms BS4 − SC: 15.5 vs. model log-difference for MC at 365 ms BS4 − SC: 6.9).

Besides these main log-evidence peaks for the rS2, IFG and MC sources, Fig. 6B indicates that also for sources S1 and lS2, the Bayesian surprise models provide better accounts of the data than the conventional/BS0 models during early processing (Fig. 6B, 109–171 ms). For intermediate processing (210–291 ms), for S1 only, the Bayesian surprise models BS3/BS4 are better than the conventional/BS0 models. The log-evidence map for the lIFG source shows only a limited degree of temporal variation and no clear superiority for any of the models. Finally, over all sources and time-windows, model BS0 finds the least support by the data.

Table 4
Statistical evaluation of the evoked 'Deviant–Standard' difference waveform.

Electrode label	Time window	T-value (df)	p-value
FCz	29–44 ms	$T(14) = -3.96$	$p = 0.001$ (sig.)
C4	78–93 ms	$T(14) = -2.86$	$p = 0.012$ (n.s.)
C6	129–145 ms	$T(14) = -4.33$	$p = 0.001$ (sig.)
F4	209–225 ms	$T(14) = -2.36$	$p = 0.033$ (n.s.)
FCz	344–359 ms	$T(14) = 5.55$	$p < 0.001$ (sig.)

Table 4 reports the results of a series of paired two-sample t-tests for statistical significance of the difference between 'Deviant' and 'Standard' waveforms in the time-windows expressing the most pronounced potentials across subjects (see Fig. 3B). Columns 1 and 2: For each time-window, the average potential at the electrode expressing the maximum potential at the group level was computed for each subject and condition (standard/deviant) and subjected to a paired two-sample t-test. Columns 3 and 4: Statistical significance (sig.) was established using Bonferroni correction for multiple testing over electrodes at a level of $p_{FWE} < 0.05$.

Discussion

In the current study we have shown that EEG markers of central somatosensory processing exhibit dynamics that are consistent with the Bayesian brain hypothesis of perceptual learning. Using a model-based approach for the analysis of trial-by-trial EEG source activity, we were able to demonstrate spatiotemporal specific encoding of Bayesian surprise as expressed by the present sequence processing models (Eqs. (1) to (6)). Moreover, we could show that the resulting models are, for some critical processing stages, better in explaining cortical source activity than conventional models that are typically used to analyze mismatch negativity studies.

Specifically, we find that a lower level source (rS2), at an early processing stage (140 ms), is more prominently involved in the representation of Bayesian surprise than in the representation of modulated stimulus change as assessed by conventional models. Critically, in the

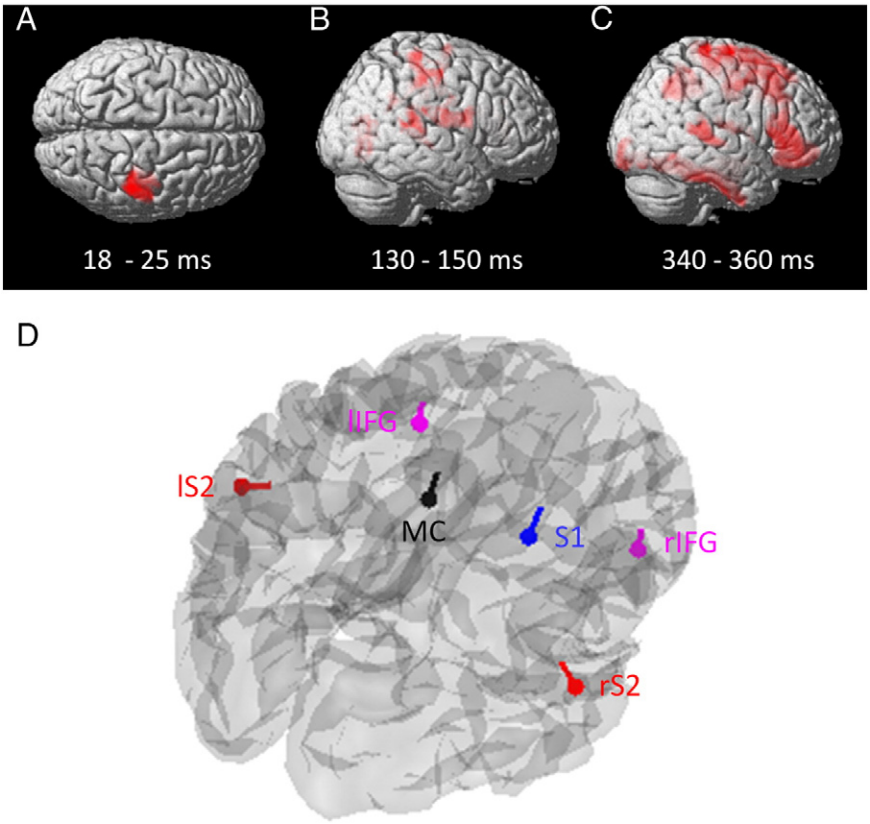


Fig. 4. Distributed source reconstruction and ECD fitting results. A–C) Statistical group distributed source reconstruction results for time-windows of 18–25 ms, 130–150 post-stimulus of the 'SEP' (A and B) and 340–360 ms post-stimulus of the 'Deviant' ERP. The p-value maps displayed were thresholded at $p<0.001$ (uncorrected) and overlaid onto SPM8's standard single subject brain. D) The complete six source ECD model with normalized moments obtained by VB-ECD overlaid on the standard MNI cortical mesh provided by SPM8 (S1: right primary sensory cortex, rS2: right secondary somatosensory cortex, lS2: left secondary somatosensory cortex, rIFG: right inferior frontal gyrus, lIFG left inferior frontal gyrus, MC: medial cingulate cortex).

present study, Bayesian surprise acts as a marker of perceptual learning by signaling the adjustment of the brain's internal generative model. Our finding is hence in line with previous imaging and electrophysiological studies that implicate secondary somatosensory cortex in fast perceptual learning (Pleger et al., 2003; Romo et al., 2003). In contrast, for a subsequent stimulus processing stage around 250 ms, a high level frontal source (rIFG) is more strongly involved in the representation of stimulus change as compared to Bayesian surprise. This finding is in accordance with the "salience network"- theory, which implicates higher level frontal/insular cortex in the bottom-up detection of salient events and switching between other large-scale networks to facilitate access to attention and working memory resources (Menon and Uddin, 2010; Vinod, 2011). Finally, the model analyses indicate additional perceptual learning attributable to the cingulate cortex (MC) at a later processing step (around 360 ms). This result lends support to recent suggestions that perceptual learning manifests itself at several temporal stages of the EEG response (Hamamé et al., 2011; Song et al., 2005), where later stages are thought to predominantly reflect learning-induced modulations of attention (Gilbert et al., 2001). The reduced superiority of Bayesian surprise over stimulus change for this late processing step might thus be indicative of an interaction between perceptual learning and saliency detection. Taken together, the present study provides novel evidence for spatiotemporal/functional segregation in human somatosensory processing. More specifically, early-processing/low-level-cortical stages (S1, contralateral S2) may implement passive short-term Bayesian perceptual learning, downstream intermediate-processing/high-level-cortical stages (rIFG) index active stimulus engagement, and late-processing/high-level-cortical stages (MC) may reflect learning-induced updating of top-down attentional control mechanisms. Importantly, without computational modeling of single trial EEG data at the source level, we would not have found the evidence for these anatomic-temporal functional differences in the somatosensory system.

In contrast to standard Bayesian online-learning schemes (Bishop, 2007), the present study shows that the cortical learning signal is not likely to arise from a Bayesian learner that incorporates all previously observed stimuli of an experimental session with equal weight (model BS0). Rather, we find strong evidence for temporally-adaptive Bayesian learners (BS1–BS4) that give higher weight to temporally close, rather than distant, observations. Specifically, as implied by the generally low explanatory power of model BS0 and by the lower explanatory power of models BS1/BS2 compared to models BS3/BS4 (for a number of ECDs and time windows, see Fig. 6A), the time window of stimulus integration in the current experimental paradigm is probably shorter than 30 s and closer to the 5–10 s range. As shown in the fourth column of Table 3, models BS3 and BS4 exhibit almost complete suppression of stimuli more than 8.6 s and 6.5 s in the past, respectively. A comparison with the timing parameters of the experimental paradigm (inter-stimulus interval 0.65 s, average length of identical stimulation ~5 s, maximal length of identical stimulation ~10 s) suggests that the somatosensory system may use an optimized integration window for the average temporal statistics of the stimulation sequence.

A number of studies have previously addressed the encoding of surprise in the auditory and visual domain using model-based trial-by-trial analyses of evoked EEG and fMRI data (Harrison et al., 2006, 2011; Mars et al., 2008; Strange et al., 2005). Besides providing an extension to the somatosensory domain, the present study makes the following novel contributions: first, the only previous EEG study that used a trial-by-trial model-based analysis of surprise encoding

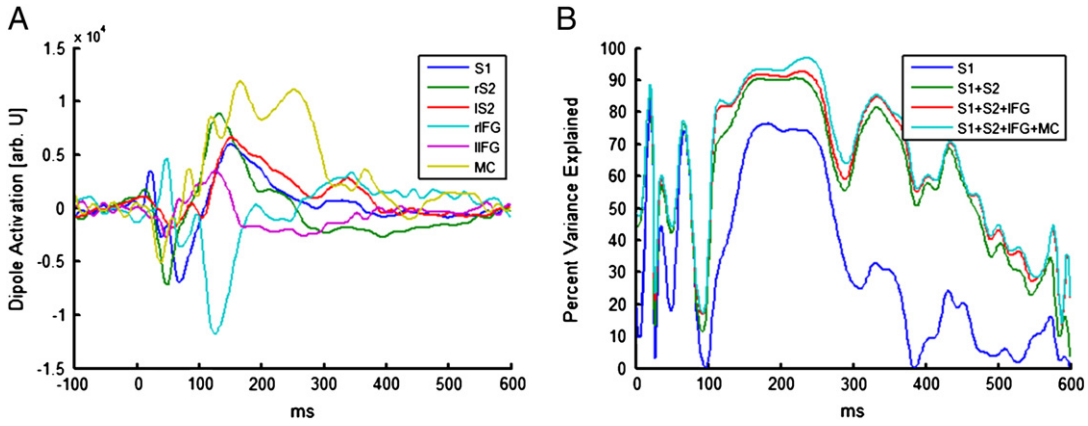


Fig. 5. Plausibility of the six source ECD model. A) ECD activity waveforms obtained by projecting the grand mean SEP onto the oriented ECDs. B) The channel percent variance explained (PVE) for the complete six ECD model (S1 + S2 + IFG + MC) and three hierarchical subversions of this model comprising only the S1 ECD (S1), the S1 and bilateral S2 ECDs (S1 + S2), and the S1, bilateral S2, and bilateral IFG ECDs (S1 + S2 + IFG + MC).

similar to the one employed here focused on a single spatiotemporal feature, namely the P300 amplitude at electrode Pz (Mars et al., 2008). The present study goes beyond this approach by explicitly analyzing all peri-stimulus time bin features of cortical source activity, thereby allowing for a comprehensive analysis of the spatiotemporal activity in brain space. Second, with respect to the model-based regressors, most previous studies used the negative log of the current stimulus probability estimates as measure of surprise (Harrison et al., 2006; Mars et al., 2008; Strange et al., 2005). While we are not claiming that Bayesian surprise is necessarily a better measure of surprise in perceptual learning schemes, it has the benefit of representing the degree to which the internal model is updated on a given trial, rather than the improbability of a stimulus under the current state of the model. This has the advantage that it reflects updating of the internal model over all possible states of external causes, rather than being conditioned on the point estimate of a single cause. In this sense, Bayesian surprise allows the observer to economically evaluate a broad range of possible external scenarios weighted by their inferred uncertainty, rather than relying on a possibly unreliable point estimate of the external cause for future predictions. Third, most previous studies (with the notable exception of Harrison et al., 2011) did not explicitly address the question of the temporal scale of stimulus integration, as provided by the exponential forgetting kinetic used here. We found clear evidence that somatosensory processing represents the stimulus sequence at a specific time-scale which may be related to the temporal statistics of the input

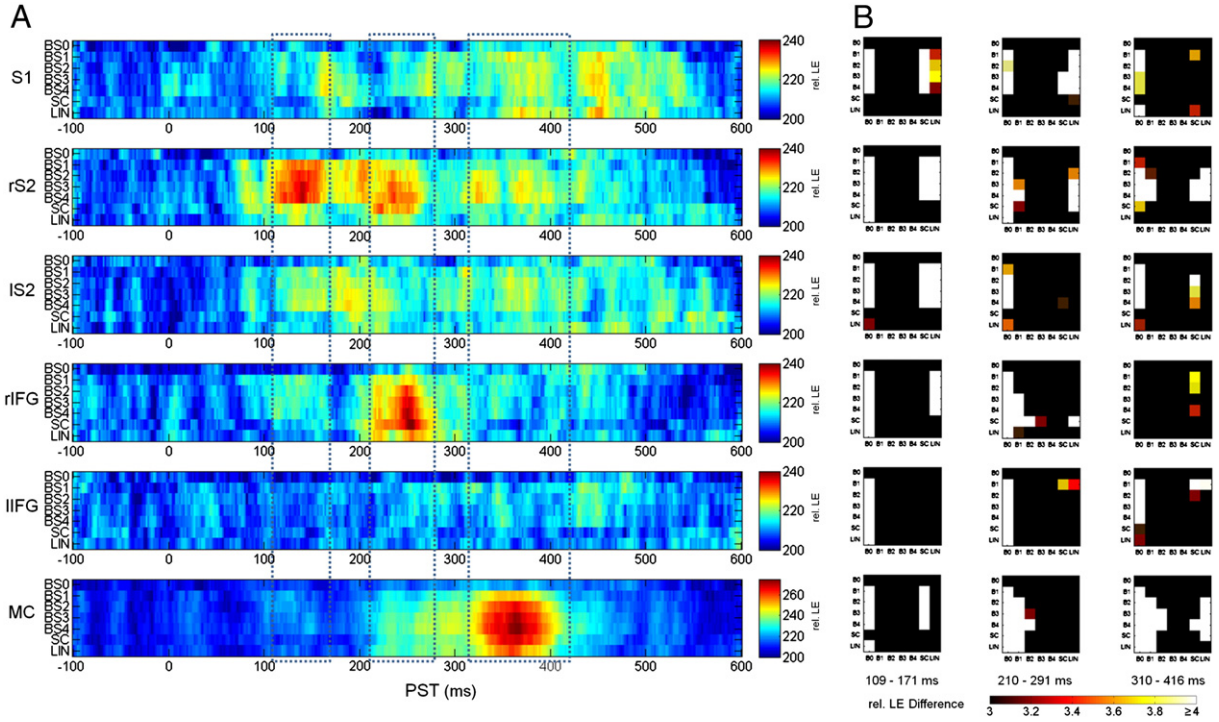


Fig. 6. Computational modeling results. A) Relative group model log evidences for all ECD sources and peri-stimulus time (PST) −100 to 600 ms (S1: right primary sensory cortex, rS2: right secondary somatosensory cortex, lS2: left secondary somatosensory cortex, rIFG: right inferior frontal gyrus, lIFG: left inferior frontal gyrus, MC: medial cingulate cortex). Each row within each ECD panel depicts the model log-evidence over peri-stimulus time for a specific model relative to the constant null model M0 (BS0–BS4: Bayesian surprise models with different time constants for the exponential forgetting constant, see Table 3. SC: stimulus change model, LIN: linear model). The dotted rectangles indicate time-windows of interest further evaluated in panel B. B) Pair-wise model log evidence comparisons for the three time-windows identified in A (109–171 ms, 210–291 ms, 310–416 ms). The squares of each panel color code the difference in model log evidence between the model indicated as row and the model indicated as column. For example, the color in the square at location (B1, SC) denotes the group model log evidence $\ln(y|B1) - \ln(y|SC)$.

sequences. Fourth, and probably most importantly, mismatch responses in EEG have traditionally been assumed to be elicited by stimulus change (Näätänen et al., 2011), whereas more recent accounts have explicitly addressed the theoretical notion that mismatch responses may be evidence of internal model adjustments in response to unexpected input (Garrido et al., 2009a; Winkler et al., 2009). The computational modeling approach employed here enables us to formally define and statistically test this notion using single trial EEG data.

We conclude the discussion with some considerations of the methodological approach employed: First, the current model-based trial-by-trial EEG analyses are conditioned on the selection of the anatomical location and orientation parameters of the ECDs used for data set projection. While this technique allows us to selectively monitor 'virtual' neural activity in specified brain areas, the data reduction entailed by this procedure might provide different results when using a different source model. Although we have taken great care to appropriately motivate the selection of sources based on both previous literature as well as on source reconstruction results of the present data set, we cannot rule out that there is some other source model that is more plausible than the one employed. Second, the underlying assumption of the present Bayesian surprise model is that the generative model used by the brain samples each stimulus identically and independently. This assumption enabled us to employ a simple update rule and is reasonably plausible given that we explicitly model sequences of stimuli. Third, fitting separate predictor functions derived from the Bayesian surprise model with different time constants τ of 'forgetting' allowed us to obtain a rough estimate of the optimal forgetting constant for a given ECD and time-window. This approach may be replaced by treating the time constant τ as a free parameter of a nonlinear optimization problem. In such a framework, it would be possible to evaluate a continuous parameter space, possibly in a Bayesian fashion, and to directly obtain source-specific estimates of the temporal dynamics of perceptual learning.

Conclusion

In summary, our current study indicates that the dynamics of single-trial somatosensory EEG responses can be explained by a formal model of Bayesian perceptual learning. Specifically, we have shown that in a somatosensory mismatch paradigm Bayesian surprise signals are encoded by multiple cortical regions of the somatosensory network in a temporally specific manner and found that Bayesian surprise signals can provide a better explanation for source-reconstructed single-trial EEG signals than conventional models typically employed for mismatch negativity studies.

Acknowledgments

This work was supported by the BMBF Bernstein II initiative (Förderkennzeichen: 01GQ1001C).

References

Akatsuka, K., Wasaka, T., Nakata, H., Kida, T., Hoshiyama, M., Tamura, Y., Kakigi, R., 2007a. Objective examination for two-point stimulation using a somatosensory oddball paradigm: an MEG study. Clin. Neurophysiol. 118, 403–411.
Akatsuka, K., Wasaka, T., Nakata, H., Kida, T., Kakigi, R., 2007b. The effect of stimulus probability on the somatosensory mismatch field. Exp. Brain Res. 181, 607–614.
Baldeweg, T., Klugman, A., Gruzelier, J., Hirsch, S.R., 2004. Mismatch negativity potentials and cognitive impairment in schizophrenia. Schizophr. Res. 69, 203–217.
Baldi, P., Itti, L., 2010. Of bits and wows: a Bayesian theory of surprise with applications to attention. Neural Netw. 23, 649–666.
Berg, P., Scherg, M., 1994. A multiple source approach to the correction of eye artifacts. Electroencephalogr. Clin. Neurophysiol. 90, 229–241.
Bishop CM., 2007. Pattern Recognition and Machine Learning. First ed. 2006. Corr. 2nd printing. ed. Springer.
Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory, Ninety-ninth ed. Wiley-Interscience.
Doya, K., Ishii, S., Pouget, A., 2007. Bayesian Brain: Probabilistic Approaches to Neural Coding, first ed. MIT Press.
Eickhoff, S.B., Stephan, K.E., Mohlberg, H., Grefkes, C., Fink, G.R., Amunts, K., Zilles, K., 2005. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage 25, 1325–1335.
Friston, K., 2003. Learning and inference in the brain. Neural Netw. 16, 1325–1352.
Friston, K.J., 2010. The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. 11, 127–138.
Friston, K.J., Dolan, R.J., 2010. Computational and dynamic models in neuroimaging. Neuroimage 52, 752–765.
Friston, K.J., Glaser, D.E., Henson, R.N.A., Kiebel, S., Phillips, C., Ashburner, J., 2002a. Classical and Bayesian inference in neuroimaging: applications. Neuroimage 16, 484–512.
Friston, K.J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002b. Classical and Bayesian inference in neuroimaging: theory. Neuroimage 16, 465–483.
Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2007. Variational free energy and the Laplace approximation. Neuroimage 34, 220–234.
Friston, K., Harrison, L., Daunizeau, J., Kiebel, S., Phillips, C., Trujillo-Barreto, N., Henson, R., Flandin, G., Mattout, J., 2008. Multiple sparse priors for the M/EEG inverse problem. Neuroimage 39, 1104–1120.
Fuchs, M., Wagner, M., Kohler, T., Wischmann, H.A., 1999. Linear and nonlinear current density reconstructions. J. Clin. Neurophysiol. 16 (3), 267–295.
Garrido, M.I., Kilner, J.M., Kiebel, S.J., Stephan, K.E., Friston, K.J., 2007. Dynamic causal modelling of evoked potentials: a reproducibility study. Neuroimage 36, 571–580.
Garrido, M.I., Kilner, J.M., Kiebel, S.J., Friston, K.J., 2009a. Dynamic causal modeling of the response to frequency deviants. J. Neurophysiol. 101, 2620–2631.
Garrido, M.I., Kilner, J.M., Kiebel, S.J., Stephan, K.E., Baldeweg, T., Friston, K.J., 2009b. Repetition suppression and plasticity in the human brain. Neuroimage 48, 269–279.
Garrido, M.I., Kilner, J.M., Stephan, K.E., Friston, K.J., 2009c. The mismatch negativity: a review of underlying mechanisms. Clin. Neurophysiol. 120, 453–463.
Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman and Hall, Boca Raton.
Gilbert, C.D., Sigman, M., Crist, R.E., 2001. The neural basis of perceptual learning. Neuron 31, 681–697.
Götz, T., Huonker, R., Miltner, W.H.R., Witte, O.W., Dettner, K., Weiss, T., 2011. Task requirements change signal strength of the primary somatosensory M50: oddball vs. one-back tasks. Psychophysiology 48, 569–577.
Hamamé, C.M., Cosmelli, D., Henriquez, R., Aboitiz, F., 2011. Neural mechanisms of human perceptual learning: electrophysiological evidence for a two-stage process.
Harrison, L.M., Duggins, A., Friston, K.J., 2006. Encoding uncertainty in the hippocampus. Neural Netw. 19, 535–546.
Harrison, L.M., Stephan, K.E., Rees, G., Friston, K.J., 2007. Extra-classical receptive field effects measured in striate cortex with fMRI. Neuroimage 34, 1199–1208.
Harrison, L.M., Bestmann, S., Rosa, M.J., Penny, W., Green, G.G.R., 2011. Time scales of representation in the human brain: weighing past information to predict future events. Front. Hum. Neurosci. 5, 37.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 14, 382–417.
Itti, L., Baldi, P., 2009. Bayesian surprise attracts human attention. Vision Res. 49, 1295–1306.
Kekoni, J., Hämäläinen, H., Saarinen, M., Gröhn, J., Reinikainen, K., Lehtokoski, A., Näätänen, R., 1997. Rate effect and mismatch responses in the somatosensory system: ERP-recordings in humans. Biol. Psychol. 46, 125–142.
Kersten, D., Mamassian, P., Yuille, A., 2004. Object perception as Bayesian inference. Annu. Rev. Psychol. 55, 271–304.
Kida, T., Nishihira, Y., Wasaka, T., Nakata, H., Sakamoto, M., 2004. Passive enhancement of the somatosensory P100 and N140 in an active attention task using deviant alone condition. Clin. Neurophysiol. 115, 871–879.
Kiebel, S.J., Friston, K.J., 2004. Statistical parametric mapping for event-related potentials: I. Generic considerations. Neuroimage 22, 492–502.
Kiebel, S.J., Daunizeau, J., Friston, K.J., 2008a. A hierarchy of time-scales and the brain. PLoS Comput. Biol. 4, e1000209.
Kiebel, S.J., Daunizeau, J., Phillips, C., Friston, K.J., 2008b. Variational Bayesian inversion of the equivalent current dipole model in EEG/MEG. Neuroimage 39, 728–741.
Knill, D.C., Pouget, A., 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci. 27, 712–719.
Linden, D.E.J., 2005. The P300: where in the brain is it produced and what does it tell us? Neuroscientist 11, 563–576.
Litvak, V., Friston, K., 2008. Electromagnetic source reconstruction for group studies. Neuroimage 42, 1490–1498.
Litvak, V., Komssi, S., Scherg, M., Hoechstetter, K., Classen, J., Zaaroor, M., Pratt, H., Kahkonen, S., 2007. Artifact correction and source analysis of early electroencephalographic responses evoked by transcranial magnetic stimulation over primary motor cortex. Neuroimage 37, 56–70.
Litvak, V., Mattout, J., Kiebel, S., Phillips, C., Henson, R., Kilner, J., Barnes, G., Oostenveld, R., Daunizeau, J., Flandin, G., Penny, W., Friston, K., 2011. EEG and MEG data analysis in SPM8. Comput. Intell. Neurosci. 2011, 852961.
Mars, R.B., Debener, S., Gladwin, T.E., Harrison, L.M., Haggard, P., Rothwell, J.C., Bestmann, S., 2008. Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. J. Neurosci. 28, 12539–12545.
Mars, R.B., Shea, N.J., Kolling, N., Rushworth, M.F.S., 2010. Model-based analyses: promises, pitfalls, and example applications to the study of cognitive control. Q J Exp Psychol (Hove) 1–16.

Menon, V., Uddin, L.Q., 2010. Saliency, switching, attention and control: a network model of insula function. Brain Struct. Funct. 214, 655–667.
Michel, C.M., Murray, M.M., Lantz, G., Gonzales, S., Spinelli, L., de Peralte, R.G., 2004. EEG source imaging. Clin. Neurophysiol. 115, 2195–2222.
Mouraux, A., Iannetti, G.D., 2008. A review of the evidence against the "first come first served" hypothesis Comment on Truini et al. [Pain 2007; 131:343–7] Pain 136, 219–225.
Mumford, D., 1992. On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol. Cybern. 66, 241–251.
Näätänen, R., 2009. Somatosensory mismatch negativity: a new clinical tool for developmental neurological research? Dev. Med. Child Neurol. 51, 930–931.
Näätänen, R., Gaillard, A.W., Mäntysalo, S., 1978. Early selective-attention effect on evoked potential reinterpreted. Acta Psychol. (Amst) 42, 313–329.
Näätänen, R., Kujala, T., Winkler, I., 2011. Auditory processing that leads to conscious perception: a unique window to central auditory processing opened by the mismatch negativity and related responses. Psychophysiology 48, 4–22.
Niedermeyer, E., Silva, F.L.D., 2004. Electroencephalography: Basic Principles, Clinical Applications, and Related Fields, Fifth ed. Lippincott Raven.
Penny, W.D., 2012. Comparing dynamic causal models using AIC, BIC and free energy. Neuroimage 59, 319–330.
Penny, W.D., 2001. Kullback-Leibler Divergences of Normal, Gamma, Dirichlet and Wishart Densities (Technical Report). Wellcome Department of Cognitive Neurology.
Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Comparing dynamic causal models. Neuroimage 22, 1157–1172.
Pitt, M.A., Myung, I.J., 2002. When a good fit can be bad. Trends Cogn. Sci. 6, 421–425.
Pleger, B., Foerster, A.F., Ragert, P., Dinse, H.R., Schwenkreis, P., Malin, J.P., Nicolas, V., Tegenthoff, M., 2003. Functional imaging of perceptual learning in human primary and secondary somatosensory cortex. Neuron 40, 643–653.
Polich, J., 2007. Updating P300: an integrative theory of P3a and P3b. Clin. Neurophysiol. 118, 2128–2148.
Rao, R.P., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2, 79–87.
Restuccia, D., Della Marca, G., Valeriani, M., Leggio, M.G., Molinari, M., 2007. Cerebellar damage impairs detection of somatosensory input changes. A somatosensory mismatch-negativity study. Brain 130, 276–287.
Restuccia, D., Zanini, S., Cazzagon, M., Del Piero, I., Martucci, L., Della Marca, G., 2009. Somatosensory mismatch negativity in healthy children. Dev. Med. Child Neurol. 51, 991–998.
Rinne, T., Alho, K., Ilmoniemi, R.J., Virtanen, J., Näätänen, R., 2000. Separate time behaviors of the temporal and frontal mismatch negativity sources. Neuroimage 12, 14–19.
Rinne, T., Degerman, A., Alho, K., 2005. Superior temporal and inferior frontal cortices are activated by infrequent sound duration decrements: an fMRI study. Neuroimage 26, 66–72.
Rinne, T., Särkkä, A., Degerman, A., Schröger, E., Alho, K., 2006. Two separate mechanisms underlie auditory change detection and involuntary control of attention. Brain Res. 1077, 135–143.
Romo, R., Hernández, A., Zainos, A., Salinas, E., 2003. Correlated neuronal discharges that increase coding efficiency during perceptual discrimination. Neuron 38, 649–657.
Shinozaki, N., Yabe, H., Sutoh, T., Hiruma, T., Kaneko, S., 1998. Somatosensory automatic responses to deviant stimuli. Brain Res. Cogn. Brain Res. 7, 165–171.
Song, Y., Ding, Y., Fan, S., Qu, Z., Xu, L., Lu, C., Peng, D., 2005. Neural substrates of visual perceptual learning of simple and complex stimuli. Clin. Neurophysiol. 116, 632–639.
Spackman, L.A., Boyd, S.G., Towell, A., 2007. Effects of stimulus frequency and duration on somatosensory discrimination responses. Exp. Brain Res. 177, 21–30.
Spackman, L.A., Towell, A., Boyd, S.G., 2010. Somatosensory discrimination: an intracranial event-related potential study of children with refractory epilepsy. Brain Res. 1310, 68–76.
Strange, B.A., Duggins, A., Penny, W., Dolan, R.J., Friston, K.J., 2005. Information theory, novelty and hippocampal responses: unpredicted or unpredictable? Neural Netw. 18, 225–230.
Tarkka, I.M., Micheloyannis, S., Stokić, D.S., 1996. Generators for human P300 elicited by somatosensory stimuli using multiple dipole source analysis. Neuroscience 75, 275–287.
Thees, S., Blankenburg, F., Taskin, B., Curio, G., Villringer, A., 2003. Dipole source localization and fMRI of simultaneously recorded data applied to somatosensory categorization. Neuroimage 18, 707–719.
Tse, C.-Y., Penney, T.B., 2008. On the functional role of temporal and frontal cortex activation in passive detection of auditory deviance. Neuroimage 41, 1462–1470.
Vinod, M., 2011. Large-scale brain networks and psychopathology: a unifying triple network model. Trends Cogn. Sci. 15, 483–506.
Wang, A.L., Mouraux, A., Liang, M., Iannetti, G.D., 2010. Stimulus novelty, and not neural refractoriness, explains the repetition suppression of laser-evoked potentials. J. Neurophysiol. 104, 2116–2124.
Winkler, I., Denham, S.L., Nelken, I., 2009. Modeling the auditory scene: predictive regularity representations and perceptual objects. Trends Cogn. Sci. 13, 532–540.
Woolrich MW., in press. Bayesian inference in FMRI. Neuroimage. DOI: 10.1016/j.neuroimage.2011.10.047.
Worsley, K.J., 1994. Local maxima and the expected Euler characteristic of excursion sets of χ^2, F and t fields. Adv. Appl. Probab. 26, 13–42.

6.5 Mesolimbic confidence signals guide perceptual learning in the absence of external feedback

Manuscript version as of 8 August 2015

Submitted

Mesolimbic confidence signals guide perceptual learning in the absence of external feedback

Matthias Guggenmos[1,2], Gregor Wilbertz[2], Martin N. Hebart[3†], Philipp Sterzer[1,2†]

[1] Bernstein Center for Computational Neuroscience, 10115 Berlin, Philippstraße 13, Haus 6, Germany
[2] Visual Perception Laboratory, Charité Universitätsmedizin, 10117 Berlin, Charitéplatz 1, Germany
[3] Department of Systems Neuroscience, Universitätsklinikum Hamburg-Eppendorf, 20251 Hamburg, Martinistraße 52, Haus W34, Germany

[†] equal contribution

Short title: Confidence-based learning in the mesolimbic system

Classification: BIOLOGICAL SCIENCES / Neuroscience

Keywords: perceptual learning, confidence, reinforcement learning, feedback, ventral striatum

Corresponding author:

Matthias Guggenmos
Bernstein Center for Computational Neuroscience, Philippstraße 13, Haus 6, 10115 Berlin
Phone: +49 (0) 30 450 517131
E-Mail: matthias.guggenmos@bccn-berlin.de

ABSTRACT

It is well established that learning can occur without external feedback. Normative reinforcement learning theories, according to which we learn from the consequences of our actions, have difficulties to explain such instances of learning. Here we tested the hypothesis that learning may be guided by self-generated confidence signals that serve as internal feedback. Human participants performed a challenging visual perceptual learning task and reported their confidence after each trial. Brain activity during learning was measured with functional magnetic resonance imaging (fMRI) and analyzed with a novel computational model in which perceptual learning was guided by the combination of a confidence-based reinforcement signal and Hebbian plasticity. Model-based fMRI data analysis showed that activation in mesolimbic brain areas reflected pre-stimulus anticipation of confidence and a subsequent stimulus-related confidence prediction error, revealing a striking similarity in the neural coding of internal confidence-based and external reward-based feedback. Importantly, mesolimbic confidence prediction error modulation predicted individual learning success, establishing the behavioral relevance of these self-generated feedback signals. Together, our results provide evidence for an important role of confidence-based mesolimbic feedback signals in perceptual learning and extend reinforcement-based models of learning to cases where no external feedback is available.

SIGNIFICANCE STATEMENT

A fundamental question of human behavior and learning is how we learn without external feedback. Here we pursued the idea that such learning may be guided by self-generated confidence signals that serve as internal feedback. Using psychophysics, functional magnetic resonance imaging and computational modeling we tracked confidence signals during a challenging perceptual learning task. We found that the neural substrate and the encoding pattern of confidence signals were in striking analogy to findings from classical studies with reward-based feedback. Importantly, the strength of these confidence signals predicted individual learning success. Together, our results provide evidence that reward-based and confidence-based feedback signals rely on a common neural mechanism, and indicate a general mechanism for learning in the absence of external feedback.

INTRODUCTION

Learning is an integral part of our everyday life and necessary for survival in a dynamic environment. The behavioral changes arising from learning have quite successfully been described by the reinforcement learning principle (1), according to which biological agents continuously adapt their behavior based on the consequences of their actions. Thus, reinforcement learning models and most other learning models depend on feedback from the environment, such as reward. Yet, there are important instances of learning where no such external feedback is provided, challenging the generality of these learning models in shaping our behavior.

A well-studied case of learning is the improvement of performance in perceptually demanding tasks through training or repeated exposure (2). Such *perceptual learning* has repeatedly been demonstrated to occur without feedback (3–6) and is therefore ideally suited as a test case to study learning in the absence of external feedback. Previous work has emphasized the role of reinforcement learning in perceptual learning (7, 8). However, these accounts were based on perceptual learning *with* external feedback and therefore cannot account for instances in which learning occurs *without* external feedback. Here we pursued the idea that, in the absence of external feedback, learning is guided by internal feedback processes that evaluate current perceptual information in relation to prior knowledge about the sensory world. We reasoned that introspective reports of perceptual confidence could serve as a window into such internal feedback processes. In this scenario, low or high confidence would correspond to a negative or positive self-evaluation of one's own perceptual performance, respectively. Accordingly, confidence could act as a teaching signal in the same

way as external feedback (reward) in normative theories of reinforcement learning (9, 10). Applied to the case of perceptual learning, a confidence-based reinforcement signal could serve to strengthen neural circuitry that gave rise to high-confidence percepts and weaken circuitry that led to low-confidence percepts, thereby enhancing the quality of future percepts.

We tested this idea in a challenging perceptual learning task, in which participants continuously reported their confidence in perceptual choices while undergoing functional magnetic resonance imaging (fMRI). No external feedback was provided; instead, confidence ratings were used as a proxy of internal feedback processes. To account for perceptual learning in the absence of feedback, we devised a confidence-based associative reinforcement learning model. In the model, confidence prediction errors (9) serve as teaching signals that indicate the mismatch between the current level of confidence and a running average of previous confidence experiences (expected confidence). Based on recent evidence of confidence signals in the mesolimbic dopamine system (9–11), we hypothesized to find neural correlates of confidence prediction errors in mesolimbic brain areas such as the ventral striatum and the ventral tegmental area. Since confidence prediction errors act as a teaching signal in our model, we hypothesized that the strength of these mesolimbic confidence signals should be linked to individual perceptual learning success.

RESULTS

Human participants (N=29) learned to detect the orientation of peripheral noise-embedded Gabor patches relative to a horizontal or vertical reference axis while undergoing functional magnetic resonance imaging (fMRI). Overall, the experiment comprised four sessions: (i) an initial behavioral test session to establish participants' baseline contrast thresholds for a performance level of 80.35% correct responses, (ii) an intensive perceptual learning session (training) in the MRI scanner with a continuous threshold determination, and two behavioral post-training test sessions to examine (iii) short-term and (iv) long-term stimulus-specific training effects (Figure 1A). While the training session was based on one reference axis (randomly assigned to each participant), all test sessions comprised a contrast threshold determination for both reference axes. The training session additionally included a control condition in interleaved presentation, for which the contrast was kept constant to enable an exploratory multivariate analysis of changes in neural stimulus representation. The Gabor stimuli were flashed briefly in the upper right quadrant and participants had to judge their orientation with respect to the current reference axis (Figures 1B and 1C). Eyetracking ensured that participants maintained fixation throughout the training session (see Figure S1). Importantly, participants did not receive external cognitive or rewarding feedback during the entire experiment. Rather, in addition to their choice, they reported their confidence about the stimulus orientation on a visual analogue scale (for a verification of accurate usage, see Figure S2). The confidence reports were used to compute the internal feedback in our model on a trial-by-trial basis.

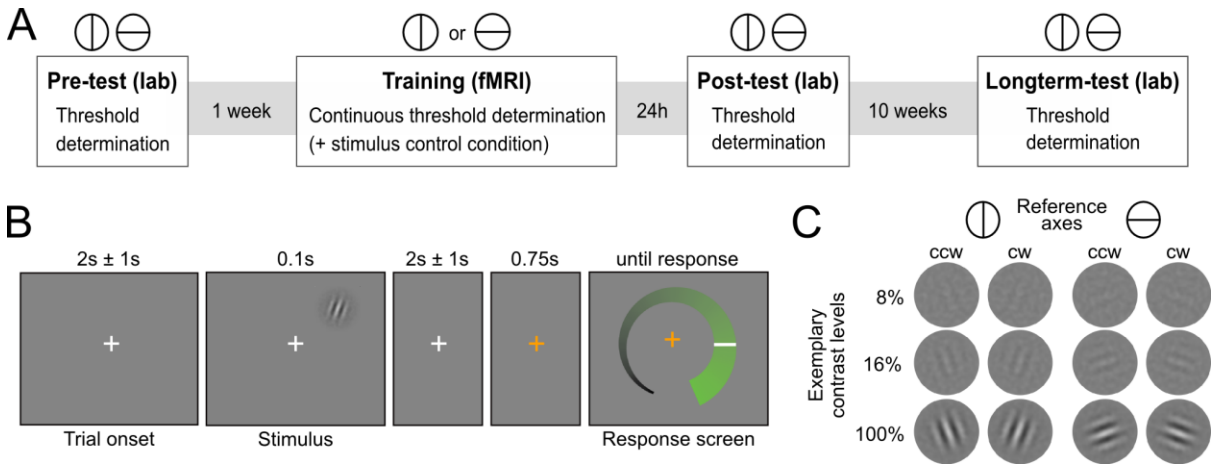


Figure 1. Experimental design. (A) Overview over experimental sessions. The experiment consisted of one training session and three test sessions (pre-test, post-test and longterm-test). The test sessions included both reference axes and were used to determine the contrast threshold for a performance of 80.35 percent correct at different stages of the experiment. In the training session only one reference axis was shown. Here too, a staircase procedure was used to continuously determine the contrast threshold for a performance level of 80.35%. In addition, the training session included a condition with constant contrast as a control for stimulus factors. Overall, this design enabled us to assess the specificity of perceptual learning one day (post-test) and 10 weeks (longterm-test) after training relative to the baseline of the pre-test. (B) Procedure of an experimental trial. Participants were presented with Gabor stimuli, which were oriented either clockwise or counterclockwise with respect to a reference axes. In the unspeeded response phase participants indicated their level of confidence about the stimulus orientation on a visual analogue scale and subsequently made a binary orientation judgment. (C) Examples of the stimuli. Gabor patches were oriented 20° clockwise (cw) or 20° counterclockwise (ccw) relative to either the vertical or the horizontal reference axis. Three exemplary contrast levels are shown, where 8% corresponds to the participant average during training, 16% to the highest obtained thresholds and 100% to full contrast.

Stimulus-specific perceptual learning. To establish stimulus-specific perceptual learning, we compared perceptual thresholds in pre- and post-experimental sessions between the trained and untrained reference axis (Figure 2A). The contrast thresholds improved for the trained (t_{28} = 6.73, $p < 0.001$, two-tailed), but not for the untrained reference axis (t_{28} = 0.41, $p = 0.68$; interaction of training × time: $F_{1,28}$ = 14.2, $p < 0.001$), demonstrating clear and specific effects of perceptual learning. These stimulus-specific training effects could still be detected 10 weeks later ($F_{1,28}$ = 4.3, $p = 0.047$), indicating long-term stability and thus demonstrating a key characteristic of perceptual learning (12). To test whether the effects of learning could already be detected during the training session, we linearly fitted the contrast thresholds across trials

in the critical constant performance condition. The analysis showed that contrast threshold consistently decreased across runs (linear slope: −0.006 ± 0.002, $t_{28} = -2.38$, $p = 0.024$), from 8.64% ± 0.47 (mean ± SEM) in the first training run to 7.68% ± 0.52 in the last training run.

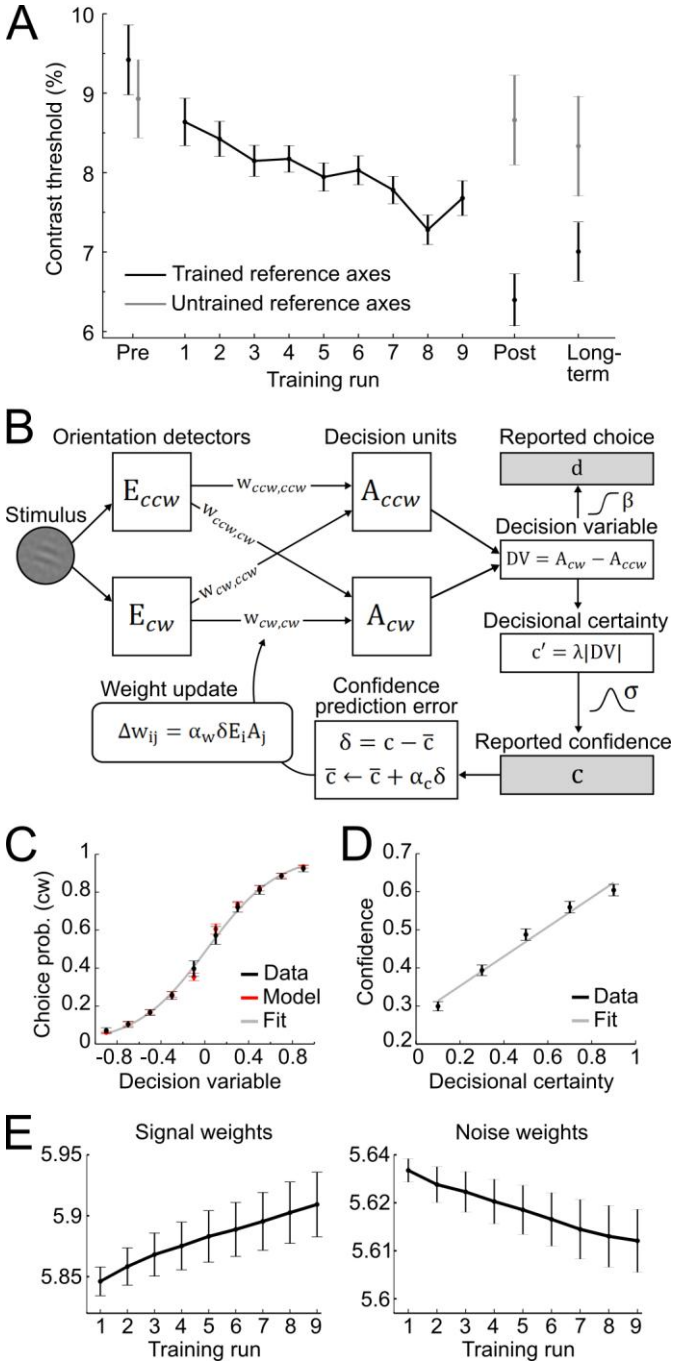


Figure 2. Behavior and confidence-based model of perceptual learning. (A) Contrast thresholds across the runs of the training session and in the three test-sessions (pre/post/long-term). (B) Model. Counterclockwise (E_{cw}) and clockwise (E_{ccw}) orientation energy detectors of a dedicated representational subsystem are connected via *signal weights* (horizontal) and *noise weights* (diagonal) to decision units (A_{ccw}, A_{cw}). Reported choices (decisions) *d* are probabilistically modeled by a *decision value* $DV = A_{ccw} - A_{cw}$ and the reported confidence *c* is modeled through the absolute value of *x*. Weights are updated through an associative reinforcement learning update rule. The reinforcement component is based on a *confidence prediction error δ*, reflecting the difference between reported confidence and a weighted running average of previous confidence experiences (*expected confidence* $\bar{c}$). The Hebbian component ($E_i \times A_j$) ensures that the update more strongly affects those connections that contribute more to the final choice. Grey-shaded boxes indicate observed variables. An exemplary time course of model variables and behavior is shown in Figure S4. (C) Binned choice probabilities (clockwise) for observed data (black) and model predictions (red) as a function of the model-derived DV (gray: logistic fit to data). (D) Correspondence between participants' binned confidence ratings and model-based decisional certainty (gray: linear fit). (E) Change of signal and noise weights across training runs. All error bars denote SEM corrected for between-subject variance (35).

A confidence-based model of perceptual learning. To account for perceptual learning without external feedback, we devised an associative reinforcement learning model with confidence as internal feedback (Figure 2B). Learning in this model was guided by the combination of a confidence-based reinforcement signal and Hebbian plasticity, inspired by the previously proposed *three-factor learning rule* (dopaminergic reinforcement signal, pre-synaptic activity, post-synaptic activity) for neural plasticity in the mesolimbic system (13, 14). Our model assumes that observers seek to maximize perceptual confidence by optimizing a filter on incoming sensory evidence. The filter is represented by two components: *signal weights* for clockwise (cw) and counterclockwise (ccw) stimulus orientation ($w_{ccw,ccw}$; $w_{cw,cw}$), connecting orientation energy detectors $E_{ccw/cw}$ to decision units $A_{ccw/cw}$ with same orientations; and *noise weights* ($w_{ccw,cw}$; $w_{cw,ccw}$), connecting detectors $E_{ccw/cw}$ to decision units $A_{cw/ccw}$ with opposing orientations. The clockwise and counterclockwise orientation energy contained in the stimuli is computed by a simple model of primary visual cortex (15; see Figure S3 for a validation of this representational subsystem). The weighted sums of $E_{ccw/cw}$ determine the activities of decision units $A_{cw/ccw}$ which, in a next step, are integrated in a *decision value* $DV = A_{cw} - A_{ccw}$. DV translates into probabilities for clockwise or counterclockwise choices via a softmax action selection rule, and to the model's equivalent of confidence—*decisional certainty*—through its absolute value. Finally, perceptual learning is based on an associative reinforcement learning rule with two separate components: a reinforcement component utilizes a confidence prediction error (CPE; denoted as δ) as internal feedback, representing the mismatch between current confidence and a long-term estimate (via a learning rate α_c) of expected confidence; and a Hebbian component ensures that the weights are updated in proportion to how strongly orientation detectors and decision units co-activate. In addition, a learning rate α_w accounts

for inter-individual differences in learning speed.

The model parameters were fitted to participants' orientation and confidence reports in the training session using maximum likelihood approximation (median ± SE of the median: α_w = 0.0018 ± 0.0007, α_c = 0.533 ± 0.077; see Table S1 for other parameters and supporting section "Relationship between learning rates and the learning process" for a brief discussion of α_w/α_c). To test the validity of the model, we correlated model-based choice probabilities with participants' actual choices. This analysis showed that the model accounted well for participants' choices (mean ± SE of $r_{pearson}$ = 0.64 ± 0.03, t_{28} = 26.2, $p < 0.001$, one-sample t-test against Fisher z' = 0). This correspondence is reflected in the fact that participants' and model-based choice probabilities show a nearly identical (sigmoidal) dependency of *DV* (Figure 2C). We next assessed whether the model could predict participants' trial-wise confidence reports. A correlation between the model-based decisional certainty and participants' confidence reports confirmed that confidence, too, was captured the by the model (mean ± SE of $r_{pearson}$ = 0.32 ± 0.02, t_{28} = 15.4, $p < 0.001$; Figure 2D). Control analyses confirmed that this model explained the data better than models with alternative feedback signals (supporting section "Alternative feedback signals" and Figure S7).

To evaluate how perceptual learning was reflected in the update of the model's sensory filter, we assessed the change of signal and noise weights across runs. We expected that an increase of signal weights and a decrease of noise weights over the course of the training session was responsible for perceptual improvements. As depicted in Figure 2E, we found a linear increase for signal weights across runs (mean ± SEM of slope = 0.0147 ± 0.0036, t_{28} = 4.1, $p < 0.001$), and a linear decrease for noise weights (slope = −0.0036 ± 0.0010, t_{28} = −3.5, $p = 0.001$). Furthermore, the individual contrast threshold learning slopes

correlated negatively with the slopes of signal weights ($r_{pearson} = -0.45$, $p = 0.013$) and positively with the slopes of noise weights ($r_{pearson} = 0.46$, $p = 0.011$). Thus, individual learning was well captured by the signal and noise weights of the model.

Model-free analysis of brain activation: analogy of confidence-based and reward-based feedback signals in the ventral striatum. We reasoned that if confidence-based internal feedback and reward-based external feedback share a common neural basis, neural responses in the ventral striatum to high-, average- and low-confidence events should exhibit a qualitatively similar pattern as reported for rewarding, neutral and punishing outcomes in reward-based learning (16, 17). The results of these previous studies suggest that striatal activation reflects a positive anticipatory response at the beginning of a trial as well as a subsequent prediction error response related to the outcome. To simulate the BOLD response that arises from such a scheme, we convolved vectors coding the neural activation for an initial anticipatory response and three different outcome scenarios (positive, absent and negative prediction error) with a canonical double-gamma hemodynamic response function (Figure 3A). In accordance with the fMRI results of these previous studies, the simulation shows (i) an increase of striatal BOLD responses related to trial onset reflecting the anticipatory signal, and (ii) a subsequent positive, absent, or negative deflection of the BOLD response reflecting prediction errors.

To relate the neural signature of confidence in the present study to the simulation and to the results of previous reward-based studies, we binned the data into tertiles of the behavioral confidence rating (low, middle and high confidence) and extracted the average BOLD time course in an anatomical mask of the ventral striatum using the SPM toolbox rfxplot (18). As shown in Figure 3B, the obtained event-related BOLD time courses are in remarkable

agreement with the predictions of the simulation of Figure 3A and previous empirical findings of reward studies (e.g, see 16). Specifically, 4-6 seconds after *trial onset* (reflecting the hemodynamic delay), the BOLD time courses exhibited a first peak, consistent with an anticipatory confidence signal at the start of a trial; 4-6 seconds after *stimulus onset*, the BOLD time courses displayed a positive deflection for high-confidence trials and a negative deflection for low-confidence trials. A statistical analysis confirmed above-baseline striatal activation at trial onset, indicative of an anticipatory signal (left peak at [−10 14 −6], $t_{28} = 6.42$, $p_{rFWE} < 0.001$; right peak at [12 14 −8], $t_{28} = 7.78$, $p_{rFWE} < 0.001$), as well as a main effect of confidence at stimulus onset in the bilateral ventral striatum (left peak at [−10 14 −4], $t_{28} = 10.56$, $p_{rFWE} < 0.001$; right peak at [16 12 −8], $t_{28} = 11.46$, $p_{rFWE} < 0.001$) (see Table S2 for a whole-brain list). Note that the pattern of results is unchanged when matching the absolute orientation energy of the three confidence tertiles (Figure S6A). This model-free assessment provides initial support for the idea that reinforcement based on reward and based on confidence share a common neural substrate both in the anticipation and the outcome period. In addition, the results lend plausibility to a model utilizing confidence as a reinforcement signal.

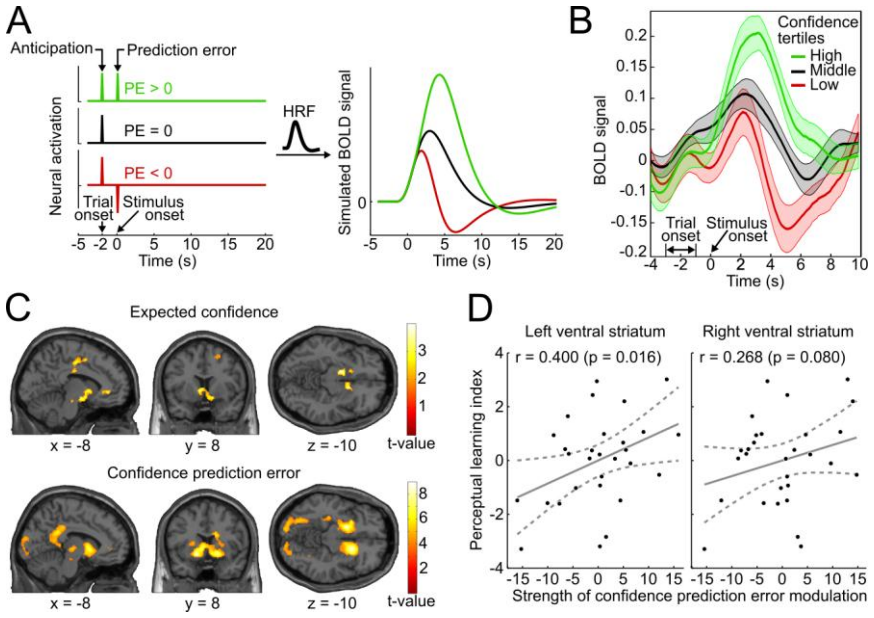


Figure 3. Confidence signals in the mesolimbic system and their relation to perceptual learning. (A) Neural activation time courses consisting of an anticipatory peak at trial onset and a positive, absent, or negative reward prediction error (PE) during outcome (stimulus onset). To simulate the associated BOLD response, the time courses were convolved with the standard canonical hemodynamic response function provided by SPM. (B) Event-related BOLD time courses in the ventral striatum for three tertiles of the behavioral confidence reports (representing "low", "middle" and "high" confidence trials). The shaded areas denote SEM. (C) Whole-brain T-maps showing brain regions with a positive relationship between BOLD signal and expected confidence at trial onset (top panel), and between BOLD signal and CPE at stimulus onset (bottom panel). The T-maps were thresholded at $p < 0.005$ (top panel) and $p < 0.001$ (bottom panel), uncorrected, for illustration purposes. (D) Scatter plot for the relation between the strength of striatal modulation by confidence prediction errors and individual perceptual learning success.

Model-based analysis of brain activation: mesolimbic correlates of expected confidence and confidence prediction errors. To link the confidence-based reinforcement learning model to brain activity, we estimated a new general linear model (GLM) using two time-varying parametric variables generated from the model: expected confidence at trial onset and CPE at stimulus onset. As conjectured, we found a significant parametric modulation of striatal activation by expected confidence at trial onset (right peak at [8 14 −4], $t_{28} = 4.12$, $p_{rFWE} = 0.018$; left peak at [−12 20 −2], $t_{28} = 3.97$, $p_{rFWE} = 0.026$; see also Table S3 and Figure S5). This model-based result corroborates the notion that striatal activation at trial onset reflects anticipated confidence for the upcoming stimulus presentation, analogous to previous

findings of an anticipatory reward signal in the ventral striatum (16, 17, 19).

Using the parametric CPE regressor we next tested for a positive linear relationship between BOLD and the CPE at the time of stimulus presentation (Figure 3C). As hypothesized, the bilateral ventral striatum showed a strong positive relationship with the CPE (left peak at [−16 8 −10], $t_{28} = 7.64$, $p_{rFWE} < 0.001$; right peak at [16 14 −8], $t_{28} = 7.81$, $p_{rFWE} < 0.001$). This modulation was also present in our second region of interest, the ventral tegmental area (peak at [−6 −22 −16], $t_{28} = 3.02$, $p_{rFWE}=0.027$). Control analyses showed that the modulation of mesolimbic activity was not merely a reflection of varying stimulus energy (Figure S6B). In sum, the CPE was represented by the same key brain structures that have been implicated in signaling reward prediction errors (20–22).

The striatal confidence prediction error signal correlates with individual perceptual learning success. Finally, we investigated whether the strength of the striatal confidence prediction error modulation translates into improvements in perceptual performance. For that purpose, we correlated the parameter estimates at the peaks of the bilateral striatal CPE contrast (coordinates [−16 8 −10] and [16 14 −8], see above) with an index for the perceptual learning success that quantified the threshold change for the trained reference axis while accounting for baseline thresholds. Age was included as an additional factor in the regression model to preclude confounding effects of age-related variation in striatal reinforcement signals (23). As hypothesized, we found a significant relationship between the striatal CPE signal and individual perceptual learning success in the left ventral striatum ($r_{pearson} = 0.400$, $p = 0.016$, one-tailed) and a trend in the right ventral striatum ($r_{pearson} = 0.268$, $p = 0.080$) (see Figure 3D). This result indicates that humans utilize this CPE signal for perceptual learning in the absence of external feedback.

Identifying the neural basis of sensory and decisional model variables. To assess whether the model variables associated with the sensory and decisional subsystem would be reflected in brain activity, we performed a multivariate whole-brain searchlight (24) analysis using cross-validated MANOVA (cvMANOVA, 25). As a variable describing the sensory we used *orientation energy* (OE), defined as the sign of the stimulus orientation (ccw = −1, cw = 1) in conjunction with the corresponding orientation energy tertile (low = 0.5, middle = 1.5, high = 2.5). CvMANOVA was then used to search for brain areas showing a multivariate linear relationship with OE. We found that OE was encoded in left occipital cortex (peak at [−44 70 0], $t_{28} = 4.77$, $p_{cFWE} = 0.033$), an area overlapping with voxels activated by the stimulus localizer (Figure 4A). An additional ROI analysis to track the distinctness of stimulus-coding patterns across training runs within the control condition (constant contrast) was not feasible due to a lack of statistical power (see SI section “Multivariate ROI analysis”).

For the analysis of the decision value (DV), trials were sorted in an analogous manner, such that negative and positive DVs (reflecting the model’s tendency for counterclockwise and clockwise choices) were separately sorted into tertiles depending on the respective absolute value. Interestingly, decoding of DV showed only a trend in occipital cortex (peak at [−40 −72 26], $t_{28} = 3.35$, $p = 0.001$, uncorrected). Instead, we found significant encoding of the DV in right middle frontal gyrus (peak at [32 14 44], $t_{28} = 5.87$, $p_{FWE} = 0.026$; Figure 4B), in line with a recent report (10). Thus, perceptual and decision variables of our model can be mapped to visual and frontal cortex, respectively.

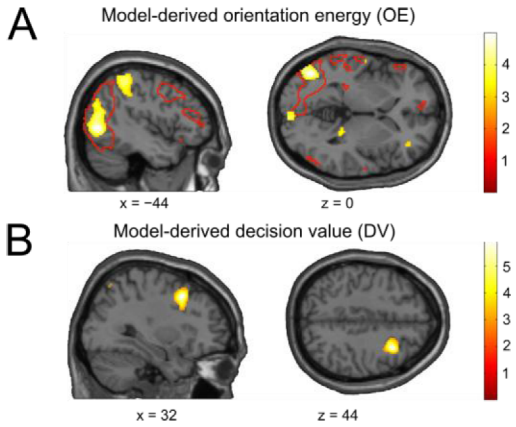


Figure 4. The neural basis of perceptual and decisional model variables. (A) Model-derived signed orientation energy (OE). The panel shows the t-map for multivariate decoding of OE. Red outlines indicate areas generally responding to the stimulus as measured with the independent stimulus localizer (t-contrast: stimulus > baseline). (B) Model-derived decision value (DV). T-map for multivariate decoding of the model-derived DV. All t-maps are thresholded at $p < 0.005$, for illustration.

DISCUSSION

In this study we addressed the question of how the human brain improves perception in the absence of external feedback. Previous reinforcement learning accounts of perceptual learning were based on external cognitive and rewarding feedback (7, 8) and could not explain the established phenomenon of perceptual learning without such feedback (3–6). Here we suggest that observers are capable of generating internal feedback by utilizing *confidence signals* that provide a graded evaluation of the correctness of a perceptual decision. In this way confidence may serve as a reinforcement signal similar to reward and guide perceptual learning in cases where no external feedback is provided.

In support of this view, our model-free fMRI analyses revealed that mesolimbic confidence signals mirror those typically found for reward, both in the anticipation period (16, 17, 19) and for prediction errors (20–22). To establish a mechanistic ground for this suggested parallel, we devised an associate reinforcement learning model, which links behavior to computational variables that each account for a different aspect of the learning process. CPEs served as feedback in the model, defined as the difference between the current level of

confidence and a long-term estimate of expected confidence. The model successfully described the learning process as a continuous adjustment of a perceptual filter linking sensory and decision units. Our model-based fMRI analyses confirmed and extended the results of the model-free analyses by demonstrating a parametric modulation in mesolimbic brain areas both by expected confidence and confidence prediction. Overall, the observed pattern of results as well as the co-modulation of the ventral tegmental area by CPEs fit well with the prediction error hypothesis of dopamine, which posits that dopaminergic midbrain neurons and their targets respond at two time points during a learning trial (20): (i) an anticipatory signal triggered by an outcome-predicting cue, and (ii) a surprise signal (prediction error) triggered by the actual outcome. Importantly, the strength of the striatal modulation by CPEs predicted participants' perceptual improvements, further corroborating the behavioral relevance of these internally-generated feedback signals.

A number of previous studies have used reinforcement learning models to capture the neural underpinnings of perceptual learning (7, 8) and category learning (9). In particular, an fMRI study by Kahnt and colleagues (7) investigated perceptual learning with external reward and found that behavioral improvements were well explained by a reinforcement learning model. Their results exhibit a notable parallel to the present findings: the authors reported stimulus information encoded in visual cortex and model-derived decision value in frontal cortices, in agreement with the findings of the present study. In addition, this previous study identified a perceptual learning-related reward prediction error in the ventral striatum, dovetailing with our finding of a perceptual learning-related confidence prediction error in the same brain region. Importantly, our combined Hebbian and reinforcement learning model extends and improves previous models in several ways. First and foremost, by implementing

confidence prediction errors in replacement of reward prediction errors, it extends previous reward reinforcement learning models of perceptual learning (7, 8) to cases without feedback. Second, these previous models were based on the assumption that perceptual performance is determined by a single "readout weight", representing the amplification of stimulus information in sensory areas. While the simplicity of these models is appealing, they are limited in the sense that negative prediction errors have an unreasonable influence on behavior: according to these models, worse-than-expected feedback reduces the readout weight, which leads to an additional reduction in performance. This property runs counter to the idea that reinforcement learning agents improve their behavior through both positive and negative prediction errors. By contrast, the *associative reinforcement learning rule* of the present model entails a behaviorally advantageous and plausible function of negative prediction errors: inhibition of sensory noise. Third, a conceptually related reinforcement learning model for perceptual categorization (9) implies that stimuli exclusively activate the correct stimulus category, an assumption that disregards the fact that the ambiguity of incoming stimulus information is an essential property of perceptually demanding tasks. In contrast, the present model utilizes a dedicated *representational subsystem* (15) to estimate the activation of all implemented input units, and it is their differential activity that determines perceptual choices.

The present model and results are biologically plausible and fit well with theoretical accounts of the neural basis of learning. The associative reinforcement learning rule in the model was inspired by the three-factor learning rule (13, 14), which has been proposed to underlie the potentiation of synapses in the striatum. It proposes that changes in neural transmission in cortico-striatal synapses not only depend on coincident presynaptic and

postsynaptic activity (Hebbian learning), but also on the presence of dopamine error signals. Indeed, Ashby and colleagues (26, 27) have previously suggested that the basal ganglia, which represent the predominant site of dopaminergic synaptic plasticity, are themselves a key region for learning. They proposed that (i) the basal ganglia serve to activate the appropriate target regions in executive frontal cortices shortly after sensory cortex activation; and (ii) such basal ganglia learning is superseded by cortico-cortical Hebbian learning, once the correct cortico-cortical synapses are built. This account fits well with the present model, in which perceptual learning corresponds to the process of reweighting connections between sensory and decisional units. These considerations in combination with the present results thus lend support to the hypothesis that the optimization of perceptual read-out (as implicated by our model) could be mediated via reinforcement learning in the basal ganglia.

It is worth noting that a few additional brain regions were positively related to CPEs (Table S3). The strongest modulation was found in posterior cingulate cortex, a region that has been related to different aspects of reinforcement learning such as value, outcome monitoring and action evaluation (28). Significant modulation was also found in the mediodorsal thalamus which is thought to bridge basic reward signals with higher cognitive functions via the striatal–thalamo–cortical loop (29). Our finding of a CPE-related activation of the mediodorsal thalamus may reflect a similar role in the case of confidence-based reinforcement signals. Finally, we found that activation in infracalcarine occipital cortex was modulated by CPEs. The fact that the activity of this early visual area was best explained by CPEs and not stimulus energy per se (see analyses associated with Figure S6B) may indicate a feedback-driven modulation, in line with previous studies reporting a modulation of V1 by surprise or prediction error (30, 31). Whether this putative feedback modulation is critical for

perceptual learning (e.g., reflecting increased cortico-cortical Hebbian learning during high-confidence events) should be addressed in future studies.

It should be noted that the present results do not constitute causal evidence for a role of CPEs in perceptual learning. It is possible that CPEs reflect a non-reinforcing monitoring signal, which, for hitherto unknown reasons, would be enhanced in individuals also showing better perceptual learning. Mesolimbic dopamine neurons have also been found to signal *motivational salience*, an unsigned variable that is high for both rewarding and aversive events and which can lead to attentional orienting and increased motivational drive (32). If the parallel between reward for external feedback and confidence for internal feedback holds for the aversive domain, this would suggest the existence of an internal punishment signal reflecting the degree of certainty of having made a mistake, a concept closely related to error awareness (33). This conceptual distinction offers a prospect for future studies to disentangle internal monitoring signals and motivational salience. Irrespectively, the strength of our model-based approach is that it provides a mechanistic, rather than a mere correlative link between confidence-based feedback signals and perceptual learning. To corroborate the relationship between confidence and perceptual learning, future studies could investigate whether biases in metacognitive awareness influence mesolimbic confidence signals and, as a consequence, impact perceptual learning.

In summary, our study devised and tested a novel model of perceptual learning in the absence of external feedback, utilizing confidence prediction errors to guide the learning process. Our analyses revealed a compelling analogy between confidence-based and reward-based feedback, suggesting a similar neural mechanism for learning with and without external feedback. Future work could investigate whether a learning mechanism based such on such

self-generated feedback is also applicable outside the realm of perception, where learning without feedback has likewise been a long-standing puzzle (34).

MATERIALS AND METHODS

Observers (N=29) performed a challenging orientation discrimination task without feedback, while undergoing fMRI scanning. A computational model was used to explain behavior (choices, confidence reports) and to perform a model-based analysis of brain activity. Detailed methods for experimental procedures, the perceptual learning model and fMRI data analyses are provided in the SI Materials and Methods.

ACKNOWLEDGEMENTS

This study was supported by the DFG Research Training Group GRK 1589/1 and by the DFG grants STE 1430/6-1 and STE 1430/7-1. We thank Shea Karst for her assistance during the experiments.

REFERENCES

1. Sutton RS, Barto AG (1998) *Reinforcement Learning: An Introduction* (Bradford Books, MIT Press, Cambridge, MA).

2. Gibson EJ (1963) Perceptual Learning. *Annu Rev Psychol* 14:29–56.

3. Herzog MH, Fahle M (1997) The role of feedback in learning a vernier discrimination task. *Vision Res* 37:2133–41.

4. Gibson JJ, Gibson EJ (1955) Perceptual learning; differentiation or enrichment? *Psychol Rev* 62:32–41.

5. McKee SP, Westheimer G (1978) Improvement in vernier acuity with practice. *Percept Psychophys* 24:258–262.

6. Karni A, Sagi D (1991) Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity. *Proc Natl Acad Sci U S A* 88:4966–70.

7. Kahnt T, Grueschow M, Speck O, Haynes J (2011) Perceptual learning and decision-making in human medial frontal cortex. *Neuron* 70:549–59.

8. Law C-T, Gold JI (2009) Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nat Neurosci* 12:655–63.

9. Daniel R, Pollmann S (2012) Striatal activations signal prediction errors on confidence in the absence of external feedback. *Neuroimage* 59:3457–67.

10. Hebart MN, Schriever Y, Donner TH, Haynes J-D (2014) The Relationship between Perceptual Decision Variables and Confidence in the Human Brain. *Cereb Cortex*.

11. Schwarze U, Bingel U, Badre D, Sommer T (2013) Ventral Striatal Activity Correlates with Memory Confidence for Old- and New-Responses in a Difficult Recognition Test. *PLoS One* 8:e54324.

12. Karni A, Sagi D (1993) The time course of learning a visual skill. *Nature* 365:250–252.

13. Reynolds JN, Hyland BI, Wickens JR (2001) A cellular mechanism of reward-related learning. *Nature* 413:67–70.

14. Schultz W (2002) Getting formal with dopamine and reward. *Neuron* 36:241–263.

15. Petrov AA, Dosher BA, Lu Z-L (2005) The dynamics of perceptual learning: an incremental reweighting model. *Psychol Rev* 112:715–43.

16. Delgado MR, Nystrom LE, Fissell C, Noll DC, Fiez JA (2000) Tracking the Hemodynamic Responses to Reward and Punishment in the Striatum. *J Neurophysiol* 84:3072–3077.

17. Knutson B, Adams CM, Fong GW, Hommer D (2001) Anticipation of Increasing Monetary Reward Selectively Recruits Nucleus Accumbens. *J Neurosci* 21:1–5.

18. Gläscher J (2009) Visualization of group inference data in functional neuroimaging. *Neuroinformatics* 7:73–82.

19. Preuschoff K, Bossaerts P, Quartz SR (2006) Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* 51:381–90.

20. Schultz W, Dayan P, Montague PR (1997) A Neural Substrate of Prediction and Reward. *Science* 275:1593–1599.

21. O'Doherty JP et al. (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304:452–4.

22. Berns GS, McClure SM, Pagnoni G, Montague PR (2001) Predictability modulates human brain response to reward. *J Neurosci* 21:2793–2798.

23. Duijvenvoorde ACK Van et al. (2014) A cross-sectional and longitudinal analysis of reward-related brain activation: Effects of age, pubertal stage, and reward sensitivity. *Brain Cogn* 89:3–14.

24. Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. *Proc Natl Acad Sci U S A* 103:3863–3868.

25. Allefeld C, Haynes JD (2014) Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. *Neuroimage* 89:345–357.

26. Ashby FG, Ennis JM, Spiering BJ (2007) A neurobiological theory of automaticity in perceptual categorization. *Psychol Rev* 114:632–656.

27. Hélie S, Ell SW, Ashby FG (2015) Learning robust cortico-cortical associations with the basal ganglia: An integrative review. *Cortex* 64:123–135.

28. Pearson JM, Heilbronner SR, Barack DL, Hayden BY, Platt ML (2011) Posterior cingulate cortex: Adapting behavior to a changing world. *Trends Cogn Sci* 15:143–151.

29. Sescousse G, Caldú X, Segura B, Dreher J (2013) Processing of primary and secondary rewards: A quantitative meta-analysis and review of human functional neuroimaging studies. *Neurosci Biobehav Rev*:1–16.

30. Den Ouden HEM, Friston KJ, Daw ND, McIntosh AR, Stephan KE (2009) A dual role for prediction error in associative learning. *Cereb Cortex* 19:1175–1185.

31. Kok P, Rahnev D, Jehee JFM, Lau HC, de Lange FP (2012) Attention reverses the effect of prediction in silencing sensory signals. *Cereb cortex* 22:2197–206.

32. Bromberg-Martin ES, Matsumoto M, Hikosaka O (2010) Dopamine in Motivational Control: Rewarding, Aversive, and Alerting. *Neuron* 68:815–834.

33. Hajcak G, Foti D (2008) Errors Are Aversive. *Psychol Sci* 19:103–108.

34. Köhler W (1925) *The mentality of apes* (Liveright, New York, NY).

35. Cousineau D (2005) Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutor Quant Methods Psychol* 1:42–45.

36. Schlagenhauf F et al. (2013) Ventral striatal prediction error signaling is associated with dopamine synthesis capacity and fluid intelligence. Hum Brain Mapp 34:1490–9.

37. Kaernbach C (1991) Simple adaptive testing with the weighted up-down method. Percept Psychophys 49:227–9.

436 38. García-Pérez MA (1998) Forced-choice staircases with fixed step sizes: asymptotic and
437 small-sample properties. Vision Res 38:1861–81.

438 39. Liu J, Lu Z-L, Dosher BA (2010) Augmented Hebbian reweighting : Interactions between
439 feedback and training accuracy in perceptual learning. J Vis 10:1–14.

440 40. Petrov AA, Dosher BA, Lu Z-L (2006) Perceptual learning without feedback in non-
441 stationary contexts: data and model. Vision Res 46:3177–97.

442 41. Barto AG, Sutton RS, Brouwer PS (1981) Associative search network: A reinforcement
443 learning associative memory. Biol Cybern 40:201–2.

SUPPORTING INFORMATION

SI Materials and Methods

Participants. Thirty healthy, right-handed participants (all female) took part in the experiment in return for payment after giving written informed consent. To reduce between-subject variance, we included female participants only. One participant was excluded due to fixation failure, leaving 29 valid participants (24.1 ± 2.5 years, range 19 – 31 years). The sample size was determined based on a study that reported an association between a striatal prediction error signal and fluid intelligence (N = 28; 36). The study was conducted according to the declaration of Helsinki, and approved by the ethics committee of the Charité Universitätsmedizin Berlin.

Setup and schedule. Each participant came in for 4 sessions. The training session took place in an fMRI scanner with a back-projection screen setup (Sanyo PLC-XT21L, 60 Hz, resolution 1024 x 768). Participants responded with their right hand using an MR-compatible trackball (Current Designs Inc., Philadelphia, PA) by pressing the left button with the thumb, the right button with the middle finger and navigating the trackball with the index finger. Around one week (7.1 ± 0.2 days) before and one day after the training session, the behavioral pre- and post-test took place in a darkened room, in which the participant sat in front of a 17'' LCD monitor (LG Flatron L1750S, 60 Hz, resolution 1024 x 768) and operated with an equivalent trackball device. Around 10 weeks (70.6 ± 2.1 days) after the training session, a long-term test was conducted with a setup identical to the pre- and post-test sessions.

Trial. Each trial started with a fixation screen for 2000 ms ± 1000 ms, followed by the presentation of the stimulus for 100 ms. After 2000 ms ± 1000 ms the white fixation cross

turned orange or blue (depending on the response mapping; see section “Training”) for 750 ms to signalize the appearance of the visual analogue confidence rating scale appeared. The rating scale was visualized as a half-open circle (radius r = 4° of visual angle) that linearly increased in width (0.1° to 1° visual angle) and color (black to green), whereby the orientation and angular direction of the circle changed randomly from trial to trial. Participants were instructed that the confidence scale started at the thin black end, representing “very low confidence”, and linearly increased (both in terms of appearance and the corresponding confidence level) to the thick green end, representing “very high confidence”. Participants used the trackball to adjust a sliding white bar according to their experienced level of confidence about the perceived stimulus orientation. No time pressure was imposed. After adjusting the confidence rating bar, participants made a binary judgment about the stimulus orientation using the two buttons of the trackball device. After the button press, the response screen remained up for 1 s.

Test sessions. The aim of the pre-, post- and long-term test (henceforth test sessions) was to determine individual contrast thresholds in the orientation discrimination task. The test sessions were divided into blocks of 16 trials with alternating reference axes and continued until the termination criterion of a staircase procedure was reached. Each block began with a start screen that indicated the reference axis of the upcoming block. The staircase procedure started with a one-up-one-down staircase with a relative stepsize of 0.05 log units to rapidly approximate the rough threshold range (start contrast: c_p=20%; see Equation S1). After three reversals, the algorithm switched to a weighted one-up-two-down staircase (37) for fine-tuning. The ratio of *stepsize down* / *stepsize up* was set to 0.5488 (38) with *stepsize down* set to 0.33%, leading to convergence at a performance of 80.35 percent correct. The termination

criterion was the 9th reversal. Thresholds for the horizontal and vertical references axes were independently adjusted. In order to familiarize with the stimulus materials, all participants performed an additional 8 blocks at maximal contrast (c_p=100%) prior to the pre-test.

Training. The training session in the fMRI scanner comprised an initial adjustment run and 9 training runs, each with 48 trials. During the adjustment run and all training runs the participants viewed only one reference axis ('trained reference axis'), which was assigned to participants based on the parity of their consecutively numbered participant IDs. Participants performed the adjustment run in the scanner prior to the first training run in order to accommodate with the scanner environment and to fine-tune the initial contrast level for the training runs. The adjustment run started at the determined contrast threshold of the pre-test and was subsequently adapted with the same weighted one-up-one-down staircase procedure used in the test sessions, targeting a performance of 80.35 percent correct. In the critical condition during the training runs, performance was kept constant at 80.35 percent correct by continuously adapting stimulus contrast through the above described one-up-one-down staircase procedure. The training runs included an additional control condition with constant stimulus contrast in an interleaved half of the trials in order to permit an assessment of orientation information encoded in activation patterns of visual cortex without the confound of a changing stimulus contrast. The instruction for the response mapping between stimulus orientation (counterclockwise/clockwise) and response button (left/right) was alternated between runs. The response mapping was indicated at the beginning of a run and additionally in each trial through the color (blue/orange) of the fixation cross (the assignment of color and orientation being counterbalanced across participants).

Note that, apart from the possibility to continuously assess participants' perceptual

sensitivity during the training session, the staircase served two additional purposes: (1) by keeping participants performance just above threshold, we expected to induce high variance in our confidence ratings, thereby increasing the sensitivity to detect neural confidence signals; and (2), by using a relatively high target performance for the staircase procedure, we accommodated a previous study that showed enhanced perceptual learning for training at high relative to low accuracies (39).

Stimuli. The stimuli were based on an additive mixture of a Gabor patch at four possible orientations (+20° or −20° from the vertical or horizontal reference axis) and phase-randomized spectrally filtered noise (40). Each stimulus consisted of a grayscale Gabor patch $G(x,y)$ embedded in a larger field of filtered grayscale noise $N(x,y)$ (40). The luminance $L(x,y)$ of each pixel was an additive mixture of a Gabor term $G(x,y)$ and noise $N(x,y)$, where L_0 was the background luminance of the screen (51.9 Cd/m²):

$$L(x,y) = \left[1 + \frac{c_p}{100}\mathrm{G}(x,y) + \frac{c_n}{100}N(x,y)\right]\mathrm{L}_0 \qquad \text{(Equation S1)}$$

$$G(x,y) = \mathrm{e}^{-\frac{x^2+y^2}{2\sigma^2}}\sin[2\pi f(x\cos\theta + y\sin\theta) + \psi] \qquad \text{(Equation S2)}$$

The peak target contrast c_p, which could take values between 0% and 100%, was continuously adapted by a staircase procedure. The Gabor patches had a fixed phase ψ of 0.25, a spatial frequency f of 1.25 cycles per degree visual angle, and the standard deviation σ of their Gaussian envelope was set to 0.6. The orientation ϑ was set to +20° or −20° from the vertical (0°) or horizontal (90°) reference axis. The Gabor patch was trimmed at values smaller than 0.005, resulting in a radius of 1.87° visual angle.

The noise field $N(x,y)$ was constructed from a random phase and a bandpass-filtered

power spectrum. The power spectrum was generated by subtracting two Butterworth low-pass filters (two-dimensional, resolution 300 x 300 pixel, corresponding to 5.3° x 5.3°; order 3) with cut-off frequencies one octave below and one octave above the spatial frequency of the Gabor patch. The phase spectrum was sampled as a 300 x 300 matrix of uniformly distributed random numbers. The noise contrast c_n was fixed at 15%. Inverse Fourier transformation of the power and phase spectrum resulted in a noise field that effectively interfered with the spatial frequency of the Gabor patch. The additive mixture of the Gabor patch and the noise field was multiplied with a circular filter and cropped at a radius of 2.5° visual angle, such that the stimulus became circular and smoothly faded out to background luminance. The filter was constructed as the inverse of a two-dimensional 300 x 300 pixel Butterworth low-pass filter of order 7 with cut-off frequency 0.275 cycles / degree visual angle. The value of the cut-off frequency ensured that the fading zone overlapped with the outer border of the Gabor patch and the high order of the filter ensured that the fading zone was relatively steep. Note that the luminance of the monitor in test sessions and the projection setup in the training session was equalized through pre-measured color look-up tables.

Stimulus localizer. To independently identify stimulus-responsive regions for a multivariate analysis, we conducted a localizer run with 18 stimulus blocks and 18 baseline blocks of 12s duration in pseudo-randomized order. In the stimulus blocks the Gabor patch was shown with maximal contrast (c_p=100%) at an eccentricity of 5° visual angle and alternated every 250ms between phase ψ and counterphase 1-ψ (phase and eccentricity were identical to the test and training sessions). The orientation of the Gabor patches alternated block-wise between the two orientations shown in the training session of the respective participant (±20° with respect to the trained reference axis). The baseline blocks consisted of the fixation cross only.

To hold participants' attention, they performed an independent color change detection task on the central fixation cross. The task was to press one of the buttons of the trackball device as soon as the fixation cross turned from white to red. They were instructed that, while fixating and performing the task, they should still note and make themselves aware of the Gabor stimuli.

Perceptual learning index. To quantify participants' perceptual learning success for the trained reference axis, the respective pre-test contrast thresholds were subtracted from post-test thresholds (threshold improvement). However, an analysis of the relationship between pre-test thresholds and threshold improvements showed a strong positive correlation ($r_{pearson}$ = 0.73, $p < 0.001$), suggesting that participants starting at higher thresholds had more room for performance to improve. Thus, to correct our perceptual learning index for this substantial learning-unrelated dependency, pre-test thresholds were regressed out from threshold improvements across participants.

Eyetracking. Eyetracking data were successfully collected in 24 participants during the fMRI training session using an infrared video eyetracking system (iView XTM MRI 50 Hz, SensoMotoric Instruments, Teltow, Germany). For 6 other participants eye tracker calibration failed. As a measure of fixation reliability, we computed the percentage of recorded eye gaze positions during stimulus presentation within a circle of 2.5° visual angle in radius around the center of the fixation cross. This radius corresponded to the eccentricity of the first stimulus pixel. The cut-off for exclusion was a percentage of below 95%.

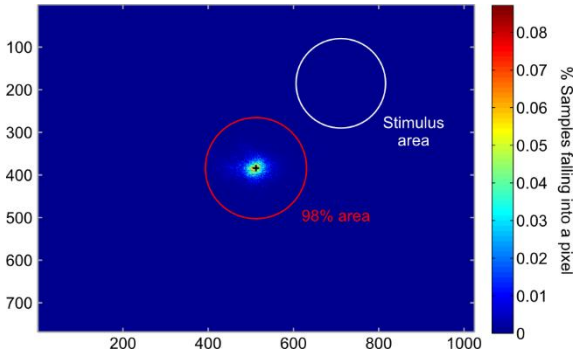


Figure S1. Eyetracking. Heatmap indicating the percentage of eye gaze position at every pixel of the screen. The red circle indicates the area that contained 98% of all eye gaze positions and the white circle depicts the area covered by the Gabor patch. On average, 98.5 ± 0.6% of recorded eye gaze positions during the training session were within the fixation area (radius r = 2.5° of visual angle), demonstrating that the participants maintained fixation throughout the fMRI experiment. One participant was excluded due to fixation failure (< 95%).

Confidence-based perceptual learning model. ***Representational subsystem.*** To compute the orientation energy in the stimuli in each trial, we used a Matlab toolbox developed by Petrov and colleagues (15, 40) (online available at alexpetrov.com/proj/plearn). The relevant component of the toolbox is the "representational subsystem", which takes raw images as input and computes the stimulus energy for pre-specified orientations and spatial frequencies as output. The underlying model is based on an input layer of orientation- and frequency-selective V1 simple cells, which subsequently feed into phase- and location-invariant activation maps. The output of the model is a single energy value for each specified orientation (here: −20°, 20°, 70° or 110°) and spatial frequency (here: 1.25 cycles / degree).

Associative reinforcement learning model. The orientation energy detectors of the representational subsystem (see above) are connected to decision units through *signal weights* (connecting detectors to decision units of the *same orientation*) and *noise weights* (connecting detectors to decision units of the *opposing orientation*). The activities of clockwise (A_{cw}) and counterclockwise (A_{ccw}) decision units are computed through weighted sums:

$$A_{ccw} = E_{ccw}w_{ccw,ccw} + E_{cw}w_{cw,ccw}$$
$$A_{cw} = E_{cw}w_{cw,cw} + E_{ccw}w_{ccw,cw} \quad \text{(Equation S3)}$$

The difference of these output activities constitutes the decision value *DV*:

$$DV = A_{cw} - A_{ccw} \quad \text{(Equation S4)}$$

In addition, the model computes its decisional certainty *c'* proportional to the absolute value of *x* with scaling parameter λ:

$$c' = \lambda|DV| \quad \text{(Equation S5)}$$

The decisional certainty is used to fit the model to participants' confidence reports (see below), which are key for the confidence-based learning rule. The general idea of the confidence-based learning rule is to reinforce circuitry giving rise to higher-than-expected confidence and to weaken circuitry giving rise to lower-than-expected confidence. For this purpose, the model continuously estimates the expected level of confidence ($\bar{c}$) by means of a Rescorla-Wagner rule (Eq. S6) with learning rate α_c. In this way, the CPE δ can be computed as the difference between actual confidence c and expected confidence $\bar{c}$ (Eq. S7).

$$\bar{c} \leftarrow \bar{c} + \alpha_c\delta \quad \text{(Equation S6)}$$

$$\delta = c - \bar{c} \quad \text{(Equation S7)}$$

The model uses an associative reinforcement learning rule (41) to update weights both in relation to the CPE and proportional to the correlated activity of presynaptic activations $E_{ccw/cw}$ and postsynaptic activations $A_{ccw/cw}$:

$$w_{cw,choice} \leftarrow w_{cw,choice} + \alpha_w \delta E_{cw} A_{choice}$$
$$w_{ccw,choice} \leftarrow w_{ccw,choice} + \alpha_w \delta E_{ccw} A_{choice} \ , \qquad \text{(Equation S8)}$$

whereby *choice* represents either *cw* or *ccw*, i.e. indicating the observer's perceptual decision for clockwise or counterclockwise orientation, respectively. The Hebbian component ensures that the update more strongly affects those connections that contribute more to the final choice. The sign of the CPE δ determines whether connections are strengthened or weakened and its absolute value modulates the extent of the update.

Model fit to behavior and model initialization. The likelihood for participants' choices (decisions) *d* is computed through a softmax action selection rule:

$$\mathrm{p}(d|d = cw) = 1 - \mathrm{p}(d|d = ccw) = \frac{1}{1+e^{-\beta x}} \qquad \text{(Equation S9)}$$

The likelihood for participants' confidence reports $c = [0;1]$ is assumed to be normally distributed with standard deviation σ around the model's decisional certainty c':

$$\mathrm{p}(c) \sim \mathcal{N}(c', \sigma) \quad \text{if } c \in]0;1[\qquad \text{(Equation S10)}$$

For the boundary cases $c = 0/1$ the likelihood is computed as the area under the normal density (Eq. S11) in the range $]-\infty; 0]$ for $c = 0$, and $[1; \infty[$ for $c = 1$, respectively.

The free model parameters (α_w, α_c, β, λ, σ) of each participant, as well as the initial values of the signal weights ($w^0_{\mathrm{signal}} = w^0_{ccw,ccw} = w^0_{cw,cw}$) and noise weights ($w^0_{\mathrm{noise}} = w^0_{ccw,cw} = w^0_{cw,ccw}$) were estimated in a two-stage maximum likelihood estimation (MLE) procedure (maximizing the likelihoods p(*d*) and p(*c*)). The first MLE stage served to estimate the initial values of the weights and was based on the pooled data of all participants. We introduced this group-level MLE stage to achieve maximal power for the estimation of the initial weight values. An initial attempt to estimate the weight values at the participant level produced

unreliable estimates, likely due to the non-independence of initial noise weight values and the inverse temperature parameter β (both parameters influence the noise in the decision process). The estimates of w^0_{signal} and w^0_{noise} were then used as initial values in the second subject-level MLE stage, in which the parameters were estimated individually for each participant. See Table S1 for results of this two-stage MLE procedure.

Exemplary time course of behavioral data and model variables. Figure S4 shows an exemplary time course of subjects' behavioral reports (orientation response, confidence rating), the corresponding predictions of the model (choice probability, confidence) and hidden model variables (orientation energy, signal weights, noise weights, expected confidence, confidence prediction error). To convey an intuitive understanding of the model's inner working, we take a closer look at trial 6. In this trial, the presented stimulus happens to have a relatively large difference between clockwise and counterclockwise orientation energy (in favor of clockwise orientation, panel A). Clockwise and counterclockwise energy represent the input to the model, which in a next step, is weighted by the signal and noise weights provided in panels B and C. As to be expected from the relatively large sensory evidence in favor of clockwise orientation, the model predicts a clockwise choice with high probability (panel D, orange line) and high confidence (panel E, orange line), and indeed the participant also responded with clockwise orientation (panel D, blue line) and high confidence (panel E, blue line). The high confidence rating relative to the expected confidence results in a positive confidence prediction error (panel F, green line). After the trial, two components of the model are updated, the expected level of confidence and the weight matrix. The update of expected confidence is proportional to the confidence prediction error and is thus increased (panel F, trial 7, golden line). Finally, the positive confidence prediction error results in an increase of

both the signal weight ($w_{cw,cw}$, panel B, trial 7, turquoise line) and the noise weight ($w_{ccw,cw}$, panel C, trial 7, purple line) of the clockwise decision unit. However, through the Hebbian learning component the weight update is proportional also to the input activity, resulting in a larger increase of the signal weight (originating from the clockwise detector) relative to the noise weight (originating from the counterclockwise detector).

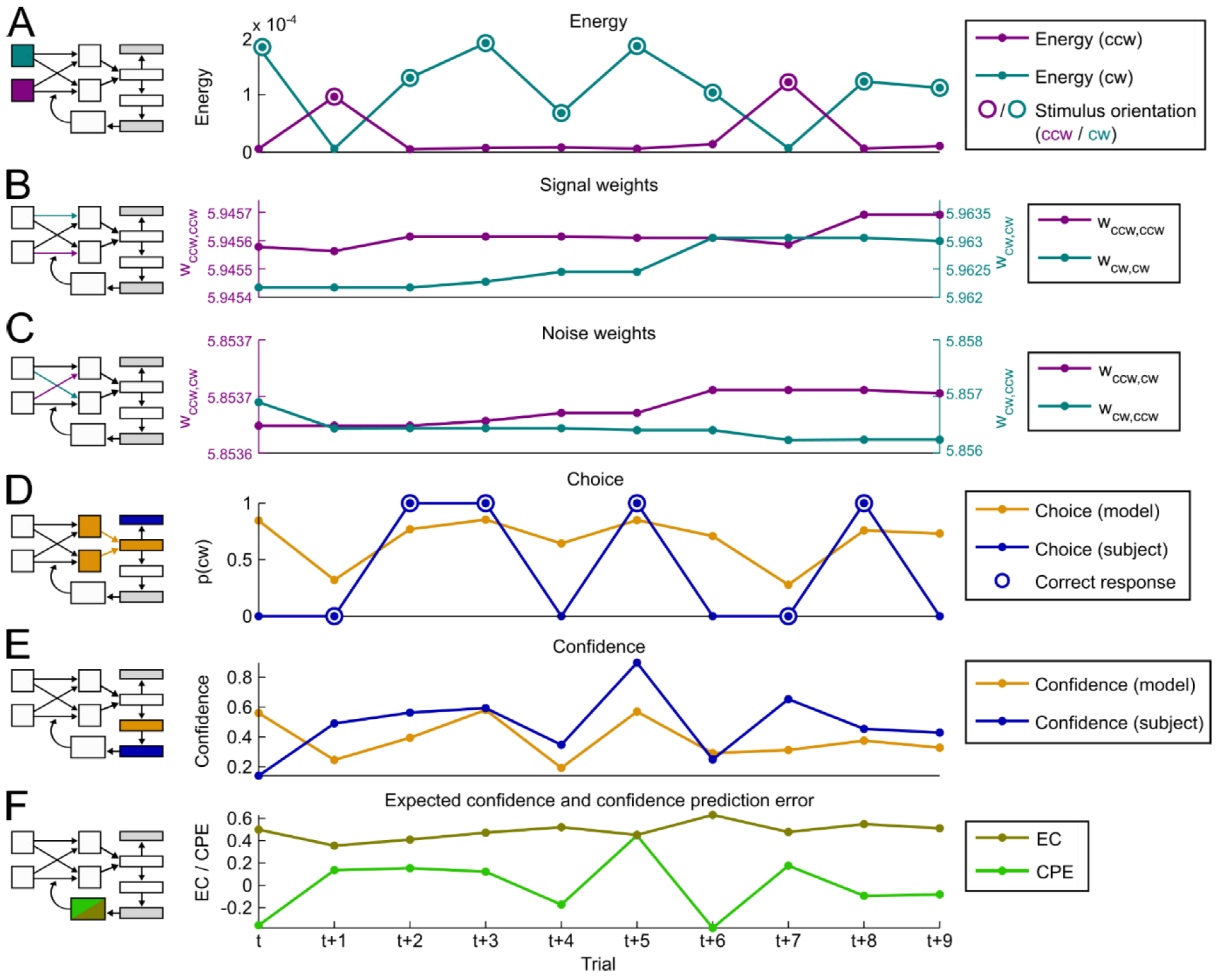


Figure S4. Exemplary time course of model variables and behavioral reports. (A) Energy. Stimulus energy for clockwise (cw) and counterclockwise (ccw) orientation as computed by the representational subsystem. (B) Signal weights. Strength of weights connecting orientation detectors to decisional units of the same orientation. (C) Noise weights. Strength of weights connecting orientation detectors to decisional units of the opposing orientation. (D) Choices. Depicted are the model's choice probability for clockwise choices and the subject's actual choices (cw = 1, ccw = 0). Correct subject choices are marked by a circle. (E) Confidence. Confidence ratings predicted by the model (corresponding to $\lambda \cdot |DV|$) and subject's actual confidence ratings. (F) Confidence prediction error and expected confidence. Depicted are the hidden model variables for the confidence prediction error (CPE) and expected confidence (EC).

FMRI data acquisition and analysis. ***FMRI data acquisition.*** Functional MRI data were acquired on a 3-Tesla Siemens Trio (Erlangen, Germany) scanner using a gradient echo planar imaging sequence and a 12-channel head-coil. In each of the 9 experimental runs on average 201 ± 2 (mean ± SEM) whole-brain volumes were acquired (TR = 2 s, echo time (TE) 25 ms, flip angle 78°, 36 slices, descending acquisition, 3mm isotropic resolution, interslice gap 0.45 mm, tilt angle −20° from ac–pc line). The exact number of volumes could vary from run to run and depended on participants' response times. Additionally, we recorded a high-resolution T1-weighted image (TR = 1.9 s, echo time (TE) 2.51 ms, flip angle 9°, 192 slices, resolution 1 mm isotropic) and a functional localizer run (220 volumes). Preprocessing was performed using SPM8 (www.fil.ion.ucl.ac.uk/spm) and included realignment to the first image, coregistration with the structural image, spatial normalization into the Montreal Neurological Institute reference system and smoothing with an 8 mm Gaussian kernel.

Univariate fMRI data analysis. Two different GLMs were used to model the BOLD response for the univariate model-free (GLM1) and model-based fMRI analysis (GLM2). Both GLMs comprised onset regressors for the stimulus and the response screen and six motion regressors from the realignment analysis. The stimulus regressor was modeled as a stick function and the response screen regressor as a boxcar function with durations equal to the appearance time of the response screen. Both regressors were convolved with a canonical hemodynamic response function. For GLM1 the stimulus regressor was split into three regressors, each representing a tertile of the behavioral confidence reports in a given run (low, middle and high confidence tertiles). In GLM2 an additional regressor for the trial onset (modeled as a stick function) was included, as well as two parametric regressors, accounting for a modulation of the trial onset regressor by expected confidence ($\bar{c}$) and a modulation of

the stimulus onset regressor by the CPE (δ). At the group level, GLM1 was used to test for above-baseline striatal activation at trial onset and for a main effect of confidence at stimulus onset. GLM2 was used in the model-based analysis to test for a positive linear relationship between BOLD signal and expected confidence at trial onset, and for a positive linear relationship between BOLD signal and CPEs at the time of stimulus presentation.

Multivariate fMRI data analysis. In addition, to identify the neural basis of additional model variables we performed a searchlight-based cross-validated multivariate analysis of variance using the cvMANOVA toolbox (25). CvMANOVA provides unbiased estimates of multivariate effects in terms of so called pattern distinctness and allows analyzing arbitrary estimable contrasts. CvMANOVA was performed either on the basis of the voxels within a given ROI or on the basis of a whole-brain searchlight to generate spatial information maps (24). Two separate GLMs were estimated for the analysis of orientation energy (OE) and decision value (DV). As for the univariate analyses, both GLMs included regressors for stimulus onset and the response screen, as well as six motion regressors. In case of the OE model, experimental trials across all runs were sorted into three energy tertiles for clockwise (cw) and counterclockwise (ccw) orientation, leading to the following six stimulus onset regressors: ccw (orientation), high (energy); ccw, middle; ccw, low; cw, low; cw, middle; cw, high. For the DV model we used an equivalent binning procedure: ccw (negative sign of the DV), high (absolute value of the DV); ccw, middle; ccw, low; cw (positive sign of the DV), low; cw, middle; cc, high. As a contrast matrix for the cvMANOVA we used [−2.5 −1.5 −0.5 0.5 1.5 2.5] for both GLMs, i.e. we tested for brain regions exhibiting a linear multivariate relationship with OE or DV. In case of the searchlight analysis we used spheres with a radius of 4 voxels, each containing 257 voxels.

Group-level inference. In all cases, the resulting first-level images (contrast images or

distinctness images) were submitted to a group-level t-test. The statistical cutoff was set to $p < 0.05$, family-wise-error-corrected for multiple comparisons within a region of interest (p_{rFWE}), at the cluster level (p_{cFWE}) with a cluster-defining threshold of $p < 0.001$, or at the whole-brain level (p_{FWE}).

Simulation of the BOLD time course. To simulate the BOLD time course that arises from an initial anticipatory neural response and subsequent prediction error responses, we first defined vectors coding the time course of neural activation for such scenarios. These vectors were 24 s in length (240 data points, i.e.10 Hz sampling rate) and represented the activation between 4 s before and 20 s after stimulus onset. The vectors were all zeros, except for the trial onset at $t = -2$ s, where the vectors were set to +1, and for the time of stimulus presentation at $t = 0$ s, where the vectors were set to +1 for positive prediction errors and −1 for negative prediction errors. Subsequently these vectors were convolved with a canonical double-gamma hemodynamic response function provided by SPM8.

Region of interest procedures. The ventral striatum ROI was based on the Harvard-Oxford cortical and subcortical structural atlases (www.cma.mgh.harvard.edu/fsl_atlas.html) and the ventral tegmental area ROI was derived from the Talairach Atlas (http://www.talairach.org/). For exploratory multivariate analysis we additionally generated a functional ROI based on the stimulus localizer. In a first step we computed a group-level functional localizer T-map based on normalized first-level contrast images (*stimulus>baseline*). After thresholding at $p < 0.001$, we constricted the localizer-based ROI with an anatomical mask based on occipital cortex and fusiform gyrus (using the Anatomical Automatic Labeling atlas available from http://www.gin.cnrs.fr/AAL-217). The final ROIs were created in native subject space by intersecting the reverse-normalized localizer T-map and the anatomical mask.

SI Results

Validation of the representational subsystem. To validate the representational subsystem, we computed the average energy content for a range of spatial frequencies (one octave above and below the actual frequency of 1.25 cycles/degree) and for a range of orientations (−40° to +40° relative to the reference axes). As expected, the energy content was highest for the spatial frequency and orientations used to generate the Gabor patches (Figure S2A), thereby validating the computed orientation energies. In a next step, we computed orientation energies separately for correct and incorrect responses. We found that the orientation energies computed for the *true orientations* (i.e., the orientations of the presented Gabor patches) were *higher for correct than for incorrect responses* of participants ($t_{28} = 8.1$, $p < 0.001$). In contrast, the orientation energies computed for the *opposite orientations* (i.e., ∓20° if ±20° was presented) were *lower for correct than for incorrect responses* ($t_{28} = -3.6$, $p = 0.001$) (Figure S2B). This pattern demonstrated that the varying orientation energy was directly associated with behavior and adds to the validation of the model.

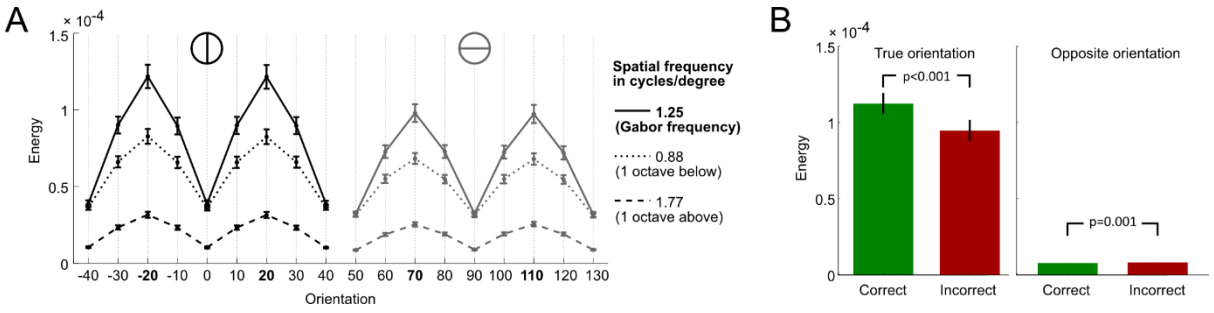


Figure S3. Validation of the representational subsystem. (A) Stimulus energy for three spatial frequencies (Gabor frequency, ± 1 octave) and for orientations ± 20° around the Gabor orientations of the vertical (−20°, 20°) and horizontal (70°, 110°) reference axis. Error bars represent SEM. The stimulus energy is highest for the spatial frequency (continuous lines) and the orientations (bold label) of the Gabor patches. (B) Stimulus energy depending on the correctness of participants' responses for the true orientation of the presented Gabor patches (left panel) and for the opposing orientation (right panel).

To test whether the computed stimulus energy explained participants' performance over and above the effect of stimulus contrast itself, we computed the orientation energies separately for correct and incorrect choices in the training control condition with constant stimulus contrast. Again, the orientation energy computed for the actual orientations was significantly higher for correct versus incorrect responses ($t_{28} = 6.64$, $p < 0.001$), whereas the energies computed for opposing orientations was significantly lower for correct versus incorrect responses ($t_{28} = -2.38$, $p = 0.024$). Thus, the representational subsystem accounted for variance in participants' performance over and above the variance due to varying stimulus contrast.

Confidence reports.

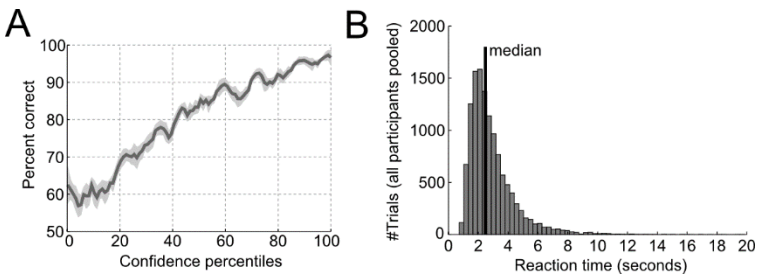


Figure S2. Confidence reports. (A) Performance as a function of confidence percentiles. Confidence ratings during the training session. To assess whether participants correctly used the confidence rating scale and accurately reported their confidence, the relationship between reported confidence and performance (proportion correct) was analyzed by means of a sliding window across sorted confidence values (window size: 5% of all trials). The performance increased monotonically with confidence (main effect of percentile: $F_{22,2442} = 5.76$, $p < 0.001$, one-way ANOVA with repeated measures), approaching chance at low levels of confidence, without showing ceiling effects at high levels of confidence. This pattern indicates that participants accurately reported their level of confidence over the whole spectrum of decisional certainty. (B) Distribution of the pooled response times of all participants. The median response time was 2.47 s.

Relationship between learning rates and the learning process. Although the learning rate α_w influences the speed of weight changes, it is important to note that these weight changes may

not always optimize the filter. This is because stimulus noise can corrupt the Hebbian learning component (leading to a stronger increase of noise weights relative to signal weights) and because confidence noise (reflected as σ in the model) can corrupt the computation of confidence prediction errors. Nevertheless, our data indicate a weak, but significant, relationship between learning rates α_w and threshold slopes during training ($r_{pearson} = -0.37$, $p = 0.048$; please note that more negative slopes correspond to faster learning). The case of the learning rate α_c, which determines the volatility of the expected confidence estimation, is similar: with too high an α_c, the expected confidence changes very erratically, continuously producing high confidence prediction errors, and, since weight changes are proportional to confidence prediction errors, an unstable perceptual filter. Indeed, there was no consistent relationship between α_c and threshold slopes ($r_{pearson} = 0.21$, $p = 0.28$). However, one could expect that participants with a more *reliable* confidence signal (lower σ) tend to update their expected confidence more vigorously (higher α_c). An exploratory analysis revealed a trend in the predicted direction ($r_{pearson} = -0.34$, $p = 0.072$).

Average BOLD time courses for different levels of expected confidence. While in Figure 3B, experimental trials were sorted according to the level of reported confidence, they can also be sorted according to the level of model-derived *expected confidence*. We therefore extracted the BOLD time courses from spheres (radius 3mm) around the bilateral striatal peak voxels of the expected confidence contrast. As evident from Figure S5, expected confidence positively modulates the peak amplitude of the BOLD response, in line with an anticipatory signal. Please note that the slow return to baseline in case of the 'low' tertile (red line) is due to an inherent dependency with the CPE: low expected confidence is more likely to be followed by a positive CPE. In addition, the figure primarily serves an illustrative purpose, as

the signal was extracted from the most sensitive voxels.

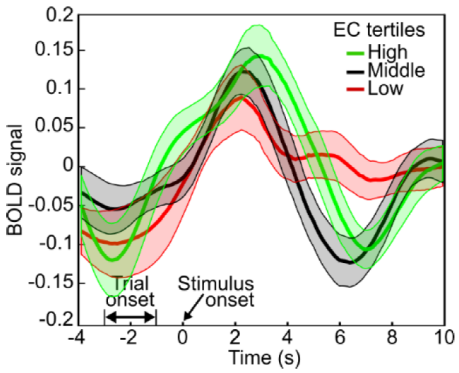


Figure S5. Event-related BOLD time courses in the ventral striatum in dependence of expected confidence (EC). Data were sorted into tertiles of expected confidence, labeled as "low", "middle" and "high" expected confidence. The shaded areas denote SEM.

Could stimulus energy alone explain the modulation of the mesolimbic BOLD signal? Considering previous reports that dopaminergic responses may be modulated by non-rewarding physical stimulus salience, we performed control analyses to account for the effect of stimulus energy in mesolimbic confidence signals. In a first step, we tested whether the unequal distribution of stimulus energy across the three confidence tertiles (low confidence: average energy = $1.0 \cdot 10^{-4}$, middle: $1.1 \cdot 10^{-4}$, high: $1.2 \cdot 10^{-4}$) could explain the *model-free* results of Figure 3B. For that purpose we iteratively sampled trials without replacement from each tertile to achieve roughly equal stimulus energy. In this way, the previous congruency between confidence and energy was successfully removed to the point of being reversed (low: $1.1 \cdot 10^{-4}$, middle: $1.0 \cdot 10^{-4}$, high: $1.0 \cdot 10^{-4}$). Figure S6A shows the BOLD time courses after the correction. While the data are noisier due to the removal of data, the division by confidence is still clearly visible. Thus, absolute orientation energy does not explain the observed effects described in Figure 3B.

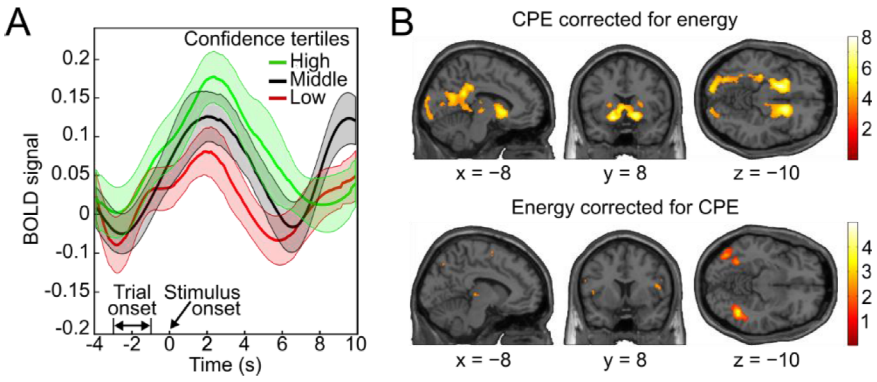


Figure S6. Control analyses accounting for effects of absolute orientation energy. (A) Event-related BOLD time course in the ventral striatum for three tertiles of the behavioral confidence reports, corrected for absolute orientation energy. (B) Top panel: whole-brain T-map for a positive relationship between BOLD signal and confidence prediction error (CPE), after accounting for absolute orientation energy (threshold: $p < 0.001$, uncorrected). Bottom panel: whole-brain T-map for a positive relationship between BOLD signal and absolute orientation energy, after accounting for CPE (threshold: $p < 0.05$, uncorrected).

In a second step, to account for absolute orientation energy in our *model-based* analysis of the confidence prediction error, we extended the GLM of the model-based analysis with a second parametric regressor for absolute orientation energy (i.e., energy for the *presented* orientation) in a way that any variance shared between the energy and the CPE regressor would be accounted for by the energy regressor. We found a significant modulation by energy in bilateral ventral striatum (left: peak at [−8 14 −2], $t_{28} = 3.85$, $p_{rFWE} = 0.036$; right: peak at [10 12 −4], $t_{28} = 5.82$, $p_{rFWE} < 0.001$), but not in the ventral tegmental area. However, the critical analysis was to determine whether CPEs accounted for the modulation of mesolimbic activity over and above the effect of orientation energy. Testing the (residual) modulation by CPEs (Figure S6B, top panel), we still found a strong positive relationship in bilateral ventral striatum (left: peak at [−16 8 −10], $t_{28} = 7.34$, $p_{rFWE} < 0.001$; right: peak at [14 14 −6], $t_{28} = 7.53$, $p_{rFWE} < 0.001$) and in the ventral tegmental area (peak at [−6 −22 −16], $t_{28} = 2.98$, $p_{rFWE} = 0.028$). Conversely, when we tested for a significant modulation of energy in a model, in which variance was first accounted for by the CPE regressor and second by the

energy regressor (i.e., reverse order of the parametric regressors), we found no residual activation in the mesolimbic ROIs (even at a liberal threshold of $p < 0.05$, uncorrected; Figure S6B, bottom panel). The strongest trends for a modulation by stimulus energy on top of CPEs was present in voxels located within our stimulus localizer ROI (left occipital cortex: peak at [−42 −74 −8], $t_{28} = 2.85$, $p = 0.004$, uncorrected; left posterior fusiform gyrus: peak at [−32 −56 −12], $t_{28} = 2.60$, $p = 0.007$, uncorrected). Interestingly, the modulation of activity in putative V1 by CPEs (cf. Table S3) appears to be entirely accounted for by CPEs, as no significant modulation by energy was detectable in this analysis ($p > 0.05$, uncorrected). In conclusion, the mesolimbic activation in our study is predominantly driven by CPEs, whereas a unique modulation by orientation energy may be restricted to high-level stimulus-driven visual areas.

Alternative feedback signals. In the following we discuss and address two alternative feedback signals that potentially could explain behavioral learning and the observed neural responses. ***Salience prediction error (SPE).*** A first possibility is that participants relied merely on stimulus *salience* as internal feedback, which in the context of our model corresponds to the maximum of clockwise and counter-clockwise orientation activity. To test this possibility we implemented a model with salience prediction errors in replacement of confidence prediction errors. (see Table S1 for the formal definition of stimulus salience). A direct comparison between the CPE and the SPE model showed superior model evidence for the original CPE model ($z = 2.89$, $p = 0.004$, Wilcoxon signed rank test; see also Table S1).

Staircase-based feedback. A second possibility is that participants inferred the correctness of their responses from changes in stimulus energy due to the staircase procedure. To quantitatively test the behavioral evidence for staircase-based feedback, we devised a model where we operationalized “difficulty” by the change in the absolute decision variable between

neighboring trials. In our model, difficulty can be operationalized through decisional certainty (|DV|), which corresponds to the amount of task-relevant information that can be read out by the subject. In contrast to our proposed model, the weight update is carried out with a delay of one trial, because only once the stimulus is perceived the feedback about the change in difficulty is available to the subject. The feedback formula and the results of this model are shown together in Table S1. We found that the model evidence for the staircase model is inferior to our original CPE model ($p = 0.008$, $z = 2.7$, Wilcoxon signed rank test; see also Table S1), demonstrating that CPE-based feedback better accounts for the behavioral data.

Beyond this quantitative analysis, there are a number of additional reasons why it is quite unlikely that participants could exploit the staircase algorithm in a meaningful way. First and foremost, in order to obscure the effects of the underlying staircase procedure, the Gabor patches were embedded in a noise field (see section Stimuli in the SI methods), which introduced considerable variability in the effective orientation energy contained in the stimuli. This variability is depicted in Figure S7B, which shows the distribution of orientation energy content separately for the main experimental condition (constant performance) and the control condition with constant contrast. As evident from the figure, the orientation energy is highly variable even in the constant contrast condition. This variability would have considerably impeded the ability of participants to infer changes effected by the staircase procedure, in which the contrast and not the energy content was manipulated directly. Second, since constant contrast and constant performance trials were presented with equal frequency, this implicit feedback would be of use only in one half of the trials, further restricting the usefulness of this information. Third, this half of the trials could only meaningfully be used if participants were able to distinguish between constant performance

and constant contrast trials. This was not the case, as a debriefing showed that not a single participant was aware of the mix of constant contrast and constant performance trials. Thus participants would not have been able to differentiate between contrast changes due to the staircase procedure and contrast changes due to the alternation of conditions. Finally, it should be noted that a staircase-based model would make opposite predictions with respect to the anticipatory confidence signal. If participants had learned that high-confidence trials are more likely to be followed by lower-confidence trials (due to the staircase algorithm), their expected confidence should decrease after a high-confidence trial. Conversely, they should increase expected confidence after low-confidence trials. This is precisely the opposite logic as in our model, in which expected confidence is more likely to be increased after high-confidence percepts and to be decreased after low-confidence percepts. An exploratory fMRI analysis testing for a *negative* relationship with the expected confidence variable from our original model did not reveal any neural basis of such an inverse expectancy signal, neither in our regions of interest (at a liberal threshold of $p < 0.05$, uncorrected) nor at the whole-brain level.

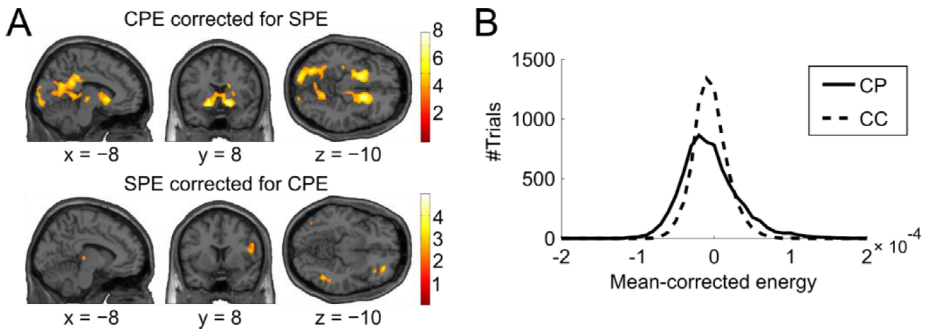


Figure S7. Control analyses addressing alternative feedback signals. (A) Top panel: whole-brain T-map for a positive relationship between BOLD signal and confidence prediction error (CPE), after accounting for salience prediction errors (SPE) (threshold: $p < 0.001$, uncorrected). Bottom panel: whole-brain T-map for a positive relationship between BOLD signal and SPE, after accounting for CPE (threshold: $p < 0.05$, uncorrected). (B) Histogram for the orientation energy in the main experimental condition (constant performance, CP) and the

control condition (constant contrast, CC) pooled across trials of all participants. To pool trials from different participants the mean energy was subtracted separately from each participant's CP and CC condition.

Multivariate ROI analysis. Based on the successful decoding of orientation energy in an occipital region overlapping with the localizer ROI, we explored the possibility of an ROI-based analysis. We reasoned that the distinctness of stimulus-coding patterns within the ROI should remain unchanged across trials of the constant contrast condition, if perceptual learning relies on readout modification and not sensory representation modification. However, an initial analysis based on all data (i.e. pooled data of all runs and not separating between the constant contrast and the constant performance condition) showed only borderline significant decoding of orientation energy (mean pattern distinctness ± SEM, $D = 0.0095 \pm 0.0053$, $t_{28} = 1.80$, $p = 0.041$, one-tailed t-test against the null hypothesis of a pattern distinctness equal to zero). Any further investigation of learning-related changes of decoding accuracies across runs was therefore not reasonable.

Tables

Table S1. Comparison and parameters of perceptual learning models. Parameter values are shown both for the participant-level MLE stage (median ± SE of the median) and for the group-level MLE stage (in brackets). As to be expected, the initial weight values showed a bias towards a higher value for signal weights, reflecting participants' above-chance performance for their starting contrasts (which corresponded to 80.35% correct responses in the pre-test). AIC values indicate mean ± SE, corrected for between-subject variance (39).

	CPE model	**SPE model**	**staircase model**
Feedback type	confidence	stimulus salience	change in difficulty
	c	$\max(A_{cw}, A_{ccw})$	$\lvert DV_t\rvert - \lvert DV_{t-1}\rvert$
α_w	0.0018 ± 0.0007 (0.0013)	0.039 ± 0.009 (0.065)	0.25 ± 0.09 (0.00003)
α_c	0.53 ± 0.08 (0.384)	0.55 ± 0.16 (0.99)	-
β	17.1 ± 1.1 (13.96)	2.80 ± 0.11 (2.64)	215.1 ± 14.0 (174.9)
λ	3.90 ± 0.70 (4.92)	0.72 ± 0.10 (1.04)	48.5 ± 6.5 (64.7)
σ	0.31 ± 0.02 (0.464)	0.29 ± 0.01 (0.56)	0.29 ± 0.01 (0.44)
w^0_{signal}	- (5.82)	- (1.21)	- (0.15)
w^0_{noise}	- (5.63)	- (0.00003)	- (0.14)
AIC	1997.1 ± 4.3	2009.4 ± 2.5	2020.4 ± 6.2

Table S2. List of active brain regions in the model-free fMRI analysis of confidence. List of brain regions exhibiting parametrically modulated activity by confidence, family-wise error corrected at the whole-brain level.

		MNI coordinates					
	Laterality	**x**	**y**	**z**	**df**	**t**	**p_{FWE}**
Positive relationship with confidence							
Ventral striatum	L	-10	14	-4	28	10.11	< .001
	R	16	12	-8	28	11.34	< .001
Posterior cingulate cortex	L	-10	-58	18	28	7.70	< .001
	R	8	-58	20	28	8.60	< .001
Anterior cingulate cortex	L	-10	44	-2	28	6.49	< .001
	R	10	42	-8	28	6.27	< .001
Dorsomedial prefrontal cortex	L	-20	36	52	28	6.38	< .001
Cerebellum	R	46	-70	-38	28	5.88	.002
Supplementary motor area	L	-6	-4	60	28	5.76	.002
	R	10	-2	56	28	5.87	.002
Caudate nucleus	L	-14	10	22	28	5.84	.002
Angular gyrus	L	-44	-72	38	28	5.70	.003
Infracalcarine occipital cortex	L	-14	-94	-6	28	4.96	.042
	R	18	-94	-4	28	5.69	.003
Precentral gyrus / motor cortex	L	-40	-10	58	28	5.35	.011
Superior frontal gyrus	L	-30	20	40	28	5.00	.036
Negative relationship with confidence							
Middle / Anterior cingulate	L	-8	20	44	28	6.73	<.001
	R	10	26	36	28	5.84	.002

Table S3. List of active brain regions in the model-based fMRI analysis of confidence prediction errors (CPEs). List of brain regions exhibiting parametrically modulated activity by CPEs, family-wise error corrected at the whole-brain level. No additional brain regions beyond those reported in the main text were significantly modulated by expected confidence.

		MNI coordinates					
	Laterality	**x**	**y**	**z**	**df**	**t**	**p_{FWE}**
Positive relationship with CPE							
Ventral striatum	L	-16	8	-10	28	7.64	<.001
	R	16	14	-8	28	7.81	<.001
Posterior cingulate cortex	L	-12	-58	22	28	6.23	.024
	R	8	-62	24	28	8.38	<.001
Thalamus		0	-16	10	28	7.81	<.001
Infracalcarine occipital cortex	L	-16	-96	2	28	6.59	.011
	R	20	-96	6	28	6.49	.014
Negative relationship with CPE: no significant effects							