Aus dem
Charité - Centrum für Neurologie, Neurochirurgie und Psychiatrie
Klinik für Psychiatrie und Psychotherapie – Campus Mitte
Direktor: Prof. Dr. med. Dr. phil. Andreas Heinz

Habilitationsschrift

The construction of unambiguous conscious experiences from ambiguous sensory information

Zur Erlangung der Lehrbefähigung
für das Fach Psychiatrie

vorgelegt dem Fakultätsrat der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Dr. med. Veith Weilnhammer

Eingereicht: **März/2023**

Dekan: **Prof. Dr. Joachim Spranger**

1. Gutachterin: **Prof. Dr. med. Alexandra Philipsen, Bad Zwischenahn**

2. Gutachter: **Prof. Dr. med. Jürgen Deckert, Würzburg**

Table of contents

Abbreviations

BOLD: blood-oxygen-level-dependent

CAPS: Cardiff Anomalous Perception Scale

CW/CCW: clock/counter-clock-wise

DCM: dynamic causal modeling

fMRI: functional magnetic responence imaging

FOV: field of view

FWE: family-wise error

GLM: general linear model

HGF: hierarchical gaussian filter

IFC: inferior frontal cortex

IFG: inferior frontal gyrus

IPL/SPL: inferior/superior parietal lobulus

MAT: matched ambiguous transition

MRT: matched replay transition

PANNS: Positive and Negative Symptom Scale

PDI: Peters Delusions Inventory

PE: prediction error

PMF: posterior-medial frontal gyrus

ROI: region of interest

SMA: supplementary motor area

SCZ: schizophrenia

SPM: statistical parametric mapping

TMS: transcranial magnetic stimulation

1. Introduction

The stream of conscious experience portrays a world that appears clear, stable, structured, and rich in detail[1]. Yet, while phenomenal consciousness seems to unfold without effort, the underlying mechanism is in fact one of the greatest mysteries of our time[1–11].

Consider the quality of conscious experiences in relation to the underlying information that is provided by the senses. In vision, we perceive objects and scenes that are typically characterized by a high degree of detail and a stable three-dimensional structure across the whole visual field[1,12]. Yet this experience of detail and stability surpasses the quality of the sensory data by far: For example, the eyes only gather fine-grained data at the central degree of the visual field, which changes in location at every saccade[12]. In addition, to infer the location and shape of objects in our three-dimensional sensory environment, the central nervous system combines the ambiguous and incomplete information from the two eyes in a computation that is, in itself, subject to noise[13,14].

Given these exemplary sources of uncertainty in perception, all signals detected by the senses are inevitably compatible with a multitude of conflicting conscious experiences[15]. The difficult task for the brain is thus to select the interpretation that is most likely to align with reality[15–20], while suppressing all competing alternatives[21]. In concert with its additional functions, the central nervous system requires approximately 20 Watts[22] (W) to execute the neural processes that generate these highly informative and usually veridical conscious experiences. In comparison, a contemporary console uses more than 200 W to render state-of-the-art video-games[23], whose graphics are still far inferior to the richness of human phenomenal consciousness.

1.1 The computational principles of perceptual inference

How does the brain accomplish the task of generating unambiguous conscious experiences from ambiguous sensory information so efficiently and swiftly? According to the influential concept of *perceptual inference*[24], a computational principle pioneered by Hermann von Helmholtz[24], conscious experiences reflect *hypotheses* or *predictions* about the most likely cause of sensory stimulation[15–21].

In the process of determining the causes of sensory stimulation[24], the brain is thought to behave much like a scientist who formulates and tests hypotheses about the environment: At the time of each measurement, the scientist's devices provide only noisy and incomplete pieces of information. Over time, however, the scientist gradually updates her predictions about the environment to better fit the available data. By accumulating her knowledge and refining her hypotheses, the scientist will thereby achieve a highly detailed picture of reality which far surpasses the quality of information associated with each individual measurement.

With respect to the construction of phenomenal consciousness, the analogy of the scientist highlights that unambiguous conscious experiences are driven not only be the ambiguous external data collected by the senses, but are decisively shaped by internal predictions about the statistical properties of the sensory environment[15–20,25]. In computational terms, the set of the predictions used for perceptual inference are often referred to as a *generative model* that emulates the causes of sensory information[18,25].

In analogy to Bayes theorem, it is proposed that the brain uses the generative model to produce internal predictions, reflecting the *prior* probability of a specific cause *c* of sensory stimulation (p(c)). The internal predictions are then integrated with the incoming sensory data that reflect the *likelihood* (i.e., the probability of the data given a specific cause of sensory stimulation *s*, p(s|c)). The result of this integration yields the *posterior* probability of a specific cause of sensory stimulation given the data[18,25] (p(c|s), Figure 1A). Based on the probabilistic integration of prior and likelihood, the generative model generates conscious experiences by selecting the cause of sensory stimulation that is associated with

the highest posterior probability[15–20,25]. Along this line of thought, conscious experiences therefore reflect *controlled hallucinations*[26], i.e., internal predictions that are continuously aligned with the sensory data.

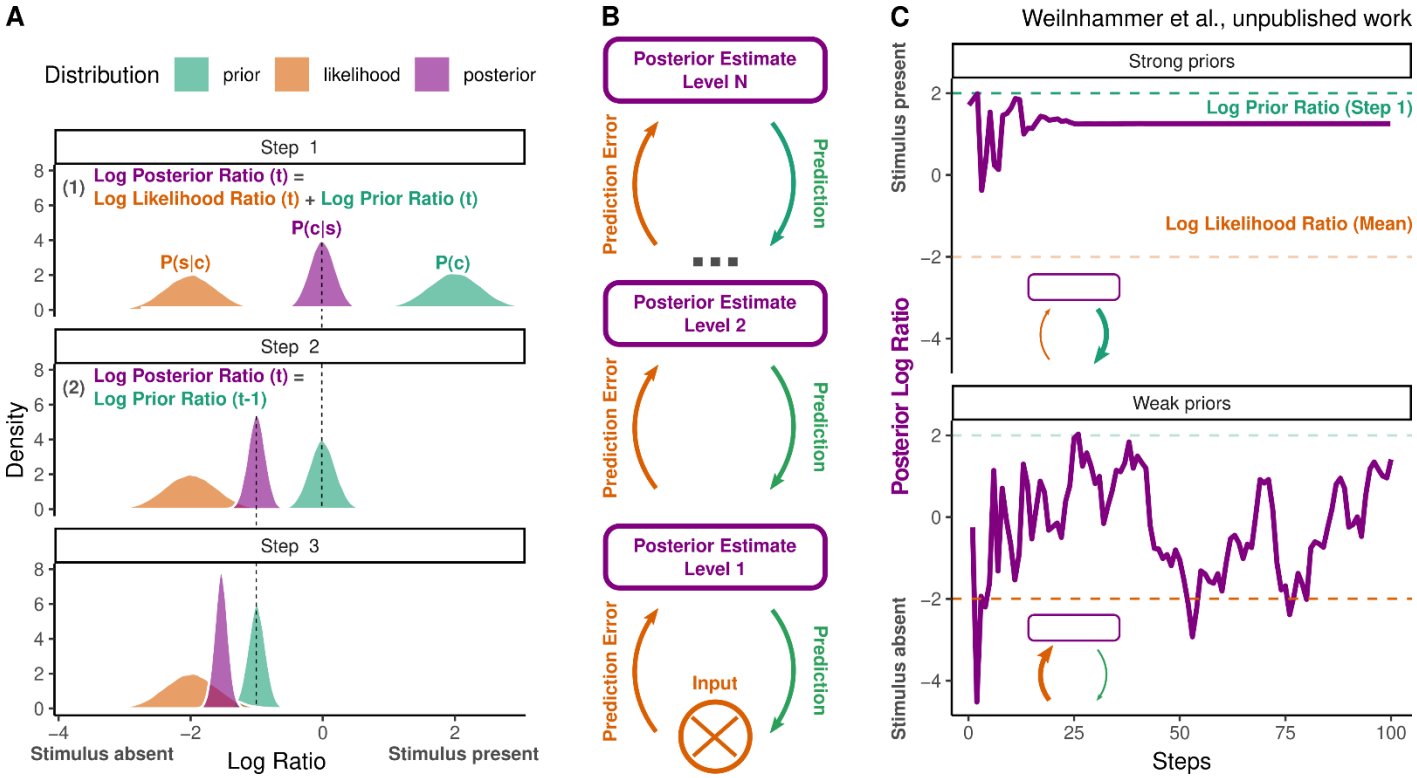


Figure 1. A. Conscious experience as perceptual inference. According to Bayes rule, perceptual inference integrates an internal prediction about the causes *c* of sensory information *s* (e.g., the prior probability of the presence of an object in the environment, p(c)) with the state's likelihood (i.e., the probability of sensory information s given c, p(s|c)) into the posterior probability of a cause of sensory information (the probability of c given s, p(c|s), equation 1). In this example, an observer expects the presence of a stimulus at step 1 (prior log ratio of 2). Sensory information, in turn, speaks against the presence of a stimulus (log likelihood ratio of -2). The posterior turns into the prior for the subsequent time step (vertical dashed lines, equation 2). Over time, the posterior approaches the true cause of sensory stimulation, which is, in this example, the absence of a stimulus (posterior log ratio at step 3 of -1.8). **B. Predictive coding.** Hierarchical predictive coding is one way to implement Bayesian perceptual inference. In this algorithm, predictions are sent from higher levels to lower levels via feedback connections. At each levels, internal predictions are compared against incoming data, starting at the sensory organs. In case of mismatch, prediction errors are fed forward from lower to higher levels. These errors induce updates to the model that continuously align the internal predictions with the incoming data. **C. Strong and weak priors.** There are a least two ways to hallucinate a stimulus: When internal predictions receive too much weight, posterior estimates are drawn away from conflicting sensory information (strong priors, upper panel). When prediction errors receive too much weight, posterior estimates are driven toward noisy fluctuations in the incoming data (weak prior, lower panel). Both scenarios may cause hallucinatory experiences (i.e., false alarms caused by positive posterior log ratios).

How can biological neural networks instantiate Bayesian perceptual inference? In the algorithm of hierarchical predictive coding[27–31], signals that are detected by the senses travel along a hierarchy of ascending processing levels (Figure 1B). Each level compares the external sensory information that is fed forward from lower levels with internal predictions that are fed back from higher levels. In case of a mismatch, each processing level generates a prediction error that is fed forward to the next higher level, inducing updates to the network that improve the alignment between internal predictions and external sensory information[27–31]. Crucially, the neural signals that carry predictions and prediction errors encode the respective precision of internal and external information: When the sensory environment is highly predictable, predictive signals have a stronger impact on inference[30,31]. Conversely, prediction errors drive larger changes in internal predictions when sensory signals are more reliable[30,31].

From the buildings blocks of neurons that represent either predictions or prediction errors, hierarchical predictive coding[27–31] creates a generative and probabilistic model of the sensory environment that increases in complexity from lower to higher levels[31]. Hierarchical predictive coding thereby explains how the brain generates unambiguous and informative conscious experiences from highly ambiguous sensory data[15–20,25]. Moreover, the compression the incoming information into prediction errors reduces the bandwidth of perceptual processing, explaining the energy-efficiency of the brain in constructing highly informative conscious experiences despite the ambiguity inherent in sensory data[22,29].

While predictive coding is only one algorithm capable of instantiating perceptual inference[32], a growing body of empirical evidence has suggested that the central nervous system may indeed realize perceptual inference via hierarchical predictive coding[30,33]. What has so far remained most controversial, however, is whether the predictive processes that shape conscious experiences are confined within primary sensory cortex or, alternatively, require supramodal brain activity in prefrontal cortex[34,35].

1.2 Hallucinations as alterations in perceptual inference

The above section illustrates how conscious experiences can be thought of as *controlled hallucinations*, i.e., internal predictions that are continuously aligned with the sensory data[26]. Hierarchical predictive coding may thereby serve adaptive functions for the central nervous system, such as stabilizing conscious experiences against uninformative fluctuations in sensory information, or reducing the overall energy demands of perception[36–38]. At the same time, however, relying on generative models may come at a cost, since imbalances in cortical feedback-feedforward loops may lead to *uncontrolled hallucinations*, i.e., internal predictions to diverge from the true cause of sensory stimulation[39–44].

How can altered perceptual inference lead to the experience of hallucinations? In general, hallucinations represent vivid and often detailed conscious experiences that occur in the absence of a corresponding external stimulus. They frequently occur in patients suffering from psychotic disorders such as schizophrenia, but are also found drug-induced states and across the neuro-psychiatric spectrum, such as in delirium, depression, mania, post-traumatic stress disorder, Parkinson's and Parkinson's Plus syndroms or Alzheimer's disease[42]. Hallucinations can even be found in the healthy population at a frequency of up to 50%[42].

Neuroimaging experiments based on symptom capture (i.e., hallucinators indicating the presence of hallucinatory experiences via button-press during a recording of brain activity) found that hallucinatory experiences correlate with increased neural activity in sensory networks dedicated to object recognition[41,45,46]. Importantly, electrical stimulation of these areas is known to induce comparable conscious experiences in the corresponding modality[47]. Like electrical stimulation, internal predictions are also capable of causing neural activity in sensory cortices[48–50] along with the respective conscious experiences[50–52]. Therefore, alterations in hierarchical predictive coding may provide a promising neurocomputational account of hallucinations[40–42].

The idea that aberrant predictive processes may cause hallucinatory experiences has been backed up by recent experiments on *induced hallucinations* in humans[39] and mice[44]. These studies created ambiguity regarding the presence of a target by titrating the presented stimuli to the psychophysical threshold. The authors then induced strong internal predictions about the presence of the target via cross-modal conditioning[39] or manipulations of signal probability[44]. In these signal detection paradigms, induced hallucinations are then defined as *false alarms* that occur when participants perceive spurious signals in sensory noise[44].

A growing body of evidence illustrates that such false alarms provide a valid phenotype for clinical hallucinations: Induced hallucinations occur more frequently in individuals who are prone to hallucinations[39,44]. Moreover, induced hallucinations correlate with neural activity in areas that overlap

with the correlates of hallucinatory experiences observed during symptom capture[39,41,45,46]. Finally, induced hallucinations can be triggered by enhancing dopamine[44,53], the primary molecular endophenotype of psychosis[54], and by ketamine[44], a NMDA-receptor antagonist that is used to model psychotic symptoms in healthy participants[55].

Yet while the general link between hallucinations and altered perceptual inference has gained growing empirical support, the neurocomputational mechanism underlying hallucinatory experiences is still heavily debated. Within the field of computational psychiatry[56–58], proponents of the so-called *strong prior account* posit that hallucinations are caused by relying too heavily on internal predictions[42] (Figure 1C, upper panel). According to this view, hallucinatory experiences occur because the impact of predictive feedback on perceptual inference is enhanced relative to the impact of prediction errors. As suggested by work on induced hallucinations in humans and mice, this over-reliance on strong internal prediction may sculpt noisy sensory information into hallucinatory experience[39,44].

The *weak prior account*, in turn, argues for the opposite scenario: In this view, hallucinatory experiences may occur because internal predictions receive too little weight relative to prediction errors[40,59]. When predictive feedback is reduced, neural activity at lower levels of cortical processing is left unexplained[59]. Such a failure of prediction-based attenuation of noise may cause perceptual inference to be driven by excessive prediction errors, thereby creating the neural substrate for hallucinatory experiences[41,45,46,59] (Figure 1C, lower panel). In line with the weak prior account, patients suffering from schizophrenia seem to be less susceptible to contextual effects in visual illusions[51] and to sensory attenuation in force-matching experiments[60], arguing for attenuated predictive feedback in people who experience hallucinations.

How could the discrepancy between strong and weak priors be resolved? Recent work has argued that, when prediction errors escalate due to attenuated predictive signaling at lower levels, overly strong internal predictions may arise at higher levels of the cognitive hierarchy as a compensatory mechanism[40–42]. This situation may lead to circular inference, where sensory information that is reverberated between adjacent levels of processing, ultimately creating hallucinatory experiences[61]. So far, however, all of the outlined neurocomputational accounts of hallucinations - the circular as well as the strong and weak prior account - await empirical confirmation.

1.3 Bistable perception as a tool to study the behavioral and neural correlates of perceptual inference in health and disease

The above sections summarize how hierarchical predictive coding may explain how the brain generates unambiguous conscious experiences from ambiguous sensory information, and how alterations in the process may lead to psychotic symptoms such as hallucinations. In the following, I will outline how the phenomenon of bistable perception[62] can be used to investigate the behavioral and neural correlates of perceptual inference in health and disease.

Bistable perception occurs when observers view an ambiguous stimulus that is compatible with two mutually exclusive interpretations, typically resulting in transitions between the two possible conscious experiences[62–65] (Figure 2A). Transitions between the two occur spontaneously and in the absence of any change in visual stimulation[62–65]. Importantly, these transitions mark a process by which our perceptual system establishes an unambiguous conscious experience in the light of ambiguous sensory information[21,66]. Thereby, bistable perception highlights a fundamental aspect of perceptual inference: As our brains do not have direct access to the events in the world, they constantly face the task of inferring the most likely cause of the ambiguous data registered by the senses. Transitions in bistable perception hence provide a unique window onto the nature of conscious experience[21,67,68] and its neural correlates[34,35].

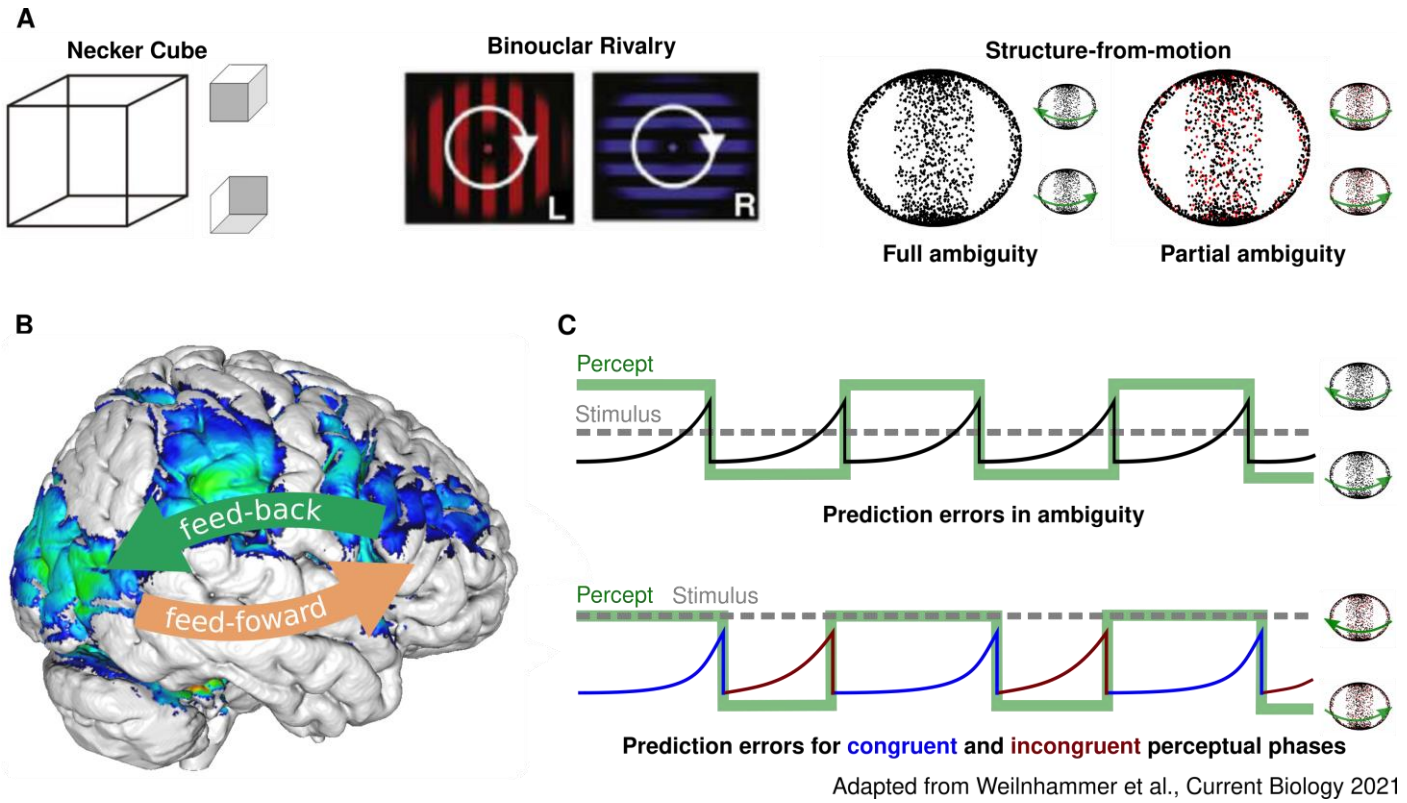


Figure 2. A. Bistable stimuli. Bistable perception arises when sensory information is compatible with two mutually-exclusive perceptual interpretations. Well-known examples include the Necker Cube (left panel), which can be seen in two spatial orientations, or binocular rivalry (middle panel), where the two eyes are presented with conflicting images that induce spontaneous transitions of experience between the monocular stimuli. During structure-from-motion (right panel), dots moving in 2D cause the illusion of a 3D object that rotates around a vertical axis. Over time, observers experience spontaneous transitions in the perceived direction of rotation. Importantly, the structure-from-motion stimulus can be partially disambiguated by adding a 3D signal to a subset of its dots (highlighted in red) using filter glasses. Increasing the number of disambiguated dots increases the signal-to-ambiguity ratio of the stimulus. Videos of the ambiguous and partially disambiguated stimuli are available on "https://veithweilnhammer.github.io/reveal/Consciousness /Content/RDK_Ambiguity_immediate.mp4" and ".../RDK_Graded_immediate.mp4". **B. The role of inferior frontal cortex in bistable perception.** Spontaneous transitions during bistable perception activate not only feature-selective regions in visual cortex, but also supra-modal areas in parietal and prefrontal cortex (i.e., the so-called frontoparietal network). Most consistently, previous research has shown increased activity in the IFC (i.e., the anterior insula and the adjacent inferior frontal gyrus) at the time of perceptual transitions during bistable perception. It has so far remained controversial whether IFC activity is a down-stream consequence of perceptual transitions realized at the level of feature-selective regions in visual cortex such as V5/hMT+ for motion (the *feedforward* model), or, alternatively, represents the cause of perceptual transitions during bistable perception (the *feedback* model). A *hybrid* model based on hierarchical predictive coding model proposes that, due to stimulus ambiguity, perceptual inference can never fully account for the sensory data. This triggers a prediction error signal that is fed forward from feature-selective regions in visual cortex to IFC. Following the prefrontal accumulation of prediction errors, feedback from IFC is thought to trigger a change in conscious experience, thereby temporarily reducing the prediction error. **C. Prediction errors during bistable perception.** During bistable structure-from-motion, observers perceive spontaneous transitions in direction of rotation (green line), alternating between left- and rightward motion of the front-surface (icons on the right). For fully ambiguous stimuli (upper panel; grey dotted line), prediction errors (black solid line) accumulate while perception remains constant (i.e. during a perceptual *phase*), until a change in conscious experience leads to a reduction in prediction errors. For partially disambiguated structure-from-motion stimuli (lower panel), conscious experience can either be congruent or incongruent with the available sensory information. When conscious experience is congruent with the disambiguating stimulus information, prediction errors are attenuated (blue line) and transitions in conscious experience occur less frequently (i.e., longer phase durations). Conversely, when conscious experience is incongruent with the disambiguating stimulus information, prediction errors are enhanced (red line) and transitions in conscious experience occur more frequently (i.e., shorter phase durations).

Over the past two decades, the controversy regarding the implication of prefrontal brain activity in conscious experience[34,35] has therefore reverberated in research on the neural processes involved in

perceptual transitions during bistability[69,70]. Functional neuroimaging in humans has pointed to a key role of the right inferior frontal cortex (IFC), a region that is consistently more active at the time of perceptual transitions during bistability as compared to perceptual events evoked by changes in visual stimulation[69] (often referred to as *replay*[71]). However, the precise role of the IFC in bistable perception is still a matter of debate. So far, it has remained elusive whether activity in this area constitutes a potential cause[67,72,73] or rather the consequence[69,71,74] of spontaneous changes in the contents of conscious experience (see Figure 2B). Progress on the neural correlates of bistable perception is intimately tied to elucidating the role of prefrontal brain activity in consciousness.

Previous theoretical and empirical work has suggested that the opposing views of IFC activity during bistable perception may be reconciled within the framework of hierarchical predictive coding[21,69,75]. The predictive coding model of bistable perception[21] is built on the general assumption that conscious experience represents the posterior prediction regarding the most likely cause of the available sensory information, i.e., the hypothesis that is best at minimizing prediction errors[21]. However, if the available sensory information is fully ambiguous and hence equally compatible with two (mutually exclusive) interpretations, a prediction based on one of two will never fully account for sensory information[21]. According to the predictive coding model of bistable perception, residual evidence for the alternative perceptual hypothesis thus constitutes a prediction error. This prediction error accumulates over time and thereby destabilizes the current conscious experience[21], eventually resulting in a transition to the alternative conscious experience[21] (see Figure 2B and C). Crucially, predictive coding therefore understands perceptual transitions during bistable perception as an attempt to minimize prediction errors by re-attributing the sensory input to the alternative interpretation of the stimulus[21].

It has repeatedly been suggested that predictive processing can help us to understand the functional significance of frontal activity during bistable perception[66,69] (Figure 2C). Previous work has proposed that the reported transition-related activity in the IFC may reflect the accumulation of prediction errors in a process that culminates in a perceptual transition[69]. This notion may resolve the above-mentioned cause-or-consequence controversy on transition-related IFC activity: On the one hand, activity in the IFC may reflect the build-up of prediction errors that originate from early processing stages and propagate up the hierarchy to frontal cortex in a feedforward manner. On the other hand, the IFC may in turn engender a feedback modulation of activity in visual cortex that facilitates a perceptual transition, thereby minimizing prediction errors[21,73].

In sum, progress in the long-standing debate regarding the role of prefrontal brain areas in bistable perception is therefore highly relevant for the inferential processes that give rise to conscious experience[21,67,68]. Elucidating the neural correlates of bistable perception will therefore expand our understanding of the neurobiological underpinnings of consciousness[34,35,76]. Given the prominent role of perceptual inference in contemporary theories on the origin of psychotic symptoms, bistable perception also represents an important experimental paradigm in the context of computational psychiatry[56–58], particularly with respect to a neurocomputational explanation of psychotic symptoms[77–79]. Experiments that modulate internal predictions and externally-driven prediction errors during bistable perception[37] will therefore be instrumental in deciphering the neurocomputational underpinnings of hallucinations.

2. Publications

2.1 *A predictive coding account of bistable perception – a model-based fMRI study*

Weilnhammer VA, Stuke H, Hesselmann G, Sterzer P, Schmack K. PLOS Computational Biology 13, e1005536 (2017). DOI: https://doi.org/10.1371/journal.pcbi.1005536

The neural correlates of bistable perception are highly relevant for understanding how the central nervous system generates the contents of conscious experience[21,34,35,68,76]. Previous research has shown that neural activity in supra-modal brain regions within the so-called *frontoparietal network*[66,69] is enhanced at the time of transitions in conscious experience during bistable perception. Crucially, functional magnetic resonance imaging (fMRI) results from my dissertation thesis revealed that effective connectivity from prefrontal to visual cortex is enhanced at the time of transitions in conscious experience during bistable perception, suggesting a causal role of frontoparietal cortex in resolving sensory ambiguities[73]. Here, we asked whether neural activity in frontoparietal cortex can be explained in terms of hierarchical predictive coding[27–31]. We developed a predictive coding model of bistable perception[21] that can be fitted to behavioral data, i.e., when participants report the content of conscious experience using button-presses. Prediction errors derived from the computational model (Figure 2C) correlated with BOLD-activity in the inferior frontal cortex (IFC, the anterior insula and the adjacent inferior frontal gyrus). This finding suggests that the active role of frontoparietal cortex during bistable perception[73] is be linked to the representation of prediction errors[21,80].

The following text corresponds to the abstract of the article[80]:

“In bistable vision, subjective perception wavers between two interpretations of a constant ambiguous stimulus. This dissociation between conscious perception and sensory stimulation has motivated various empirical studies on the neural correlates of bistable perception, but the neurocomputational mechanism behind endogenous perceptual transitions has remained elusive. Here, we recurred to a generic Bayesian framework of predictive coding and devised a model that casts endogenous perceptual transitions as a consequence of prediction errors emerging from residual evidence for the suppressed percept. Data simulations revealed close similarities between the model’s predictions and key temporal characteristics of perceptual bistability, indicating that the model was able to reproduce bistable perception. Fitting the predictive coding model to behavioural data from an fMRI-experiment on bistable perception, we found a correlation across participants between the model parameter encoding perceptual stabilization and the behaviourally measured frequency of perceptual transitions, corroborating that the model successfully accounted for participants’ perception. Formal model comparison with established models of bistable perception based on mutual inhibition and adaptation, noise or a combination of adaptation and noise was used for the validation of the predictive coding model against the established models. Most importantly, model-based analyses of the fMRI data revealed that prediction error time-courses derived from the predictive coding model correlated with neural signal time-courses in bilateral inferior frontal gyri and anterior insulae. Voxel-wise model selection indicated a superiority of the predictive coding model over conventional analysis approaches in explaining neural activity in these frontal areas, suggesting that frontal cortex encodes prediction errors that mediate endogenous perceptual transitions in bistable perception. Taken together, our current work provides a theoretical framework that allows for the analysis of behavioural and neural data using a predictive coding perspective on bistable perception. In this, our approach posits a crucial role of prediction error signalling for the resolution of perceptual ambiguities.”

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

A predictive coding account of bistable perception - a model-based fMRI study

Veith Weilnhammer[1]*, Heiner Stuke[1], Guido Hesselmann[1], Philipp Sterzer[1,2,3☯], Katharina Schmack[1☯]

1 Department of Psychiatry, Charité Universitätsmedizin Berlin, 10117 Berlin, Germany, **2** Bernstein Center for Computational Neuroscience, Charité Universitätsmedizin Berlin, 10117 Berlin, Germany, **3** Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, 10099 Berlin, Germany

☯ These authors contributed equally to this work.
* veith-andreas.weilnhammer@charite.de

OPEN ACCESS

Citation: Weilnhammer V, Stuke H, Hesselmann G, Sterzer P, Schmack K (2017) A predictive coding account of bistable perception - a model-based fMRI study. PLoS Comput Biol 13(5): e1005536. https://doi.org/10.1371/journal.pcbi.1005536

Editor: Jean Daunizeau, Brain and Spine Institute (ICM), FRANCE

Received: June 30, 2016

Accepted: April 26, 2017

Published: May 15, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: VW is a participant in the Charité Junior Clinical Scientist Program funded by the Charité Universitaetsmedizin Berlin and the Berlin Institute of Health; German Federal Ministry of Education and Research within the framework of the e:Med research and funding concept (01ZX1404A to KS); German Research Foundation (grants HE 6244/1-2 to GH, STE 1430/7-1 to PS); KS is a participant in the Charité Clinical Scientist Program funded by the

Abstract

In bistable vision, subjective perception wavers between two interpretations of a constant ambiguous stimulus. This dissociation between conscious perception and sensory stimulation has motivated various empirical studies on the neural correlates of bistable perception, but the neurocomputational mechanism behind endogenous perceptual transitions has remained elusive. Here, we recurred to a generic Bayesian framework of predictive coding and devised a model that casts endogenous perceptual transitions as a consequence of prediction errors emerging from residual evidence for the suppressed percept. Data simulations revealed close similarities between the model's predictions and key temporal characteristics of perceptual bistability, indicating that the model was able to reproduce bistable perception. Fitting the predictive coding model to behavioural data from an fMRI-experiment on bistable perception, we found a correlation across participants between the model parameter encoding perceptual stabilization and the behaviourally measured frequency of perceptual transitions, corroborating that the model successfully accounted for participants' perception. Formal model comparison with established models of bistable perception based on mutual inhibition and adaptation, noise or a combination of adaptation and noise was used for the validation of the predictive coding model against the established models. Most importantly, model-based analyses of the fMRI data revealed that prediction error time-courses derived from the predictive coding model correlated with neural signal time-courses in bilateral inferior frontal gyri and anterior insulae. Voxel-wise model selection indicated a superiority of the predictive coding model over conventional analysis approaches in explaining neural activity in these frontal areas, suggesting that frontal cortex encodes prediction errors that mediate endogenous perceptual transitions in bistable perception. Taken together, our current work provides a theoretical framework that allows for the analysis of behavioural and neural data using a predictive coding perspective on bistable perception. In this, our approach posits a crucial role of prediction error signalling for the resolution of perceptual ambiguities.

Charité Universitaetsmedizin Berlin and the Berlin Institute of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

In bistable vision, perception spontaneously alternates between two different interpretations of a constant ambiguous stimulus. Here, we show that such spontaneous perceptual transitions can be parsimoniously described by a Bayesian predictive coding model. Using simulated, behavioural and fMRI data, we provide evidence that prediction errors stemming from the suppressed stimulus interpretation mediate perceptual transitions and correlate with neural activity in inferior frontal gyrus and insula. Our findings empirically corroborate theorizations on the relevance of prediction errors for spontaneous perceptual transitions and substantially contribute to a longstanding debate on the role of frontal activity in bistable vision. Therefore, our current work fundamentally advances our mechanistic understanding of perceptual inference in the human brain.

Introduction

During bistable perception, observers experience fluctuations between two mutually exclusive interpretations of a constant ambiguous input. Remarkably, percepts evoked by ambiguous stimuli usually closely resemble the experience of unambiguous objects and thus illustrate the constructive nature of perception. However, the mechanisms driving transitions in bistable perception remain poorly understood.

Previous neuroimaging work [4, 5, 6, 7, 8, 9, 10] has sought to distill the neural processes underlying bistable perception by recurring to a 'replay' condition, in which physical stimulus changes mimic the perceptual alternations induced by ambiguous stimuli. This approach revealed a right-lateralized assembly of fronto-parietal areas whose activity is specifically enhanced during endogenously evoked transitions (ambiguity) as compared to exogenously evoked transitions (replay) [4, 5, 7, 9].

However, the functional role of fronto-parietal areas in bistable perception is a matter of ongoing debate. According to one view, transitions in bistable vision are primarily a result of adaptation and inhibition within visual cortex, while switch-related activations in fronto-parietal areas reflect a mere 'feedforward' consequence of neural events at sensory processing levels [6, 10]. Another view proposes that fronto-parietal areas may be involved in stabilizing and destabilizing perception, thus causally contributing to perceptual switching via 'feedback' mechanisms [4, 5, 11, 7]. Here, we sought to resolve this debate by using model-based fMRI to empirically test a theoretical model that has the potential to integrate these two seemingly contradictory views of perceptual bistability.

From a theoretical perspective, endogenous transitions might be explained by framing perception as an inferential process generating and testing hypotheses about the most likely causes of sensory stimulation [12, 13, 14]. Such processes can be elegantly implemented by hierarchical predictive coding [15, 16, 17]. Here, 'predictions' encoded at higher levels are compared against 'sensory input' represented at lower levels, while a mismatch between the two elicits a prediction error, updating higher-level predictions [15]. Such belief-updating schemes can be translated onto Bayes' rule, where prior distributions ('predictions') are combined with likelihood distributions ('sensory input') into posterior distributions in a sequential manner [16, 18].

Here, we tested whether this framework provides a mechanistic explanation for perceptual transitions and related neural activity during bistable perception. We devised a computational model that formalizes perceptual decisions (i.e., decisions that define the content of conscious perception, as indicated by participants' response) to be performed on the basis of posterior probability distributions. This model is a modification of an approach introduced by [19], who

propose that perceptual time-courses during bistable perception result from samples drawn subsequently from a posterior distribution. The authors implement a memory decay favoring recent over older samples as well as stationary prior capturing the effect of context on bistable perception. Our model, in turn, posits that the shape of the posterior distribution changes dynamically over time in response to prediction errors emerging from the currently suppressed interpretation of the ambiguous input. Importantly, this model has the potential to integrate feedforward and feedback mechanisms in bistable perception: The prediction errors arising from sensory processing levels may be propagated up to higher-level brain areas in a feedforward fashion. The registration of prediction errors in higher-level brain areas leads to an updating of predictions that may in turn drive perceptual switching through a feedback mechanism.

To test this hypothesis, we began with data simulations to establish that our model's predictions match the key characteristics of perceptual bistability. We proceeded by fitting our model to behavioural data from a fMRI experiment on bistable perception [7].

In this experiment, participants viewed a Lissajous figure [42] rotating either clockwise (as viewed from above, i.e. movement of the front surface to the left) or counter-clockwise (vice versa) and indicated their current perception via button-presses. Participants were presented with alternating blocks of ambiguous and disambiguated Lissajous figures: In the ambiguous condition, we presented bistable Lissajous figures which elicited spontaneous (endogenous) alternations in perception. In the disambiguated ('replay') condition, we mimicked the endogenous perceptual time-course by introducing exogenous perceptual switches. Ambiguous and disambiguated stimuli were constructed by presenting two Lissajous figures separately to the two eyes: In the ambiguous condition, both eyes received identical stimulation. In the replay condition, the two Lissajous figures were slightly phase-shifted against each other, biasing perception in the direction of the phase shift.

Having inverted our predictive coding approach based on behavioural data from this experiment, we investigated whether our model accurately explains individual perceptual time-courses during ambiguous and replay stimulation.

In a supplementary analysis (see S2 Text), we furthermore compared our model to three established models of bistable perception: Firstly, we tested an oscillator model [1], which is based on mutual inhibition between to competing neural populations coding for the alternative perceptual outcomes during bistable perception. Here, the currently dominant population suppresses activity in the alternative population. However, due to adaptation in the dominant population, this relation reverses over time, leading to regular oscillations in perception. Secondly, we constructed a noise-driven attractor model of bistable perception [2]. In this framework, internal and external sources of noise trigger transitions between two stable states in an attractor network, representing the two perceptual interpretations associated with a bistable stimulus. Thirdly, we tested an intermediate model [3], which contains both adaptive processes and noise. We validated our approach against these models by the use of Bayesian Model Comparison [20].

We then conducted a model-based fMRI-analysis [21] based on the predictive coding model to test whether prediction errors account for transition-related neural activity during bistability. Additionally, we compared the model-based fMRI analysis with conventional fMRI analyses using a Posterior Probability Map (PPM) approach [22].

Methods

Theoretical background

Our Bayesian modelling approach draws on the view that perception is an inferential process in which perceptual decisions are based on posterior distributions [13]. According to Bayes'

rule, the posterior combines information in the current sensory data (likelihood) with information from previous visual experience (prior) in a probabilistically optimal manner. Crucially, this posterior at a given moment becomes a prior for the current perceptual decision, which entails a prediction error signal that influences on the prior at the next moment. Hence, the posterior not only provides the basis for current perception, but also shapes future perception.

In line with previous theorizations [12], we reasoned that the ambiguous likelihood provides equally strong sensory evidence for two different percepts. We further hypothesized that the current percept establishes an implicit prior belief about similar percepts in the future, thereby contributing to stability of visual perception. The application of Bayes' rules combines the likelihood for ambiguous stimuli with the stability prior into a posterior that represents stronger evidence for the dominant percept, but still contains residual evidence for the suppressed percept. While the stronger evidence for the dominant percept will again favor this percept for the upcoming perceptual decision, the residual evidence for the suppressed percept is equivalent to a prediction error that leads to an update of the stability prior.

Over time, the stability prior is weakened and the posterior shifts towards the suppressed percept, paralleled by an escalating prediction error. When the residual evidence for the suppressed percept equals the evidence for the dominant percept, the prediction error reaches a maximum and a perceptual transition is most likely to occur. Once such a transition has occurred, the process starts over again, minimizing the current prediction error.

Please note that our approach was influenced by the work of [19], who argue that bistable perception is a product of Bayesian decision making in ambiguous sensory environments. They study the effects of viewpoint context on perception of the Necker Cube and propose that bistable perception arises from sampling a bimodal posterior distribution. Here, the sample with the highest ‚weight' determines the content of conscious perception. Key elements of their model are (1), a stationary prior, whose precision reflects interindividual differences in the effects of viewpoint context on perception of the Necker Cube and (2), a memory decay that discounts the weight associated with a sample drawn from the posterior distribution by its age and influences on the length of individual phase durations.

In contrast to [19], our model does not assume a specific memory decay process, but controls the length of phase durations by means of the dynamically updated stability prior. In analogy to the stationary viewpoint prior in [19], our model captures the influence of additional sensory evidence on perceptual decisions using a ‚stereodisparity' distribution, whose precision determines the effectiveness of disambiguation.

Please refer to to the mathematical appendix (see S1 Text) for a complete description, to Fig 1 for a step-by-step illustration of our approach and to Table 1 for a summary of model parameters and quantities. For computational expediency, we assume Gaussian probability distributions defined by mean and variance (or inverse precision).

Model simulation

To test whether our model is able to reproduce the temporal dynamics of bistable perception, we used it to generate perceptual time-courses from some ambiguous visual input such as the Lissajous figure. We assumed a sampling rate of 0.33 Hz, which was chosen to be close to the average overlap frequency in the behavioural experiment (see below), and simulated for a total of $6 * 10^5$ seconds. To model the ambiguous visual input, the impact of the stereodisparity weight was suppressed by setting $\mu_{stereo} = 0.5$ and $\pi_{stereo} = 0$. We further assumed fixed values for the precision π_{init}, which was set to 3.5 to match the posterior parameter value from our behavioural modelling (see *Modelling analysis of behavioural data*).

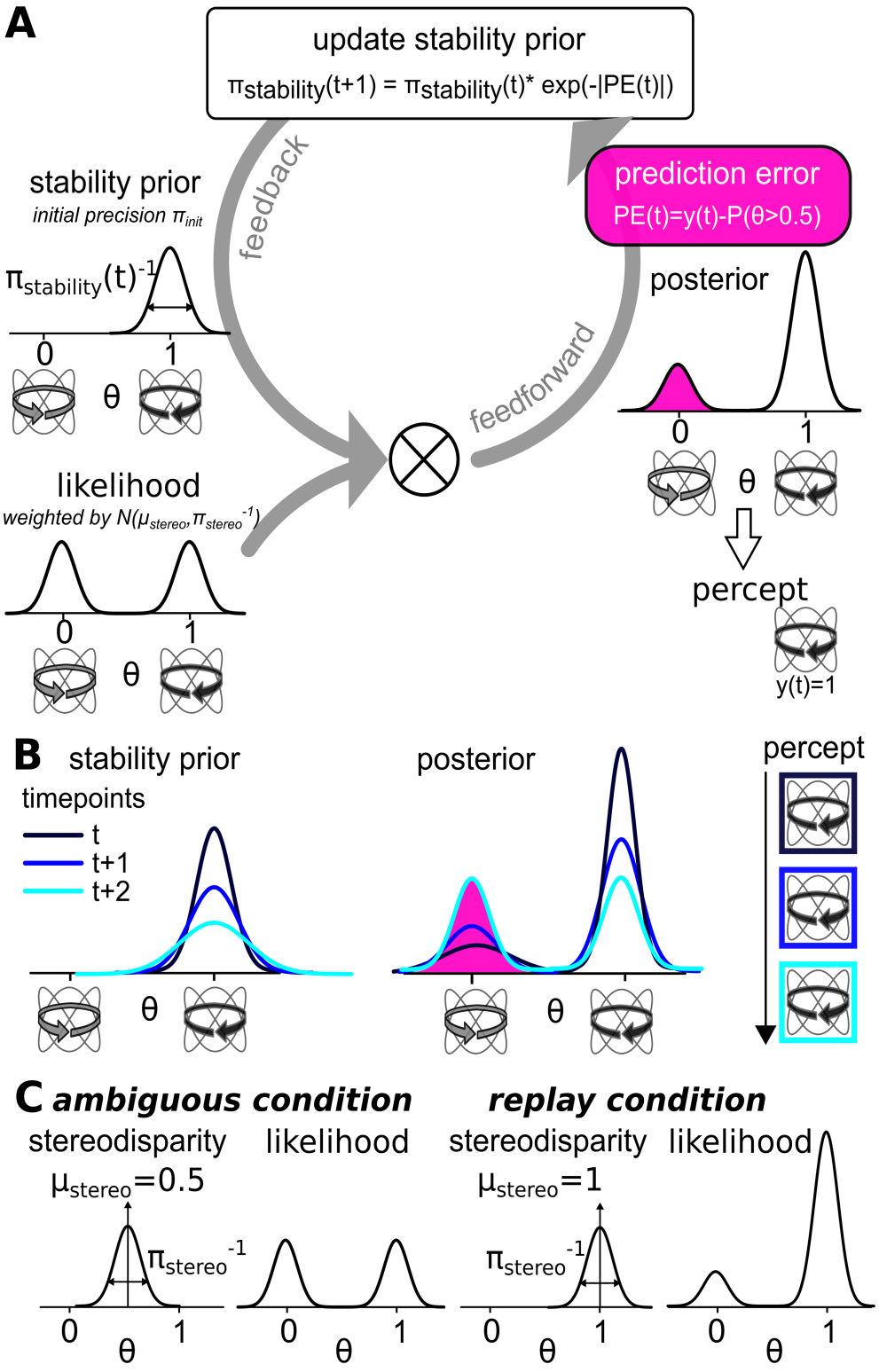


Fig 1. Modelling procedures. A. In the modelling approach illustrated here, we capture the temporal dynamics of bistable perception by changes in a continuously updated stability prior, which is combined with a bimodal likelihood representing the sensory input (see 'feedback' arrow). Under ambiguous viewing conditions, the likelihood contains equivalent evidence for both perceptual interpretations of the bistable stimulus. The mean of the prior 'perceptual stability' is defined by $\mu_{stability}$, which corresponds to the preceding

perceptual decision $y(t-1)$ (here centered around '1' for counter-clockwise rotation of the Lissajous figure). The impact of the prior on the bimodal likelihood is determined by its precision (the inverse of variance) $\pi_{stability}$. If a new perceptual decision was adopted at the preceding overlapping configuration of the Lissajous figure, this precision is set to π_{init}. Otherwise, $\pi_{stability}$ is repeatedly updated by a prediction error signal. This signal results from residual evidence for the alternative explanation of the bistable stimulus and is given by the difference between $P(\theta > 0.5)$ and the current perceptual decision $y(t)$ (see 'feedforward' arrow). In this example, the prediction error signal stems from remaining evidence for clockwise rotation (centered around '0'), as the current perceptual decision represents counter-clockwise rotation ($y(t) = 1$) of the stimulus. Overtime, the stability prior is weakened, which is accompanied by an increasing probability for a novel transition in perception. **B**. Here, we depict the temporal evolution of the stability prior (left panel) and the corresponding posterior (right panel) at three successive overlapping configurations of the Lissajous figure (dark to light blue). As the precision of the stabilizing prior is gradually reduced, the posterior relaxes to equivalent probability for both perceptual interpretations of the stimulus. This is accompanied by escalating prediction error signals and increased likelihood for a perceptual transition. **C**. Furthermore, our approach accounts for additional sensory evidence, which is realized by a stereodisparity signal and used to disambiguate the Lissajous figure in the 'replay' condition. To this end, we introduce a 'stereodisparity' distribution (characterized by mean μ_{stereo} and precision π_{stereo}), which serves as a weight on the bimodal likelihood. In the ambiguous condition (left panel), μ_{stereo} is centered around 0.5 and is thus uninformative with regard to the two perceptual interpretations of the stimulus. In the replay condition (right panel), μ_{stereo} is centered around '0' or '1' (depending on the direction of stereodisparity). The strength of the bias in the direction of either percept introduced by the stereodisparity signal scales with the precision π_{stereo}.

https://doi.org/10.1371/journal.pcbi.1005536.g001

Experimental procedures

To examine whether our prediction error model might account for bistable perception and associated neural activity in human observers, we used data from an fMRI experiment applying the Lissajous figure. Results from conventional analyses but not from behavioural modelling or model-based fMRI (see below) have been reported previously [7].

Participants. Twenty right-handed participants (11 female, mean age: 28, range: 21 -34) took part in this study, which was conducted at the Berlin Center for Advanced Neuroimaging (BCAN), Charité Universitätsmedizin Berlin, Campus Mitte. All participants had normal or corrected-to-normal vision, were naive to the purpose of the study, and provided informed written consent. The study was approved by the ethics committee of Charité Universitätsmedizin Berlin, Campus Mitte.

Stimulus. We presented stimuli generated with Psychophysics Toolbox 3 [23] running under Matlab 2007b (Mathworks inc.) on a 60 Hz Sanyo LCD projector, on which participants viewed alternating blocks of ambiguous and corresponding replay stimulation. In ambiguous blocks, we displayed two identical moving Lissajous figures formed by the intersection of two perpendicular sinusoids ($x(t) = sin(3t)$ and $y(t) = sin(6t + \delta)$; with δ increasing from 0 to 2π),

Table 1. Summary of model parameters and quantities.

	Name	Explanation
Sensory Stimulation	μ_{stereo}	Mean of sensory stimulation
Responses	y	Binary perceptual decision
Model Parameters	π_{init}	Initial precision of stability prior
	π_{stereo}	Initial precision of stability prior
	ζ	Inverse decision temperature of the reponse model
Model Quantities	$y_{predicted}$	Predicted perceptual response
	$\mu_{stability}$	Mean of the stability prior
	$\pi_{stability}$	Precision of the stability prior
	μ_m	Mean of the joint prior
	π_m	Precision of the joint prior
	$P(\theta > 0.5)$	Probability of perceiving counter-clockwise rotation

https://doi.org/10.1371/journal.pcbi.1005536.t001

separately to the two eyes. In replay blocks, a disambiguated version of the Lissajous figure mimicked the perceptual time-course participants had experienced during the preceding ambiguous block. To this end, the two dichoptically presented Lissajous figures were phase-shifted against each other by an offset of 0.04˚. This disparity cue was used to disambiguate the stimulus, biasing participants perceived direction of rotation in the direction of the phase shift. All stimuli subtended 2.05˚ visual angle.

We achieved dichoptic stimulation by placing a custom build cardboard divider between the mirror attached to the head-coil and the screen at the end of the scanners bore [24]. Participants wore prism glasses to facilitate fusion between to two eyes. All screens contained a fixation mark at the center and fusion frames surrounding the stimuli.

Task. Participants were instructed to indicate the perceived direction of rotation of the Lissajous figure by pressing a left (index finger; for clockwise rotation of the stimulus, i.e. movement of the front surface to the left) or right (ring finger) button with their right hand, responding to the first perceived direction after stimulation onset and to all additional perceptual transitions. Furthermore, they reported unclear or mixed percepts by pressing a middle button (middle finger) on a standard MRI button box.

In order to titrate individual percept durations to approximately 10 s, we adjusted the rotational speed of the stimulus for every participant to one of three levels ('overlap' frequency 0.24, 0.30, and 0.40 Hz) based on a psychophysical experiment prior to the fMRI session. In the fMRI experiment, participants were presented with three experimental runs, each containing 8 pairs of ambiguity and replay separated by 10 s fixation. Block duration amounted to 42.8, 40.90, or 41 s, depending on the individually adjusted speed. After completion of the fMRI experiment, participants answered a debriefing questionnaire (A: *Did you have the impression that some blocks were different from others?* B: *Did you perceive the transitions as instantaneous or prolonged?* C: *Were you able to tell the direction of rotation of the Lissajous figure at all times during the experiment?*).

fMRI acquisition and preprocessing

We recorded BOLD images by T2-weighted gradient-echo echo-planar imaging (FOV 192, 33 slices, TR 2000 ms, TE 30 ms, flip angle 78˚, voxel size 3 x 3 x 3 mm, interslice gap 10 percent) on a 3T MRI scanner (Tim Trio, Siemens). The number of volumes amounted to 402 (0.15 Hz and 0.2 Hz) or 415 (0.12 Hz) volumes, respectively. We used a T1-weighted MPRAGE sequence (FOV 256, 160 slices, TR 1900 ms, TE 2.52 ms, flip angle 9˚, voxel size 1 x 1 x 1 mm) to acquire anatomical images.

Image preprocessing (standard realignment, coregistration, normalization to MNI stereotactic space using unified segmentation, spatial smoothing with 8 mm full-width at half-maximum isotropic Gaussian kernel) was carried out with SPM8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8).

Modelling analysis of behavioural data

To probe whether our predictive coding model might explain perceptual time-courses during bistable perception in human observers, we fitted our model to the behavioural data collected during the fMRI experiment. We optimized our model for the prediction of perceptual outcomes, i.e. on the perception of clockwise or counter-clockwise rotation as indicated by the individual participants. To this end, participants' responses were aligned to the overlapping stimulus configurations of the Lissajous figure ('overlaps'). This refers to timepoints during presentation when fore- and background of the stimulus cannot be discerned (i.e. depth-symmetry) [25, 26]. Depending on the rotational speed of the stimulus and the associated 'overlap'

frequency, sampling rates varied across participants between 0.24 Hz and 0.40 Hz (see above). We first constructed models incorporating all combinations of the likelihood weight 'stereodisparity' and prior 'perceptual stability', yielding a total of 4 behavioural models (behavioural model 1: no stereodisparity, no perceptual stability; behavioural model 2: no stereodisparity, perceptual stability; behavioural model 3: stereodisparity, no perceptual stability; behavioural model 4: stereodisparity, perceptual stability) to be compared. The respective precision of these distributions was optimized for the prediction of perceptual outcomes based on posterior distributions using a free energy minimization approach [27]. This method minimises the surprise about the individual participants' data, thereby maximising log-model evidence.

For model inversion, precisions were modelled as log-normal distributions. π_{init} and π_{stereo} were either estimated as free parameters (π_{init}: prior mean of log(3) and prior variance of 5; π_{stereo}: prior mean of log(5) and prior variance of 5) or fixed to zero (thereby effectively removing the distribution from the model). We kept ζ, which represents the inverse decision temperature in the response model represented by Equation 11 (see Mathematical Appendix, S1 Text), fixed to 1, since we did not have a particular a-priori hypothesis regarding this parameter. Please note that when choosing ζ as a free parameter (prior mean of log(1), prior variance of 1), results remained almost identical. Parameters were optimised using quasi-Newton Broyden-Fletcher-Goldfarb-Shanno minimisation as implemented in the HGF4.0 toolbox (TAPAS toolbox, http://www.translationalneuromodeling.org/hgf-toolbox-v3-0/).

After identifying the optimal model using Random Effects Bayesian model selection [20], as implemented in SPM12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/), we analyzed its posterior parameters with regard to the respective precision of the prior distributions using classical frequentist statistics. Since parameters were estimated in log-space, we report the geometric mean (i.e. the arithmetic mean in log-space).

In a supplementary analysis (see S2 Text), we further compared the explanatory power of our predictive coding model with established models of bistable perception. To this end, we implemented models of bistable perception belonging to three different classes ([1] as an example of so-called oscillator models based on mutual inhibition and self-adaptation between two competing neuronal populations, [2] as a representative of noise-driven attractor models and [3] as an intermediate model), which can be fitted to experimental data. We conducted a Random Effects Bayesian Model Comparison [20] between the established models and our predictive coding model in order to probe the validity of our approach.

Model-based fMRI data analysis

To examine the neural correlates of prediction error time-courses from our model, we conducted model-based fMRI analyses [21] in SPM12. We adopted a general-linear-model-(GLM-) approach, constructing a total of three models:

The design matrix of the first GLM (the 'PE model') represented prediction error trajectories timepoint by timepoint. To this end, the regressor 'transitions' and the regressor 'overlaps' were modelled as stick functions. Furthermore, we extracted the individual 'Prediction Error' time-course for every participant and run and used its absolute value as a parametric modulator for the regressor 'overlaps'.

In order to enable a comparison to the conventional approach of analysing fMRI data on bistable perception, we constructed a second GLM that dissociated between transition-related activity specific to bistable perception and the replay condition [4, 5, 6, 7, 9, 10]. In addition to the regressor 'overlaps', the design matrix of this 'Conventional model' contained ambiguous and replay transitions represented by stick functions.

To further investigate the specificity of the prediction error trajectories and their neural correlates, we constructed a third GLM that took into account the presence of ambiguity inherent to the bistable condition. The design matrix contained the regressors 'transitions' as well the regressor 'overlaps' modelled as stick functions. Here, however, we used a box-car function being 1 for ambiguous and 0 for 'replay' blocks as a parametric modulator of the regressor 'overlaps'. Hence, this 'Block model' only differs from the 'PE model' in the values of the parametric modulator and serves to investigate whether correlations with the prediction error (which we assumed to be higher in the bistable condition) merely correspond to ambiguity per se.

All further analyses were conducted for all models in parallel: regressors were convolved with the canonical hemodynamic response function as implemented in SPM12. We added six rigid-body realignment parameters as nuisance covariates and applied high-pass filtering at 1/128 Hz.

In a first step, we tested which of the three models accounted best for the measured BOLD signal. Therefore, we conducted a voxel-wise model comparison of the 'PE model' with the 'Conventional model' and the 'Block model', as described in [22]. In brief, this technique uses Bayesian statistics for the construction of 'Posterior Probability Maps' (PPMs) and 'Exceedance Probability Maps' (EPMs), which enable the calculation of log-evidence maps for each participant and model separately. On a second level, these log-evidence maps can be combined, thereby enabling voxel-wise model inference at the group level. Using the 'Bayesian 1st level' procedure for model estimation, we constructed log-evidence maps for every participant and model separately and compared the 'PE model' to the other models on a group level using exceedance probabilities computed with Random Effects analyses.

In a second step, we aimed to identify regions in which prediction error trajectories ('PE model'), ambiguity per se ('Block model') or ambiguous as compared to replay transitions ('Conventional model') were correlated with the recorded BOLD signals. To this end, we estimated single-participant statistical parametric maps, then created contrast images for the parametric regressor against baseline ('PE model' and 'Block model') or ambiguous against replay transitions ('Conventional model'). These were entered into voxel-wise one-sample t-tests at the group level. Voxels were considered statistically significant if they survived family-wise-error (FWE) correction for all voxels in the brain at $p < 0.05$. Anatomic labeling of cluster peaks was performed using the SPM Anatomy Toolbox Version 1.7b [28].

In order to further visualize our results, we extracted eigenvariate time-courses (without adjustment for effects of interest) from spherical ROIs (radius: 3 mm) around peak voxels from clusters for the contrast 'Prediction Error vs baseline' (thresholded at $p < 0.05$) corresponding to left IFG (peak voxel: [-54 2 22]), right IFG (peak voxel: [51 8 10]), left insula (peak voxel: [-30 20 10]) and right insula (peak voxel: [33 23 7]). These time-courses were extracted for ambiguous stimulation only. The time-courses for all perceptual phases were aligned with the respect to the end of the perceptual phase and averaged within and across observers.

Results

Model simulation

To test whether our predictive coding model was able to reproduce perceptual switching in bistable perception, we used the model to generate perceptual time-courses during simulated viewing of an ambiguous stimulus.

The distribution of perceptual phase durations followed a sharp rise and slow fall (Fig 2) typical for bistable stimuli [29, 30]. Mean and median simulated phase durations were 10.40 and 10.00 seconds, closely matching the results from behavioural analysis (see *Modelling*

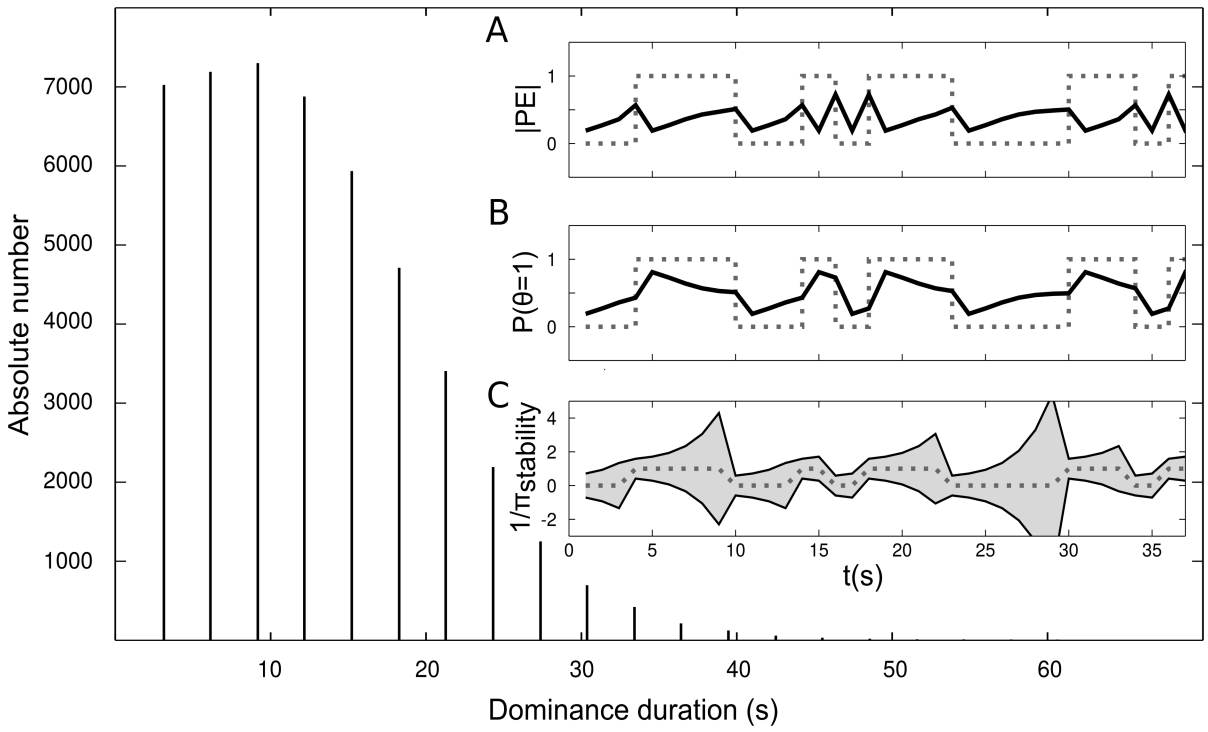


Fig 2. Simulating perceptual decisions during ambiguous stimulation. Data were simulated using π_{init} of 3.5 at a sampling rate of 0.33 Hz for a total of $6*10^5$ seconds. The distribution of phase durations followed a sharp rise and slow fall resembling a gamma-distribution. The insets A-C show simulated perceptual time-courses (grey dotted lines) next to updated model quantities (black solid lines). **A**: Prediction errors increase during a dominance phase and are reduced by perceptual transitions. **B**: Bistable perception can be conceived as resulting from subsequent sampling from a bimodal probability distribution [19], the weight of which is expressed by $P(\theta > 0.5)$. This weight is close to 0 or 1 at the beginning of a dominance phase (low transition probability) and gradually relaxes to 0.5 (high transition probability). **C**: The variance (inverse precision) of the prior distribution 'perceptual stability' increases as a consequence of prediction errors and is set to $1/\pi_{init}$ after a transition in perception.

https://doi.org/10.1371/journal.pcbi.1005536.g002

analysis of behavioural data). As illustrated by exemplary time-courses of model parameters, the prediction error *PE* (Fig 2A) increases over time while one percept is dominant and is reduced once a new percept is adopted, reflecting the accumulation of evidence from the suppressed percept. The variance ($1/\pi_{stability}$) of the prior 'perceptual stability' (Fig 2C) increases over a perceptual phase as a function of the prediction error. In line with the hypothesized role of prediction errors in driving perceptual transitions, the prediction error *PE* and, hence, the variance $1/\pi_{stability}$ are maximal when the posterior $P(\theta > 0.5)$ relaxes to 0.5 (Fig 2B), thereby increasing the probability of a new perceptual transition.

Modelling analysis of behavioural data

To investigate whether our model is able to explain the dynamics of perceptual bistability in human observers, we fitted our model to behavioural data collected from 20 healthy participants during an fMRI experiment, in which participants viewed ambiguous and unambiguous (replay) versions of a rotating Lissajous stimulus. As reported previously, perceptual transitions occurred on average every 9.3 seconds in the ambiguity condition and neither block-by-block ratings nor debriefing after the experiment revealed differences in perceived appearance between the ambiguity and the replay condition [7].

We first performed a model comparison with other models that lacked the key conceptual elements of our model. By eliminating either the likelihood weight 'stereodisparity' or the

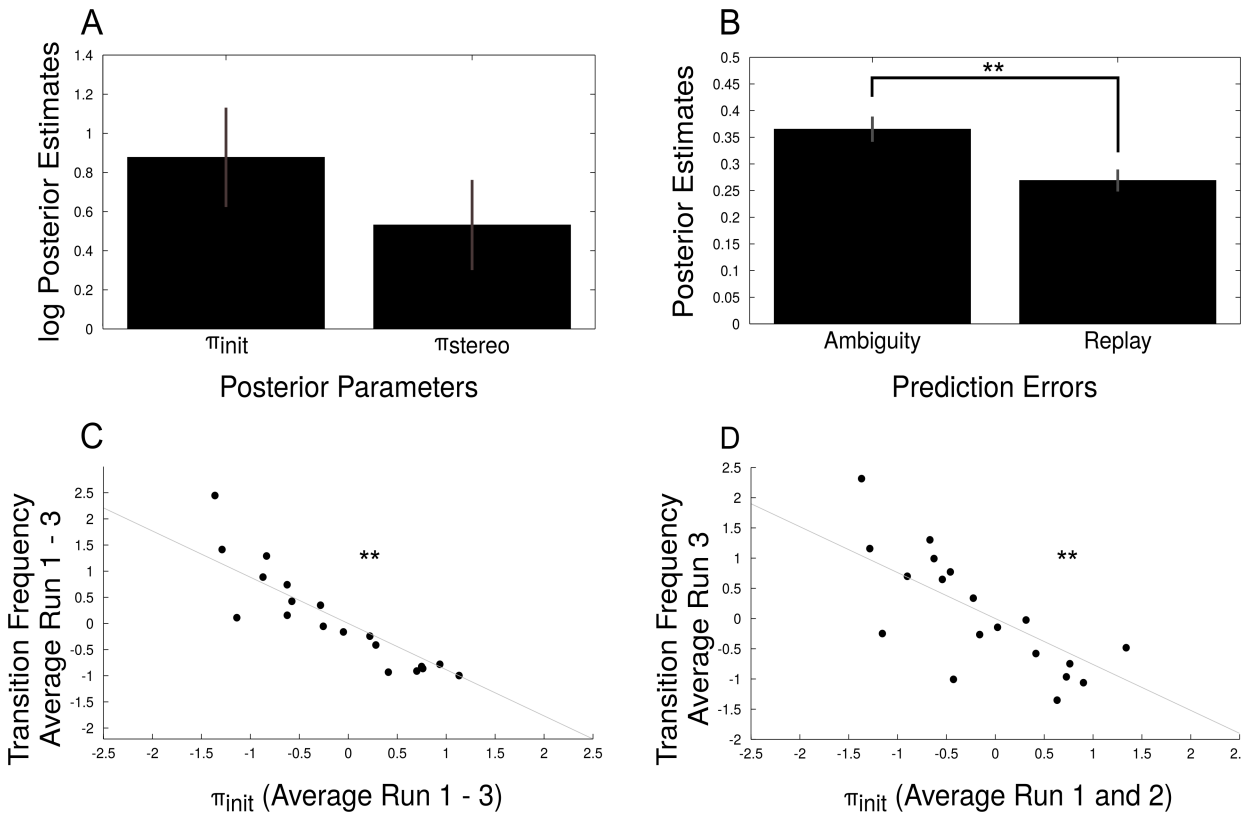


Fig 3. Posterior model parameters. A: The geometric mean (i.e. the arithmetic mean in log-space) of posterior π_{init} and π_{stereo}, averaged across runs and participants, and standard error of the mean. **B**: Mean prediction errors averaged across runs and participants for ambiguous and replay blocks and standard error of the mean. Prediction errors were significantly decreased during replay stimulation (two-sample t-test, $p < 10^{-6}$, $t_{19} = 7.69$). **C**: Average transition probabilities correlated significantly with average π_{init} for individual participants ($\rho = -0.88$, $p < 10^{-7}$, Pearson correlation), providing a sanity check for model fit. **D**: Transition probabilities from run 3 were predictive of posterior π_{init} averaged over run 1 and 2. The significant Pearson correlation between the two independent measures ($\rho = -0.76$, $p < 10^{-4}$) illustrates the predictive power of the model.

https://doi.org/10.1371/journal.pcbi.1005536.g003

prior 'perceptual stability' or both from the model, we constructed three additional models which we compared to our model using Random Effects Bayesian Model Selection. Our model (i.e. behavioural model 4) was identified as a clear winning model with a protected exceedance probability of 99.96%, demonstrating that the incorporation of both the likelihood weight 'stereodisparity' and the prior 'perceptual stability' best explained participants' perception.

From this model, we extracted the parameters for π_{init} and π_{stereo} and averaged across runs and participants (Fig 3A). We predicted average prediction errors to be lower in replay as compared to the ambiguous condition, since the presented stereodisparity reduces the ambiguity left in the experimental display, and hence, the residual evidence for suppressed percept. Consistently, mean prediction errors were significantly higher in the ambiguous condition than in the replay condition (0.36 +/- 0.03 vs. 0.26 +/- 0.02, mean +/- s.e.m., $p < 10^{-6}$, $t_{19} = 7.06$, two-sample t-test, Fig 3B), providing support for a correct implementation of our predictive coding model.

Given that π_{init} describes the strength of the initial stabilization after a switch in perception, we expected this parameter to be related to the frequency of perceptual transitions. In line with this, model parameter estimates π_{init} were negatively correlated with perceptual transition frequencies across participants ($\rho = -0.88$, $p < 10^{-7}$, Pearson correlation, Fig 3C), providing a

sanity check for model fit. Notably, this correlation was also significant when we correlated model parameter estimates for π_{init} averaged over run 1 and 2 with perceptual transition frequencies from run 3 ($\rho = -0.76$, $p = 10^{-4}$, Fig 3D), corroborating that our model successfully accounted for observers' perception evoked by an ambiguous stimulus.

We furthermore validated our approach by comparing our predictive coding model to established models of bistable perception from three different classes: oscillator models [1], attractor models [2] and intermediate models [3] (see Supplementary Methods in S2 Text). Data simulations indicated that all established models, similar to our predictive coding model, were able to produce spontaneous transitions in perception and a typical gamma-like distribution of perceptual phase durations (see Supplementary Results and Fig. A-C in S2 Text). Fitting of the behavioural data further showed that both the oscillator and the intermediate, similar to our predictive coding model, adequately accounted for the observers' perceptual decisions during bistable perception (see Supplementary Results and Fig. D-I in S2 Text). In order to validate our approach, we conducted a Bayesian Model Comparison, which showed that our predictive coding model compared to these established models was best in explaining the behavioural data collected during this experiment (see Fig. J in S2 Text).

Please note that we did not carry out these analysis to demonstrate a superiority of our approach over these earlier models, which were initially conceived mainly for binocular rivalry and not for the prediction of behavioural responses during presentation of the Lissajous figure (a specific type of structure-from-motion stimulus). On the contrary, we aimed at probing the validity of our approach and tried to ascertain that the predictive coding approach was at least equivalent to existing models of bistable perception.

Model-based fMRI analysis

One central aim of our study was to gain mechanistic insight into the neural processes underlying transition-related activity during bistable perception. We therefore performed both a model-based fMRI analysis suitable to identify the neural correlates of modelled prediction errors ('PE model'), and, for the purpose of comparison, a conventional analysis ('Conventional model') dissociating between ambiguous and replay transitions as well as a 'Block model' accounting for effects of ambiguity per se.

To test the validity of these models, we first searched for voxels that were more active during visual stimulation as compared to baseline ('overlaps vs. baseline'). For the 'PE model', this analysis revealed significant clusters ($p < 0.05$, FWE-corrected across the whole brain) bilaterally in middle occipital cortex (right: [39 -9 1], T = 10.21; left: [-30 -94 1], T = 13.30), in V5/hMT+ (right: [45 -70 1], T = 11.61; left: [-45 -73 4], T = 14.09), as well as in superior parietal cortex (right: [27 -49 58], T = 10.26; left: [-36 -46 -61], T = 8.62). The same analyses for the 'Conventional model' and the 'Block model' yielded virtually identical results (see Tables 2–4), confirming the comparability between all three models.

We then investigated which voxels were more active during perceptual transitions as compared to baseline ('transitions vs. baseline', Fig 4A): For the 'PE model', we found significant activations of motor-related areas in left precentral gyrus ([-36 -16 67], T = 12.23) extending to left postcentral gyrus ([-63 -19 25], T = 8.62) as well as significant clusters in regions previously associated with transition-related activity during bistable perception: right inferior frontal gyrus ([54 17 13], T = 7.96), right inferior parietal lobulus (54 -37 52, T = 9.32) and right middle frontal gyrus ([39 44 31], T = 7.57). Additional clusters were located in bilateral posterior-medial frontal gyrus (right: [6 2 67], T = 9.50; left: [-6 2 55], T = 12.63). Again, repeating this analysis for the 'Block model' and the 'Conventional model' yielded qualitatively very similar

Table 2. 'PE model': Overlaps vs baseline.

Cluster	T	MNI			Region
1	T = 11.61	45	-70	1	R Middle Temporal Gyrus
	T = 10.94	30	-91	-5	R Inferior Occipital Gyrus
	T = 10.21	39	-79	1	R Middle Occipital Gyrus
2	T = 10.26	27	-49	58	R Superior Parietal Lobule
	T = 0.22	30	-46	55	R Postcentral Gyrus
	T = 8.93	21	-58	58	R Superior Parietal Lobule
3	T = 11.96	-27	-52	55	L Inferior Parietal Lobule
	T = 8.62	-36	-46	61	L Superior Parietal Lobule

https://doi.org/10.1371/journal.pcbi.1005536.t002

Table 3. 'Conventional model': Overlaps vs baseline.

Cluster	T	MNI			Region
1	T = 11.64	42	-70	-2	R Middle Temporal Gyrus
	T = 10.92	30	-91	-5	R Inferior Occipital Gyrus
	T = 10.22	39	-79	1	R Middle Occipital Gyrus
2	T = 10.17	27	-52	61	R Superior Parietal Lobule
	T = 10.09	30	-49	58	R Superior Parietal Lobule
	T = 8.90	21	-58	58	R Superior Parietal Lobule
3	T = 11.82	-27	-52	55	L Inferior Parietal Lobule
	T = 8.35	-36	-46	61	L Superior Parietal Lobule

https://doi.org/10.1371/journal.pcbi.1005536.t003

results as in the 'PE model' (see Tables 5–7), thereby providing further evidence for the validity and comparability of all three models.

To formally test whether the modelled prediction error explains the BOLD signal better than the conventional comparison of ambiguous with replay perceptual switches ('Conventional model'), or the mere ambiguity of the visual display (the 'Block model'), we performed a PPM analysis [22] to compute voxel-wise exceedance probability maps for the 'PE model' against the 'Conventional model' and the 'Block model' (Fig 4C). We restricted this analysis to areas of the fronto-parietal cortex, which be delineated by intersecting the statistical-parametric maps for 'transitions vs. baseline' thresholded at $p < 0.05$ FWE for all three models considered. Remarkably, when applying a conservative threshold of an exceedance probability of $\gamma = 99\%$ and a cluster size of $n > 10$ voxels, we found clusters in right insula ([39 26 -2]) and right inferior frontal gyrus ([51 14 1]) to show strong evidence for the 'PE model' as compared

Table 4. 'Block model': Overlaps vs baseline.

Cluster	T	MNI			Region
1	T = 11.60	42	-70	-2	R Middle Temporal Gyrus
	T = 10.96	30	-91	-5	R Inferior Occipital Gyrus
	T = 10.26	39	-79	1	R Middle Occipital Gyrus
2	T = 10.19	27	-52	61	R Superior Parietal Lobule
	T = 10.10	30	-46	55	R Postcentral Gyrus
	T = 8.89	21	-58	58	R Superior Parietal Lobule
3	T = 11.78	-27	-52	55	L Inferior Parietal Lobule
	T = 8.36	-36	-46	61	L Superior Parietal Lobule

https://doi.org/10.1371/journal.pcbi.1005536.t004

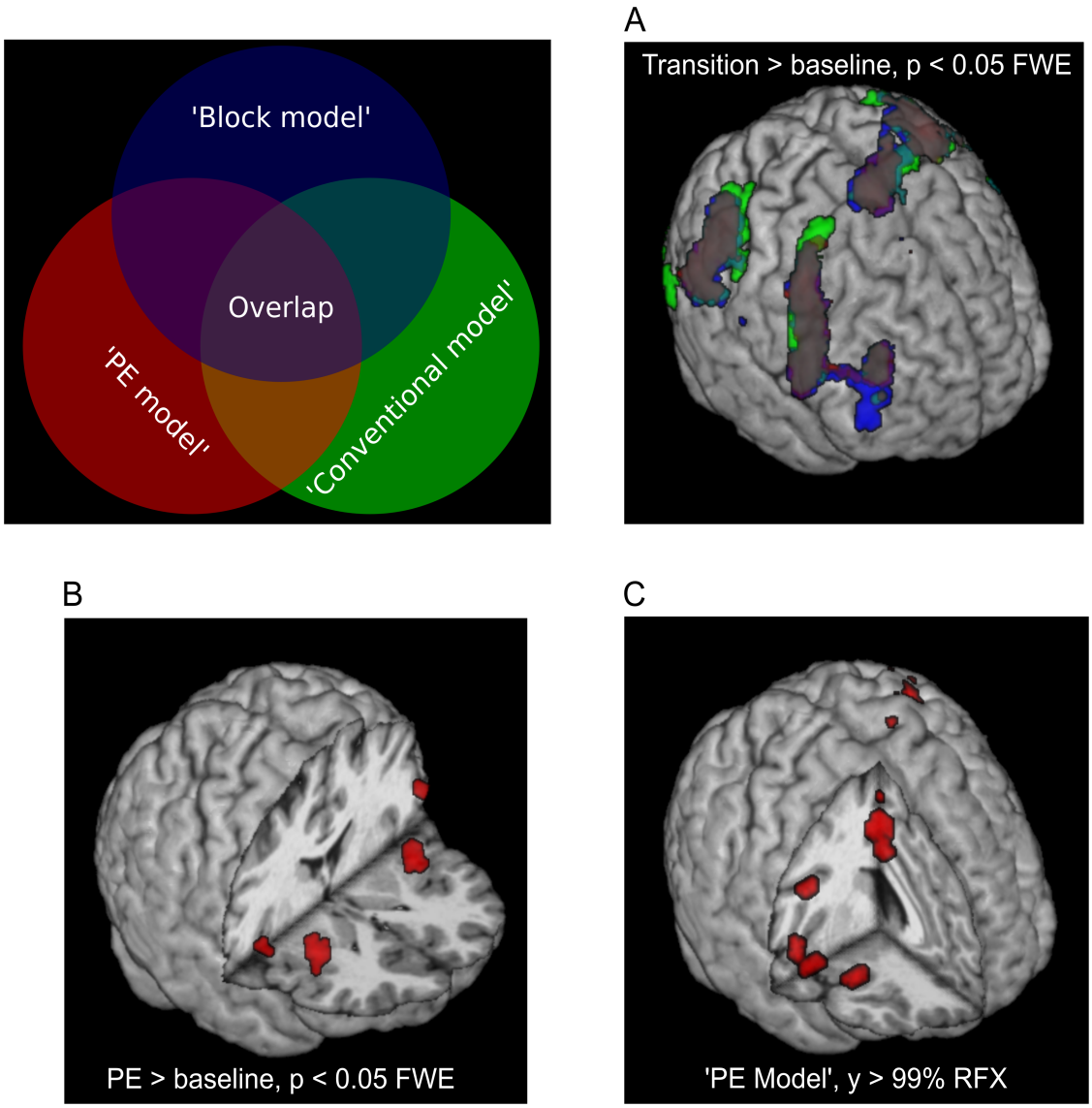


Fig 4. Model-based fMRI results from standard GLM (A, B) and PPM (C) analyses. GLMs are displayed using FWE correction at $p < 0.05$. For PPM results, we show voxels above an exceedance probability of 99% with a cluster size $n > 10$. **A**: 2-level contrast for 'Transition vs. baseline' showing activations left pre- and postcentral gyrus, right inferior frontal gyrus, right inferior parietal lobulus and right middle frontal gyrus with qualitatively equivalent results for all models. **B**: 'PE vs. baseline' ('PE model') yielded activations in bilateral insulae and inferior frontal gyri. We found no whole-brain correctable voxels for 'Ambiguity vs. baseline' ('Block model') nor for 'Ambiguous vs. replay transitions' ('Conventional model'). **C**: Group exceedance probability maps with right insula, right inferior frontal gyrus, right posterior-medial frontal gyrus as well as left precentral gyrus showed strongest evidence for the 'PE model' as compared to the control models.

https://doi.org/10.1371/journal.pcbi.1005536.g004

to the 'Block model' and the 'Conventional model'. Additional clusters were located in right posterior medial frontal gyrus ([6 5 49]) as well as left precentral gyrus ([-36 -16 52]).

Conversely, for the exceedance probability map of the 'Conventional model' compared against 'Block model' and 'PE model', no voxels survived the conservative threshold used in the main experiment. For the exceedance probability map of the 'Block model' compared against the 'Conventional model' and 'PE model', we found clusters in bilateral inferior parietal lobule at an exceedance probability of 99% and a cluster size > 10.

Table 5. 'PE model': Transitions vs baseline.

Cluster	T	MNI			Region
1	T = 12.23	-36	-16	67	L Precentral Gyrus
	T = 8.74	-51	-28	43	L Inferior Parietal Lobule
	T = 8.28	-57	-28	34	L SupraMarginal Gyrus
2	T = 11.19	42	2	46	R Precentral Gyrus
	T = 9.73	42	8	40	R Middle Frontal Gyrus
	T = 7.71	15	5	13	R Caudate Nucleus
	T = 7.96	54	17	13	R IFG (p. Opercularis)
3	T = 12.63	-6	2	55	L Posterior-Medial Frontal
	T = 9.50	6	2	67	R Posterior-Medial Frontal
4	T = 9.42	60	-40	43	R SupraMarginal Gyrus
5	T = 6.70	42	26	-5	R Insula
6	T = 7.03	-18	-97	-8	L Inferior Occipital Gyrus
7	T = 6.50	-27	-88	-2	L Middle Occipital Gyrus

https://doi.org/10.1371/journal.pcbi.1005536.t005

Table 6. 'Conventional model': Transitions vs baseline.

Cluster	T	MNI			Region
1	T = 14.73	-6	-1	55	L Posterior-Medial Frontal
	T = 12.88	-36	-16	67	L Precentral Gyrus
	T = 10.33	-42	-7	4	L Insula Lobe
2	T = 10.11	60	-40	43	R SupraMarginal Gyrus
	T = 9.57	51	-40	55	R Inferior Parietal Lobule
3	T = 7.97	42	44	28	R Middle Frontal Gyrus
4	T = 7.30	39	-46	40	R Inferior Parietal Lobule
5	T = 6.80	-21	-94	-8	L Inferior Occipital Gyrus
6	T = 6.82	33	-58	43	R Angular Gyrus

https://doi.org/10.1371/journal.pcbi.1005536.t006

For our central analysis aimed at identifying the neural correlates of modelled prediction errors, we searched for voxels in which BOLD activity was related to the parametric modulator of the 'PE model' that encoded prediction error trajectories from our Bayesian model of bistable perception (Fig 4B). We found significant clusters ($p < 0.05$, FWE-corrected across the whole brain) in bilateral insulae (right: [33 23 7], T = 7.24; left: [-30 20 10], T = 7.88) and

Table 7. 'Block model': Transitions vs baseline.

Cluster	T	MNI			Region
1	T = 14.71	-6	-1	55	L Posterior-Medial Frontal
	T = 12.58	-42	-22	58	L Postcentral Gyrus
	T = 12.57	-36	-16	67	L Precentral Gyrus
	T = 10.51	-42	-7	4	L Insula Lobe
	T = 9.65	42	8	37	R Middle Frontal Gyrus
2	T = 10.17	60	-40	43	R SupraMarginal Gyrus
	T = 9.57	51	-40	55	R Inferior Parietal Lobule
3	T = 6.89	33	-61	43	R Angular Gyrus
4	T = 6.71	-18	-97	-8	L Inferior Occipital Gyrus

https://doi.org/10.1371/journal.pcbi.1005536.t007

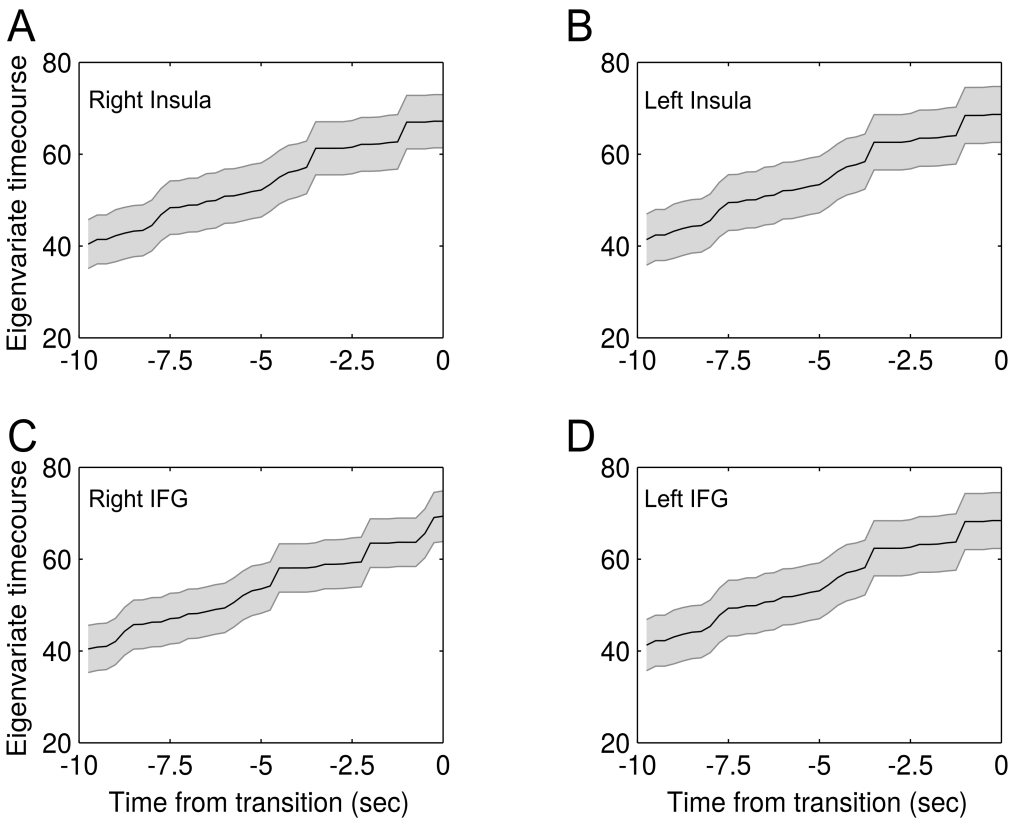


Fig 5. Average eigenvariate time-courses. For visualization, we extracted eigenvariate time-courses from right insula, left insula, right IFG and left IFG (A–D), aligned all phase durations to the timepoint of the upcoming perceptual transition and averaged within and across observers. In analogy to modelled prediction error trajectories, mean eigenvariate time-courses (± standard error of the mean) showed a gradual increase towards a transition in perception.

https://doi.org/10.1371/journal.pcbi.1005536.g005

bilateral inferior frontal gyri (right: [51 8 10], T = 6.89; left: [- 54 2 22], T = 6.67). These regions are located in close anatomical proximity to frontal regions previously suggested to mediate perceptual transitions in bistable perception [4, 5, 7].

In order to further visualize the correlation between modelled prediction error and BOLD activity, we extracted eigenvariate time-courses from right insula, left insula, right IFG as well as left IFG and averaged across perceptual phase durations and observers. As expected, these time-courses showed a gradual increase towards a transition in perception (Fig 5), nicely mirroring the build-up of prediction error during a perceptual phase.

Discussion

In this work, we present a Bayesian predictive coding model for bistable perception, which rests on the basic assumption that prediction errors are elicited by the unexplained alternative interpretation of an ambiguous stimulus and represent the driving force behind perceptual transitions during bistable perception. We found that this model is able to reproduce key temporal characteristics of human bistable perception and that it explains observers' behaviour during a bistable perception experiment. Our central finding shows that modelled prediction errors correlate with BOLD activity in bilateral insulae and bilateral inferior frontal gyrus. Remarkably, our PPM analysis revealed that modelled prediction errors best accounted for

BOLD activity as compared to mere occurrence of endogenous perceptual transitions or ambiguity of the visual display in these frontal regions. Hence, our current results suggest that prediction errors might provide the mechanistic basis for perceptual switching in bistable perception and offer a novel interpretation of frontal activity in bilateral insulae as well as the right inferior frontal gyrus during bistable perception.

The functional significance of enhanced frontal brain activity for transitions during bistability as compared to an unambiguous control condition is a matter of ongoing debate: Some authors proposed that non-sensory higher-level brain regions are actively implicated in resolving the perceptual conflict during bistable perception, thus mediating perceptual transitions [4, 31, 5, 11, 7]. Others have argued that perceptual conflicts are resolved primarily in sensory brain areas and that activity in frontal and parietal regions reflects the registration and/or report of perceptual transitions, rather than their cause [6, 8, 10]. For a detailed discussion of this debate, see "Brascamp, Sterzer, Blake and Knapen, Multistable perception and the role of frontoparietal cortex in perceptual inference, Annual Review of Psychology, in press."

Here, we provide further evidence for an active implication of frontal regions in bistable perception by functionally relating these regions to a prediction error signal. Hence, our work is in line with hybrid models that suggest bistable perception to arise from an interplay between lower-level sensory and higher-level non-sensory areas [32, 12, 11]. In this context, it might be speculated that prediction errors are computed in frontal regions based on feedforward signals from visual and parietal cortex; and that these prediction errors, in turn, modulate activity in visual cortex via feedback signals.

In addition to the prediction error, the stability prior represents an essential element of our predictive coding model of bistable perception, since its initial precision determines the frequency of perceptual transitions. The notion of such a stability prior is supported by experimental work on serial dependence in visual perception: In an orientation-judgement task, [33] showed that perceived orientation was biased by recently observed stimuli and reasoned that the visual system might use past experiences as predictors of present perceptual decisions, thereby incorporating representations of the continuity of the visual environment. Corroborating these results in a fMRI experiment, [34] found that orientation signals in early visual cortex were biased towards previous perceptual decisions. At this point, however, we can only speculate about the neural correlates of the stability prior from our model: In recent work on the role of parietal cortex in bistable perception, [35] and [9] have proposed a functional segregation of the superior parietal lobulus (SPL), which they deduced from differential effects of grey matter volume on perceptual dominance durations and analyses of effective connectivity on the basis of fMRI. By interpreting their results in a Bayesian framework, the authors argued that posterior SPL might represent a prediction error, while the anterior SPL would entertain a perceptual prediction.

A key advantage of our predictive coding model of bistable perception is that it allows us to treat ambiguous and replay stimulation within the same framework. By formalizing the disambiguating factor as a weight on a bimodal likelihood distribution, such models can be used to investigate perceptual transitions under varying degrees of ambiguity, thus dissolving the artificial dichotomy between the two conditions. Hence, such models provide a new perspective on how the brain might resolve perceptual conflicts despite the ambiguity inherent in every sensory signal and offer a generic tool for quantifying the contribution of different contextual factors on perceptual outcomes.

The major strength of predictive coding models for bistable perception, however, lies in their ability to parsimoniously link different levels in the description of perceptual dynamics in ambiguous visual environments: On a computational level, prediction errors constitute the driving force behind perceptual transitions and are substantially reduced by additional sensory

information (such as stereodisparity) during replay. On a neural level, casting frontal activity during rivalry in terms of prediction error signals nicely relates to increased transition-related activity [4, 5, 9] and connectivity [7]. On a theoretical level, viewing perceptual transitions as means of reducing prediction errors places bistable perception in the context of Bayesian theories of the brain [16, 36, 27, 37], and in particular the free-energy principle [13]. According to the latter, agents strive for a reduction of their model's free energy, which translates onto a minimization of squared prediction errors in predictive coding schemes. When sensory information is constantly ambiguous, one possibility to reduce free energy is to update beliefs about the world, which ultimately corresponds to the adoption of a new percept.

However, given that the Lissajous differs in some aspects from other types of bistable stimuli, one has to consider important limitations regarding the generalization of our findings: While being physically ambiguous for all angles of rotation, transitions almost exclusively occur at overlapping stimulus configurations, which is similar to the behaviour of some types of random dot kinematograms [26] or intermittent presentation of bistable stimuli [38] and accompanied by a reduced incidence of mixed percepts or incomplete transitions. Since these phenomena are present in many other forms of bistable perception and significantly affect frontoparietal activity during perceptual transitions [6], our current imaging results can only be interpreted in relation to the specific stimulus used here.

A similar limitation applies to the behavioural modelling presented in this manuscript: Previous work on computational modelling of bistable perception has focused on a variety of mechanisms at the heart of spontaneous perceptual transitions: Oscillator models have focused on mutual inhibition between two competing neuronal populations combined with slow adaptation of the currently dominant population [1]. [39] have studied the differential effects of short and long interruptions in intermittent bistable perception for binocular rivalry and structure-from-motion and presented a model based on adaptive processes, cross-inhibition and neural baseline levels. Importantly, this model also accounts for the possibility of voluntary control via attentional processes interacting with early processing stages.

Alternative approaches view noise as the underlying cause of perceptual transitions [2]. Importantly, models belonging to this class have also taken account of the aforementioned mixed percepts and incomplete transitions during binocular rivalry [40].

Further models have related transitions in perception to a combination of adaptation and noise [3]. In this vein, [41] have argued for a neurodynamic mechanism at the bifurcation between adaptation- and noise-driven processes to be the basis for perceptual transitions during binocular rivalry.

The majority of the models mentioned above has been developed for continuous presentation of binocular rivalry or ambiguous structure-from-motion, while [39] have also studied paradigms with intermittent presentation. As noted above, such stimuli differ significantly from the Lissajous figure used in our current study, which shares aspects with intermittent stimulation due to the existence of overlapping configurations facilitating transitions in perception. Hence, future theoretical and empirical work is needed to probe our modelling approach on paradigms such as binocular rivalry and ambiguous structure-from-motion for both continuous and intermittent presentation and to extend the predictive coding model in order to account for top-down attentional control as well as interactions at earlier processing stages.

Taken together, our current work provides theoretical and empirical evidence across different levels for a driving role of prediction errors in bistable perception, thereby shedding new light on an ongoing debate about the neural mechanisms underlying bistable perception and, more generally, opening up a novel computational perspective on the mechanisms governing perceptual inference.

Supporting information

S1 Text. Mathematical appendix. The appendix contains a detailed mathematical description of our modelling procedures.
(PDF)

S2 Text. Validation against established models of bistable perception. In this supplement, we provide a validation of our modelling approach against established models of bistable perception based on adaptation and inhibition [1], noise [2] and an intermediate model [3].
(PDF)

S1 Video. Example of a full rotation of the specific Lissajous figure used in this experiment.
(WMV)

Author Contributions

Conceptualization: VW HS GH PS KS.

Data curation: VW GH PS.

Formal analysis: VW HS GH PS KS.

Funding acquisition: GH PS KS.

Investigation: VW HS GH PS KS.

Methodology: VW HS GH PS KS.

Project administration: GH PS KS.

Resources: GH PS KS.

Software: VW.

Supervision: PS KS.

Validation: VW HS GH PS KS.

Visualization: VW.

Writing – original draft: VW HS KS.

Writing – review & editing: VW HS GH PS KS.

References

1. Wilson HR. Minimal physiological conditions for binocular rivalry and rivalry. Vision research. 2007; 47 (21):2741–50. https://doi.org/10.1016/j.visres.2007.07.007 PMID: 17764714
2. Moreno-Bote R, Rinzel J, Rubin N. Noise-Induced Alternations in an Attractor Network Model of Perceptual Bistability. Journal of Neurophysiology. 2007; 98(3):1125–1139. https://doi.org/10.1152/jn.00116.2007 PMID: 17615138
3. Lehky SR. An astable multivibrator model of binocular rivalry. Perception. 1988; 17(2):215–28. https://doi.org/10.1068/p170215 PMID: 3067209
4. Lumer ED, Friston KJ, Rees G. Neural correlates of perceptual rivalry in the human brain. Science (New York, NY). 1998; 280(5371):1930–4. https://doi.org/10.1126/science.280.5371.1930
5. Sterzer P, Kleinschmidt A. A neural basis for inference in perceptual ambiguity. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104(1):323–8. https://doi.org/10.1073/pnas.0609006104 PMID: 17190824

6. Knapen T, Brascamp J, Pearson J, van Ee R, Blake R. The Role of Frontal and Parietal Brain Areas in Bistable Perception. Journal of Neuroscience. 2011; 31(28):10293–10301. https://doi.org/10.1523/JNEUROSCI.1727-11.2011 PMID: 21753006

7. Weilnhammer VA, Ludwig K, Hesselmann G, Sterzer P. Frontoparietal cortex mediates perceptual transitions in bistable perception. The Journal of neuroscience: the official journal of the Society for Neuroscience. 2013; 33(40):16009–15. https://doi.org/10.1523/JNEUROSCI.1418-13.2013 PMID: 24089505

8. Frässle S, Sommer J, Jansen A, Naber M, Einhäuser W. Binocular rivalry: frontal activity relates to introspection and action but not to perception. The Journal of neuroscience: the official journal of the Society for Neuroscience. 2014; 34(5):1738–47. https://doi.org/10.1523/JNEUROSCI.4403-13.2014 PMID: 24478356

9. Megumi F, Bahrami B, Kanai R, Rees G. Brain activity dynamics in human parietal regions during spontaneous switches in bistable perception. NeuroImage. 2015; 107:190–7. https://doi.org/10.1016/j.neuroimage.2014.12.018 PMID: 25512040

10. Brascamp J, Blake R, Knapen T. Negligible fronto-parietal BOLD activity accompanying unreportable switches in bistable perception. Nature neuroscience. 2015; 18(11):1672–1678. https://doi.org/10.1038/nn.4130 PMID: 26436901

11. Sterzer P, Kleinschmidt A, Rees G. The neural bases of multistable perception. Trends in cognitive sciences. 2009; 13(7):310–8. https://doi.org/10.1016/j.tics.2009.04.006 PMID: 19540794

12. Hohwy J, Roepstorff A, Friston K. Predictive coding explains binocular rivalry: an epistemological review. Cognition. 2008; 108(3):687–701. https://doi.org/10.1016/j.cognition.2008.05.010 PMID: 18649876

13. Friston K. The free-energy principle: a unified brain theory? Nature reviews Neuroscience. 2010; 11 (2):127–38. https://doi.org/10.1038/nrn2787 PMID: 20068583

14. Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. The Behavioral and brain sciences. 2013; 36(3):181–204. https://doi.org/10.1017/S0140525X12000477 PMID: 23663408

15. Rao RP, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience. 1999; 2(1):79–87. https://doi.org/10.1038/4580 PMID: 10195184

16. Lee TS, Mumford D. Hierarchical Bayesian inference in the visual cortex. Journal of the Optical Society of America A, Optics, image science, and vision. 2003; 20(7):1434–48. https://doi.org/10.1364/JOSAA.20.001434 PMID: 12868647

17. Friston K. A theory of cortical responses. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2005; 360(1456):815–36. https://doi.org/10.1098/rstb.2005.1622 PMID: 15937014

18. Mathys CD, Lomakina EI, Daunizeau J, Iglesias S, Brodersen KH, Friston KJ, et al. Uncertainty in perception and the Hierarchical Gaussian Filter. Frontiers in human neuroscience. 2014; 8:825. https://doi.org/10.3389/fnhum.2014.00825 PMID: 25477800

19. Sundareswara R, Schrater PR. Perceptual multistability predicted by search model for Bayesian decisions. Journal of vision. 2008; 8(5):12.1–19. https://doi.org/10.1167/8.5.12 PMID: 18842083

20. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. NeuroImage. 2009; 46(4):1004–17. https://doi.org/10.1016/j.neuroimage.2009.03.025 PMID: 19306932

21. O'Doherty JP, Hampton A, Kim H. Model-based fMRI and its application to reward learning and decision making. Annals of the New York Academy of Sciences. 2007; 1104(1):35–53. https://doi.org/10.1196/annals.1390.022 PMID: 17416921

22. Rosa MJ, Bestmann S, Harrison L, Penny W. Bayesian model selection maps for group studies. NeuroImage. 2010; 49(1):217–24. https://doi.org/10.1016/j.neuroimage.2009.08.051 PMID: 19732837

23. Brainard DH. The Psychophysics Toolbox. Spatial vision. 1997; 10(4):433–6. https://doi.org/10.1163/156856897X00357 PMID: 9176952

24. Schurger A. A very inexpensive MRI-compatible method for dichoptic visual stimulation. Journal of neuroscience methods. 2009; 177(1):199–202. https://doi.org/10.1016/j.jneumeth.2008.09.028 PMID: 18973774

25. Weilnhammer VA, Sterzer P, Hesselmann G. Perceptual Stability of the Lissajous Figure Is Modulated by the Speed of Illusory Rotation. PloS one. 2016; 11(8):e0160772. https://doi.org/10.1371/journal.pone.0160772 PMID: 27560958

26. Pastukhov A, Vonau V, Braun J. Believable change: bistable reversals are governed by physical plausibility. Journal of vision. 2012; 12(1). https://doi.org/10.1167/12.1.17 PMID: 22267054

27. Friston KJ, Stephan KE. Free-energy and the brain. Synthese. 2007; 159(3):417–458. https://doi.org/10.1007/s11229-007-9237-y PMID: 19325932

28. Eickhoff SB, Stephan KE, Mohlberg H, Grefkes C, Fink GR, Amunts K, et al. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. NeuroImage. 2005; 25 (4):1325–35. https://doi.org/10.1016/j.neuroimage.2004.12.034 PMID: 15850749

29. Levelt WJM. Note on the Distribution of Dominance Times in Binocular Rivalry. British Journal of Psychology. 1967; 58(1-2):143–145. https://doi.org/10.1111/j.2044-8295.1967.tb01068.x PMID: 5582864

30. Logothetis NK, Leopold DA, Sheinberg DL. What is rivalling during binocular rivalry? Nature. 1996; 380 (6575):621–4. https://doi.org/10.1038/380621a0 PMID: 8602261

31. Leopold DA, Logothetis NK. Multistable phenomena: changing views in perception. Trends in Cognitive Sciences. 1999; 3(7):254–264. https://doi.org/10.1016/S1364-6613(99)01332-7 PMID: 10377540

32. Tong F, Meng M, Blake R. Neural bases of binocular rivalry. Trends in cognitive sciences. 2006; 10 (11):502–11. https://doi.org/10.1016/j.tics.2006.09.003 PMID: 16997612

33. Fischer J, Whitney D. Serial dependence in visual perception. Nat Neurosci. 2014; 17(5):738–743. https://doi.org/10.1038/nn.3689 PMID: 24686785

34. St John-Saaltink E, Kok P, Lau HC, de Lange FP. Serial Dependence in Perceptual Decisions Is Reflected in Activity Patterns in Primary Visual Cortex. Journal of Neuroscience. 2016; 36(23):6186–6192. https://doi.org/10.1523/JNEUROSCI.4390-15.2016 PMID: 27277797

35. Kanai R, Carmel D, Bahrami B, Rees G. Structural and functional fractionation of right superior parietal cortex in bistable perception. Current Biology. 2011; 21(3):R106–R107. https://doi.org/10.1016/j.cub.2010.12.009 PMID: 21300270

36. Knill DC, Pouget A. The {Bayesian} brain: the role of uncertainty in neural coding and computation. Trends Neurosci. 2004; 27(12):712–719. https://doi.org/10.1016/j.tins.2004.10.007 PMID: 15541511

37. Hohwy J. Attention and conscious perception in the hypothesis testing brain. Frontiers in psychology. 2012; 3:96. https://doi.org/10.3389/fpsyg.2012.00096 PMID: 22485102

38. Pearson J, Brascamp J. Sensory memory for ambiguous vision. Trends Cogn Sci (Regul Ed). 2008; 12 (9):334–341. https://doi.org/10.1016/j.tics.2008.05.006 PMID: 18684661

39. Klink PC, van Ee R, Nijs MM, Brouwer GJ, Noest AJ, van Wezel RJA. Early interactions between neuronal adaptation and voluntary control determine perceptual choices in bistable vision. Journal of Vision. 2008; 8(5):16. https://doi.org/10.1167/8.5.16 PMID: 18842087

40. Brascamp JW, van Ee R, Noest AJ, Jacobs RHAH, van den Berg AV, R B. The time course of binocular rivalry reveals a fundamental role of noise. Journal of Vision. 2006; 6(11):8–8. https://doi.org/10.1167/6.11.8 PMID: 17209732

41. Panagiotaropoulos TI, Kapoor V, Logothetis NK, Deco G. A Common Neurodynamical Mechanism Could Mediate Externally Induced and Intrinsically Generated Transitions in Visual Awareness. PLoS ONE. 2013; 8(1):e53833. https://doi.org/10.1371/journal.pone.0053833 PMID: 23349748

42. Weilnhammer VA, Ludwig K, Sterzer P, Hesselmann G. Revisiting the Lissajous figure as a tool to study bistable perception. Vision research. 2014; 98:107–12. https://doi.org/10.1016/j.visres.2014.03.013 PMID: 24718018

2.2 The neural correlates of hierarchical predictions for perceptual decisions

Weilnhammer VA, Stuke H, Sterzer P, Schmack K. The Journal of Neuroscience 38, 5008–5021 (2018). DOI: https://doi.org/10.1523/JNEUROSCI.2901-17.2018

The above publication links neural activity during bistable perception to prediction-error related activity in the frontoparietal network[80]. I a next step, we sought to map the neural correlates of internal predictions that are relevant for deciphering the most likely cause of ambiguous sensory information. In this study, we used cross-modal associative learning to induce internal predictions about a bistable apparent-motion stimulus[81]. Importantly, the reliability of the relation between auditory cues and visual targets changed unpredictably over time[30,31]. At the level of behavior, we found that the participants' conscious experience was strongly biased by internal predictions derived from cross-modal associative learning and perceptual history. At the neural level, we observed that predictions about the reliability of the cue-target association were reflected by BOLD-responses in supra-modal regions such as orbitofrontal cortex and hippocampus, while the lower-level conditional target probabilities correlated by BOLD-responses in retinotopic visual cortex. These findings corroborate that the brain uses *hierarchical* predictions in the resolution of sensory ambiguity.

The following text corresponds to the abstract of the article[50]:

"Sensory information is inherently noisy, sparse, and ambiguous. In contrast, visual experience is usually clear, detailed, and stable. Bayesian theories of perception resolve this discrepancy by assuming that prior knowledge about the causes underlying sensory stimulation actively shapes perceptual decisions. The CNS is believed to entertain a generative model aligned to dynamic changes in the hierarchical states of our volatile sensory environment. Here, we used model-based fMRI to study the neural correlates of the dynamic updating of hierarchically structured predictions in male and female human observers. We devised a crossmodal associative learning task with covertly interspersed ambiguous trials in which participants engaged in hierarchical learning based on changing contingencies between auditory cues and visual targets. By inverting a Bayesian model of perceptual inference, we estimated individual hierarchical predictions, which significantly biased perceptual decisions under ambiguity. Although "high-level" predictions about the cue–target contingency correlated with activity in supramodal regions such as orbitofrontal cortex and hippocampus, dynamic "low-level" predictions about the conditional target probabilities were associated with activity in retinotopic visual cortex. Our results suggest that our CNS updates distinct representations of hierarchical predictions that continuously affect perceptual decisions in a dynamically changing environment.

Behavioral/Cognitive

The Neural Correlates of Hierarchical Predictions for Perceptual Decisions

Veith A. Weilnhammer,[1] Heiner Stuke,[1] Philipp Sterzer,[1,2,3]* and Katharina Schmack[1]*

[1]Department of Psychiatry, [2]Bernstein Center for Computational Neuroscience, Charité Universitätsmedizin Berlin, 10117 Berlin, Germany, and [3]Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, 10099 Berlin, Germany

Sensory information is inherently noisy, sparse, and ambiguous. In contrast, visual experience is usually clear, detailed, and stable. Bayesian theories of perception resolve this discrepancy by assuming that prior knowledge about the causes underlying sensory stimulation actively shapes perceptual decisions. The CNS is believed to entertain a generative model aligned to dynamic changes in the hierarchical states of our volatile sensory environment. Here, we used model-based fMRI to study the neural correlates of the dynamic updating of hierarchically structured predictions in male and female human observers. We devised a crossmodal associative learning task with covertly interspersed ambiguous trials in which participants engaged in hierarchical learning based on changing contingencies between auditory cues and visual targets. By inverting a Bayesian model of perceptual inference, we estimated individual hierarchical predictions, which significantly biased perceptual decisions under ambiguity. Although "high-level" predictions about the cue–target contingency correlated with activity in supramodal regions such as orbitofrontal cortex and hippocampus, dynamic "low-level" predictions about the conditional target probabilities were associated with activity in retinotopic visual cortex. Our results suggest that our CNS updates distinct representations of hierarchical predictions that continuously affect perceptual decisions in a dynamically changing environment.

Key words: **Bayesian brain theory; hippocampus; orbitofrontal cortex; predictive coding; sensory predictions; visual perception**

Significance Statement

Bayesian theories posit that our brain entertains a generative model to provide hierarchical predictions regarding the causes of sensory information. Here, we use behavioral modeling and fMRI to study the neural underpinnings of such hierarchical predictions. We show that "high-level" predictions about the strength of dynamic cue–target contingencies during crossmodal associative learning correlate with activity in orbitofrontal cortex and the hippocampus, whereas "low-level" conditional target probabilities were reflected in retinotopic visual cortex. Our findings empirically corroborate theorizations on the role of hierarchical predictions in visual perception and contribute substantially to a longstanding debate on the link between sensory predictions and orbitofrontal or hippocampal activity. Our work fundamentally advances the mechanistic understanding of perceptual inference in the human brain.

Introduction

When dealing with complex and volatile environments, agents are faced with uncertainties introduced by imprecise sensory signals ("perceptual uncertainty"), the known stochasticity of predictive relationships within a stable environment ("expected uncertainty"), or changes in the statistical properties of the environment that compromise predictions based on previous experience ("unexpected uncertainty"; Yu and Dayan (2005)).

To make adaptive inferences about the causes of uncertain information, the brain recurs to learned predictions, which are thought to match the hierarchical structure of the world (Friston, 2005). For instance, when estimating the flight trajectory of the shuttlecock during a badminton match, the current shuttlecock position depends on previous shuttlecock positions and this dependence of current on previous positions in turn depends on the current wind situation. Sensory signals indicating the current

Received Oct. 6, 2017; revised April 1, 2018; accepted April 8, 2018.

Author contributions: V.A.W., P.S., and K.S. designed research; V.A.W., H.S., and K.S. performed research; V.A.W., P.S., and K.S. analyzed data; V.A.W., H.S., P.S., and K.S. wrote the paper.

This work was supported by the German Federal Ministry of Education and Research within the framework of the e:Med research and funding concept (Grant 01ZX1404A to K.S.) and the German Research Foundation (Grant STE 1430/7-1 to P.S.). V.A.W. is a participant in the Charité Junior Clinical Scientist Program funded by the Charité Universitaetsmedizin Berlin and the Berlin Institute of Health. K.S. is a participant in the Charité Clinical Scientist Program funded by the Charité Universitaetsmedizin Berlin and the Berlin Institute of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The authors declare no competing financial interests.

*P.S. and K.S. contributed equally to this work.

Correspondence should be addressed to Veith A. Weilnhammer, Department for Psychiatry and Psychotherapie, Charité Campus Mitte, Charitéplatz 1, 10117 Berlin. E-mail: veith-andreas.weilnhammer@charite.de.

DOI:10.1523/JNEUROSCI.2901-17.2018

 0270-6474/18/385008-14$15.00/0

shuttlecock position may be noisy (e.g., due to partial occlusion), resulting in "perceptual uncertainty". Furthermore, expected uncertainty arises from the known irregularity of the shuttlecock trajectory within stable wind conditions, whereas unexpected uncertainty results from changes in wind conditions that affect the relation between successive shuttlecock positions. To deal with these uncertainties, a badminton player cannot only rely on sensory signals generated by the shuttlecock, but also requires a "high-level" prediction about the likely shuttlecock trajectory given the current wind condition, which he or she can then use to generate a "low-level" prediction about the current shuttlecock position based on previous positions. Here, we investigated how such hierarchically related predictions are updated and maintained in the brain.

Hierarchical predictions can be elegantly formalized by Bayesian predictive coding. Bayesian theories propose that our brain entertains a predictive model of the environment, enabling inference and learning under uncertainty (Knill and Pouget, 2004; Yu and Dayan, 2005; Behrens et al., 2007; Hohwy et al., 2008; Nassar et al., 2010; Payzan-LeNestour et al., 2013). These perspectives are tightly related to hierarchical predictive coding schemes (Rao and Ballard, 1999; Lee and Mumford, 2003), which assume that predictions are serially implemented across hierarchical levels and that prediction errors are generated in cases of mismatch between predictions and incoming signals. Please note that, here, we do not use the term "predictive coding" in its narrow sense for the specific instantiation of top-down predictions proposed by Rao and Ballard (1999), but in its broader sense referring to hierarchical predictive models aiming at the minimization of prediction errors (Clark, 2013) or free energy (Friston, 2005).

To investigate the neural implementation of hierarchical predictions, we devised a crossmodal associative learning task (Fig. 1, Schmack et al. (2016)) in which participants made inferences about volatile cue–target associations. In brief, we presented participants with flashing dot quartets that elicited the perception of either clockwise (CW) or counterclockwise (CCW) tilt motion. These dot quartets were preceded by auditory cues that probabilistically predicted the tilt direction of the upcoming visual target. Over time, observers learned the relation between auditory and visual stimuli, whereas cue–target contingencies changed unpredictably at times unknown to the participants. Crucially, perceptually ambiguous dot quartets equally compatible with CW and CCW tilt were covertly interspersed in the sequence of visual stimulation. In relation to the example of the badminton match, the CW or CCW tilting dot quartet (i.e., the visual target stimulus) corresponds to the current shuttlecock position. The auditory cue in our experiment corresponds to the current wind condition in the badminton example. That is, by introducing changes in cue–target association, our paradigm induces varying degrees of predictability of the visual target given the cue, akin to changes in predictability of the shuttlecock position due to changing wind conditions. Moreover, perceptual uncertainty is introduced by the use of ambiguous visual stimuli, akin to perceptual uncertainty caused by, for example, temporary partial occlusion of the shuttlecock.

We used computational modeling in a Bayesian framework (Mathys et al., 2014a) to estimate hierarchically related predictions on a trial-by-trial basis. Correlating these trialwise estimates with fMRI time courses allowed us to dissociate the neural correlates of "high-level" predictions regarding the coupling of tones and visual stimuli from "low-level" predictions regarding the probability of binary perceptual outcomes.

Materials and Methods

Participants

Twenty-five participants took part in the experiment, which was conducted with informed written consent and approved by the local ethics committee. One participant had to be excluded because of not following the experimental instructions correctly. A second participant was excluded due to excessive movement inside the scanner (5 mm maximum average translational movement across runs. All remaining participants ($N = 23$, age 19–34 years, mean 25.6 years, 14 female) had normal or corrected-to-normal vision and no prior psychiatric or neurological medical history.

Experimental procedures

Main experiment. In this fMRI experiment, we aimed at disentangling the neural representations of continuously updated hierarchical predictions. To this end, participants performed an associative reversal learning task (Fig. 1*A*) similar to Schmack et al. (2016), which induced changing expectations about visual stimuli. High or low tones were coupled with subsequently presented CW or CCW tilting dot pairs, which could be either unambiguous or ambiguous with regard to the direction of tilt. On unambiguous trials, tilt direction was determined by a motion streak, yielding a clear impression of the corresponding movement. The association of tones with tilting directions was probabilistic (75% correct and 12.5% incorrect associations with 12.5% ambiguous trials, see below) with contingencies changing unpredictably every 16–32 trials. Ambiguous trials used the phenomenon of apparent motion (Muckli et al., 2005; Sterzer et al., 2006; Sterzer and Kleinschmidt, 2007) to induce the percept of tilting movement and were covertly interspersed in the experimental sequence (12.5% of all trials). Here, the motion streak was omitted and the physical visual stimulus was hence uninformative with regard to the direction of tilt.

During the main experiment, participants completed a total of 576 trials, which were divided into 9 individual runs of varying length with a medium duration of ~9 min. Visual and auditory stimuli were produced using MATLAB 2014b (The MathWorks) and Psychophysics Toolbox 3. Frames were projected at 60 Hz using a Sanyo LCD projector (resolution 1024 × 768 pixels) on a screen placed at 60 cm viewing distance at the Trim Trio Siemens 3T fMRI scanner's bore.

Auditory stimuli were presented binaurally at −15 dB (relative to maximum intensity) using MRI-compatible headphones powered by MR-ConFon hardware. At the beginning of every trial (Fig. 1*B*), a high (576 Hz) or low (352 Hz) tone was presented for a total of 300 ms. Immediately afterwards, participants indicated their prediction about whether the upcoming visual stimulus would tilt CW or CCW by pressing a left or right button on a standard MRI button box using the index and middle fingers of their right hand. The prediction screen was displayed for 1 s and consisted of 2 single arrows (displayed at 2.05° eccentricity right and left of fixation and turning from white to red after the response). The offset between the prediction screen and the onset of the visual stimuli was jittered between 100 and 300 ms (mean offset: 200 ms). Visual stimuli consisted of two light-gray dots of 1.2° diameter presented simultaneously at an eccentricity of 4.01° on the vertical (starting position) or horizontal (final position) meridian. The circumference of the tilting movement was depicted by a dark-gray circular streak of 1.2° width, which was displayed throughout the trial. Starting and final dot positions were presented for 600 ms, separated by trajectories of 33 ms duration.

On unambiguous trials, the tilting direction was defined by motion streaks (upper right and lower left quadrant for CW tilt, upper left and lower right for CCW tilt). On ambiguous trials, no motion streak was presented and visual stimuli were compatible with CW and CCW tilt. Immediately after presentation of the visual stimulus, participants reported their perception by pressing a left or right button using the index and middle finger of their right hand. The video response screen consisted of two double arrows (displayed at 2.05° eccentricity right and left of fixation and turning from white to red after response) and was presented for 1 s. Trials were separated by fixation intervals jittered between 0.5 and 2.5 s (mean fixation interval: 1.5 s) and amounted to a mean duration of 5.25 s each.

Perceptual rating. Subsequent to the main experiment, we aimed at assessing the perceptual quality of ambiguous and unambiguous trials in

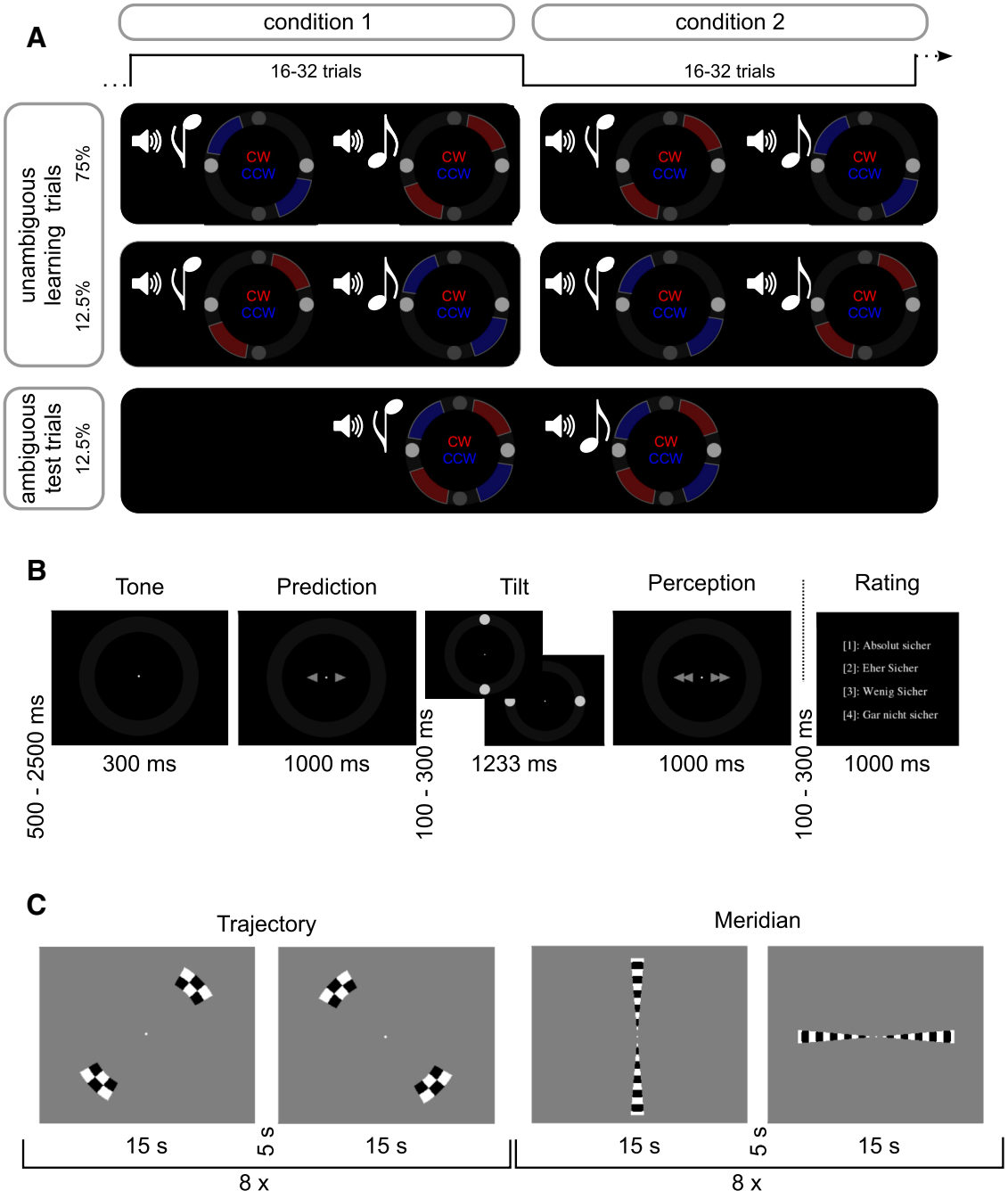


Figure 1. ***A***, Experimental paradigm. In this experiment, we coupled CW (motion trajectory highlighted in red) or CCW (motion trajectory highlighted in blue) tilting dot pairs (visual targets) with high or low tones (auditory cues), which were predictive of the upcoming visual stimulus at a contingency of 75%. Importantly, the association between tones and visual stimuli reversed unpredictably for the participants every 16–32 trials. Furthermore, 12.5% of tilting dot pairs were ambiguous with regard to the perceived direction of tilt. Such test trials enabled us to quantify the influence of predictions formed during crossmodal associative learning on visual perception. ***B***, Trial structure: Main experiment. After presentation of a high (576 Hz) or low pitch (352 Hz) auditory cue, participants indicated their predicted tilting direction. After presentation of the visual target (which could be either unambiguous or ambiguous), participants reported their perception. In an additional perceptual rating experiment, this sequence was followed by a rating on the certainty associated with the perceptual response. ***C***, Trial structure: Localizer. At the end of the fMRI experiment, we conducted localizer sessions mapping the meridians and the dot trajectories of CW and CCW tilt, respectively. Checkerboards were flickered in the respective areas eight times for 15 s in alternation while participants performed a challenging change detection task at fixation.

an additional perceptual rating experiment, which was performed during the anatomical scan. Here, trial structure was identical to the main experiment. However, the video response screen was followed by a confidence rating (offset to perceptual response jittered between 100 and 300 ms, mean offset: 200 ms), which displayed a 4-point scale where 1 = very sure, 2 = rather sure, 3 = rather unsure, and 4 = very unsure with regard to the visual percept for a total of 1 s. Participants reported their rating using the index, middle, ring and little finger of the right hand. The selected rating turned from white to red after response. In total, participants rated their perceptual confidence for a total of 60 trials.

Localizer. Given that the binary perceptual outcomes of the visual target were spatially separated, our design enabled us to investigate how activity in retinotopic stimulus representations in primary visual cortex would relate to predictive processes evoked during the main experiment. To identify voxels corresponding to CW or CCW tilt, we conducted two localizer scans (Fig. 1*C*) at the end of experimental session. The first

localizer was designed to map the dot trajectories. Black-and-white checkerboards covering the circular dot trajectories from the main experiment were flickered at a frequency of 8 Hz in each visual field quadrant, but did not cover the starting and final position of the dot pairs. Specifically, the upper right and lower left (CW tilt) as well as the upper left and lower right quadrant (CCW tilt) were flickered in alternating sequence for 15 s each for a total of 8 repetitions separated by 5 s of fixation.

The second localizer was conducted with identical temporal structure, but mapped the vertical and horizontal meridian spanning over starting and final dot positions. For both localizers, checkerboards were scaled by the cortical magnification factor and participants performed a fixation task, responding to color changes in the fixation dot (alternating between white and red in unpredictable intervals) with their right index finger.

Behavioral analysis

The behavioral analysis outlined here is directed at the influence of the current cue–target association on perceptual decisions under ambiguity. In previous work using a similar experimental design with a different visual stimulus (Schmack et al., 2016), we found that in addition to a main effect of "associative learning," perceptual history also had an influence on perceptual decisions under ambiguity in the form of "priming" and "sensory memory." Whereas "priming" refers to the influence of the immediately preceding trial on the current trial, the term "sensory memory" (Pearson and Brascamp, 2008) is defined by the influence of the preceding ambiguous trial on the current ambiguous trial and therefore acts over longer timescales. In our current work, we used an optimized experimental design with a different visual stimulus that we expected to maximize the effect of associative learning while minimizing the effects of perceptual history. Nevertheless, in our behavioral analyses, we considered not only the main effect of associative learning but also the effects of priming and sensory memory to account for variance of no interest caused by perceptual history.

Conventional analysis. To establish that prior predictions acquired during the course of the experiment biased perceptual decision under ambiguity, we performed a series of conventional behavioral analyses, which furthermore served as a validation for our inverted Bayesian model (see below). Our central interest was in the effect of learned tone–target associations on perceptual decisions under ambiguity. We therefore calculated the proportion of ambiguous percepts congruent to the currently prevalent hidden contingency (associative learning) averaged across runs and participants. Given our previous findings suggesting additional effects of perceptual history on perceptual decisions under ambiguity, we further quantified the proportion of trials perceived in congruence with the preceding unambiguous trial (priming) or the preceding ambiguous trial (sensory memory; Schmack et al., 2016).

We further investigated the effectiveness of the disambiguation by calculating the proportion of unambiguous trials perceived according to the disambiguation and averaged across runs and participants.

To assess the results from our perceptual rating experiment, we calculated the proportion of trials rated as 1 = very sure, 2 = rather sure, 3 = rather unsure, and 4 = very unsure for unambiguous and ambiguous trials separately and averaged across participants. To assess a potential mediation of the effect of predictions on perceptual decision under ambiguity by perceptual uncertainty, we conducted an across-participants correlation between average perceptual ratings and the proportion of ambiguous trials perceived according to the current cue–target contingency.

Finally, correlating the metrics for the strength of the impact of learned associations on perceptual decisions under ambiguity between conventional and model-based behavioral analyses allowed us to validate our Bayesian modeling approach.

Bayesian modeling. To investigate the neural correlates of hierarchical predictions, we adopted a Bayesian modeling approach (implemented previously in Schmack et al., 2016), which allows for the estimation of individual trial-by-trial model quantities such as the dynamic and continuously updated "high-level" prediction about the association between auditory cues and visual target or the inferred "low-level" conditional probability of a binary visual outcome given a specific auditory cue.

Our model, which is defined in detail in the section "Mathematical model description," frames perception as an inferential processes in which perceptual decisions are based on posterior distributions. According to Bayes' rule, such posterior distributions are derived from likelihood distributions representing the sensory evidence, and prior distributions, which, in the context of this experiment, can be used to formalize expectations about perceptual outcomes.

Crucially, here, we were interested in such perceptual expectations or priors that are formed by associative learning; that is, the subjects' continuously updated inference on the probabilistic coupling between tones and visual stimuli (please note that this is not equivalent to the hidden contingency used for conventional analysis, which is in principle unknown to the participant). As indicated by our previous work (Schmack et al., 2016), perception might be further influenced by priors that are derived from perceptual history: priming (the influence of a visual percept on the subsequent trial) and sensory memory (the influence of the visual percept in an ambiguous trial on the subsequent ambiguous trial). Inclusion of these priors based on perceptual history into a model helps to explain away additional variance of no interest. Please note that the factors of associative learning and priming constitute potential priors for perceptual decisions on all trials regardless of ambiguity, whereas sensory memory is defined as a prior for perceptual decisions under ambiguity only.

All of these priors (associative learning, priming, and sensory memory) can be modeled by Gaussian probability distributions, which are defined by their respective mean and precision (the inverse of variance). Importantly, the precision term represents the impact of a prior on the posterior distribution and thus relates to its influence on visual perception.

In the analysis presented here, we fitted our model on two behavioral responses given by our participants: The prediction of upcoming tilting direction $y_{\text{prediction}}$ (which we hypothesized to depend on the conditional probability of tilting direction given the tone as expressed by the prior distribution "associative learning") and the perceived tilting direction $y_{\text{perception}}$ (which we reasoned to be based on a specific combination of the prior distributions "associative learning", "priming", "sensory memory", and the likelihood-weight "disambiguation"). Therefore, our model is divided into two interacting parts: a "contingency" model, which was built to model the inferred association between tones and CW or CCW tilt and used to extract "high-level" model quantities, and a "perceptual" model, which was designed to predict the participants' perceptual choices and enabled us to assess "low-level" model quantities.

To determine which factors drive perceptual predictions relevant for perceptual decisions under ambiguity, we used Bayesian model selection. In addition to the factor "associative learning", which we were interested in primarily, we considered the additional factors "priming and sensory memory" to allow for models that account for the variance caused by perceptual history. We constructed behavioral models incorporating all combinations of the prior distributions "associative learning" (A), "priming" (P), and "sensory memory" (S), whereas all models considered incorporated the distribution "disambiguation", which adjusts the weight of the fixed bimodal likelihood. This yielded a total of eight behavioral models to be compared (A-P-S-, A-P-S+, A-P+S-, A-P+S+, A+P-S-, A+P-S+, A+P+S-, A+P+S+), which were optimized for the prediction of both behavioral responses using a free energy minimization approach. This allowed us to compare the behavioral models using random effects Bayesian model selection (Stephan et al., 2009). We used a version of the hierarchical Gaussian filter for binary inputs (Mathys et al., 2014a, 2014b), as implemented in the HGF 4.0 toolbox (distributed within the TAPAS toolbox translationalneuromodeling.org/tapas/), for model optimization and SPM12 (http://www.fil.ion.ucl.ac.uk/spm/) for model selection.

After identifying the optimal model using Bayesian model selection, we analyzed its posterior parameters using classical frequentist statistics and extracted model quantities for model-based fMRI. To test for a relation between fMRI activity and "high-level" predictions, we extracted the absolute cross-model prediction $\hat{\mu}_2$ from the contingency model. To account for additional variance in the BOLD signal, we further extracted the precision of the absolute cross-model prediction $\hat{\pi}_2$ and the precision-weighted cross-modal prediction error $|\epsilon_2|$ from the contingency model.

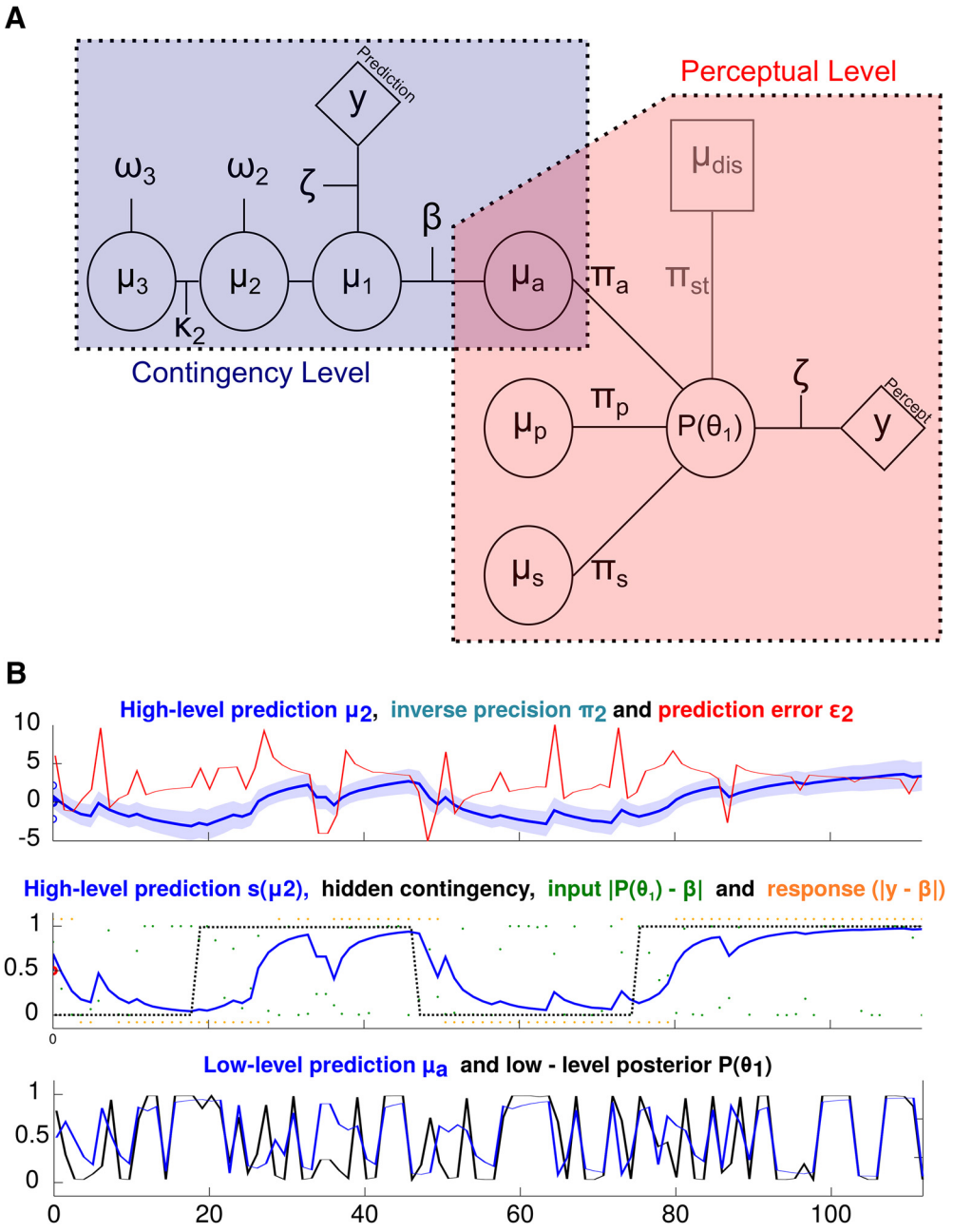


Figure 2. Hierarchical Gaussian filter. ***A***, The behavioral model consists of a standard hierarchical Gaussian filter for binary perceptual outcomes (contingency level), representing the inferred association between tones and tilting directions during the experiment. This part of the model is coupled to the perceptual level, which determines the influence of prior predictions derived from previous cue–target associations as well as priming and sensory memory on perceptual decisions. ***B***, Exemplary model quantities for one individual participant and run. The top displays the time course of the "high-level" prediction μ_2, its variance (i.e., the inverse precision π_2) as well as the precision-weighted "high-level" prediction error ε_2. The middle panel shows the sigmoid transform of the "high-level" prediction $s(\mu_2)$, the time course of the underlying contingency (black dotted line) as well as inputs and responses (both transformed on the level of the contingency between auditory cues and visual targets). The bottom displays the "low-level" conditional probability of CW tilt (in blue) as well as the "low-level" posterior probability of CW tilt (in black).

To investigate the relationship between fMRI activity and "low-level" predictive processes, we assessed the dynamic stimulus-specific prediction μ_a (i.e., the inferred conditional probability of CW tilt given the tone) and its analog $1 - \mu_a$ (i.e., the inferred conditional probability of CCW tilt given the tone) from the perceptual model. To capture additional variance in the recorded BOLD signal, we furthermore extracted the following model quantities from the perceptual model: the posterior probability of CW tilt $P(\theta_1)$ and CCW tilt $P(\theta_0)$, the choice prediction error $\varepsilon_{\text{choice}}$ and the perceptual prediction error $|\delta_q|$.

Please see the section "Mathematical model description" for a detailed definition of our modeling procedures. Figure 2*A* provides a graphical illustration of the modeling approach. We provide exemplary time courses for "high-" and "low-level" model quantities in Figure 2*B*. Table 1 provides a summary of model quantities and model parameters, including prior mean and variance for inversion as well as average posterior parameter estimates across participants.

Mathematical model description

Here, we applied a Bayesian modeling approach to assess the continuous updating of predictions about the causes of sensory input and their impact on perceptual decisions under ambiguity. We devised a model that was inverted on two behavioral responses given by the participants: The prediction of upcoming tilting direction $y_{\text{prediction}}$ and the perceived tilting direction $y_{\text{perception}}$. With this, we inferred on model parameters that

Table 1. Summary of model parameters and quantities

	Name	Explanation	Inversion		
Sensory Stimulation	μ_{dis}	Mean of sensory stimulation			
	β	High- or low-pitch tone			
Responses	$y_{prediction}$	Binary prediction			
	$y_{perception}$	Binary perceptual decision			
Model Parameters			Prior mean	Prior variance	Posterior
Perceptual Model	π_a	Associative precision	0.5	1	1.6052 ± 0.0456
	π_p	Priming precision	0.5	1	0 ± 0
	π_s	Sensory memory precision	0.5	1	0.6138 ± 0.0511
	π_{dis}	Disambiguation precision	1.5	0	1.5 ± 0
Contingency Model	ω_2	Learning rate of 2nd level	−1.28	1	-0.0483 ± 0.0713
	ω_3	Learning rate of 3rd level	−6.14	1	-6.6800 ± 0.0469
	κ_2	Coupling strength between 3rd and 2nd level	1	0	1 ± 0
	$\mu^0_{2/3}$	Initial mean of 2nd/3rd level	0/1	0/0	0/1
	σ^0_2	Initial variance of 2nd level	4.6413	1	4.5739 ± 0.0536
	σ^0_3	Initial variance of 3rd level	4	1	3.3315 ± 0.0536
Response Mapping	ζ	Inverse decision temperature (response model)	1	0	1
Selected Model Quantities					
Predicted Responses	$\hat{y}_{prediction}$	Model prediction on $y_{prediction}$			
	$\hat{y}_{perception}$	Model prediction on $y_{perception}$			
Perceptual Model	μ_a	Inferred conditional probability of CW-tilt ("low-level" prediction)			
	δ_q	Perceptual prediction error			
	ϵ_{choice}	Choice prediction error			
	$P(\theta_1)$	Posterior probability of perceiving CW-tilt			
Contingency Model	$\hat{\mu}_2$ and $\hat{\pi}_2$	Prior mean ("high-level" prediction) and precision of 2nd level			
	ϵ_2	Precision-weighted ("high-level") contingency prediction error			

Table 2. Model-based fMRI Results with thresholds $p < 0.05$, FWE, for $|\widehat{\mu_2}|$ and $p < 0.001$, uncorr., for μ_a

$|\hat{\mu}_2|$

Region	Hem.	x	y	z	T	Region	Hem.	x	y	z	T
Mid. Orbital Gy.	R	6	40	−10	11.04	Post. Hippocampus	R	10	−40	5	7.65
Mid. Orbital Gy.	L	−4	50	−8	11.44	Post. Hippocampus	L	−7	−42	5	8.70
Caudate Nucleus	L	−7	18	−8	9.06	Precuneus	R	8	−50	8	6.49
Insula	R	43	−12	5	11.01	Precuneus	L	−4	−54	12	9.04
Precentral Gy.	R	16	−24	78	9.61	Postcentral Gy.	R	48	−12	35	6.50
Post. med. frontal Gy.	R	10	−17	78	7.95	Postcentral Gy.	L	−20	−30	72	7.05

μ_a

Region	Hem.	x	y	z	T	Region	Hem.	x	y	z	T
Rolandic Operculum	L	−60	3	5	4.47	Inf. Temporal Gy.	L	−42	−62	−10	3.70
Inf. Occipital Gy.	L	−42	−70	−8	3.79	Caudate Nucleus	L	−14	6	15	4.29
Inf. Frontal Gy.	R	48	38	0	3.89	Caudate Nucleus	R	18	−17	20	4.27

govern the updates in model quantities belonging to two different interacting parts of our model: A "contingency" model dealing with the inferred contingencies between concurring auditory and visual stimuli and a "perceptual" model, which integrates different sources of prior and likelihood information to predict individual perceptual choices.

Perceptual model. At each time point t, the two alternative visual percepts are predicted on the basis of a posterior probability distribution over θ:

$$\theta = \begin{cases} > 0.5: & CW \quad tilt \\ < 0.5: & CCW \quad tilt \end{cases} \tag{1}$$

Participants responded with button presses indicating the current visual percept as follows:

$$y_{perception}(t) = \begin{cases} 1: & CW \quad tilt \\ 0: & CCW \quad tilt \end{cases} \tag{2}$$

Based on previous work (Schmack et al., 2016), we formalized a number of prior distributions that could influence on participants' perception, considering separate contributions of priming, sensory memory, and associative learning. The latter was driven by the co-ocurrence of the direction of tilt (see above) and the pitch of the preceding tone, which was defined as follows:

$$\beta(t) = \begin{cases} 1: & high \quad pitch \\ 0: & low \quad pitch \end{cases} \tag{3}$$

To map the dynamic inference on the contingency between tones β and perceived direction of tilt y, we constructed a three-level hierarchical Gaussian filter (Mathys et al. (2014b), see below for details), which received the conjunction of tone and posterior probability of tilt direction as input. From here, we extracted first level prediction $\hat{\mu}_1(t)$, which represents the inferred contingency over tones and rotations. This was transformed into the conditional probability of CW tilt given the tone as follows:

$$\mu_a(t) = \begin{cases} \hat{\mu}_1(t): & for \quad \beta(t) = 0 \\ 1 - \hat{\mu}_1(t): & for \quad \beta(t) = 1 \end{cases} \tag{4}$$

This defines the mean of the prior distribution "associative learning" (associative learning $\sim \mathcal{N}(\mu_a, \pi_a^{-1})$), while π_a represents its precision. Please note that the conditional probability of CCW tilt is given by $1 - \mu_a$. We refer to these model quantities as "low-level perceptual predictions".

Likewise, the mean of the prior distribution "priming" (priming $\sim \mathcal{N}(\mu_p, \pi_p^{-1})$) in trial t was defined by the visual percept in the preceding trial:

Table 3. Explorative model-based fMRI Results. Statistical thresholds are $p < 0.05$ FWE for the regressors "Tone", "Tile", $|\widehat{\mu}_2|$, $|\delta_q|$ and $|\epsilon_2|$ as well as $p < 0.001$ uncorr. for the remaining regressors

Tilt						Tone					
Region	Hem.	x	y	z	T	Region	Hem.	x	y	z	T
Inf. Occipital Gy.	R	38	−74	−12	12.30	Sup. Temporal Gy.	R	53	−14	0	14.29
Inf. Occipital Gy.	L	−37	−84	−5	11.75	Sup.Temporal Gy.	L	−60	−42	15	12.06
Mid. Occipital Gy.	R	30	−92	0	12.60	Cerebelum (VI)	R	33	−62	−28	12.54
Mid. Occipital Gy.	L	−22	−97	−2	11.49	Cerebelum (VI)	L	−30	−67	−28	9.20
Fusiform Gy.	R	30	−72	−18	12.81	Thalamus	R	8	−12	2	8.64
Fusiform Gy.	L	−32	−77	−18	15.10	Thalamus	L	−7	−17	0	9.82
Lingual Gy.	R	23	−90	−5	14.45	Post. med. frontal Gy.	R	8	23	52	7.47
hMTI+/V5	R	50	−70	5	12.29	Post. med. frontal Gy.	L	−7	3	55	7.28
						Precentral	L	−37	−17	55	7.78
						Insula	R	30	23	−2	7.06
$\lvert\widehat{\mu}_2\rvert$											
Region	Hem.	x	y	z	T	Region	Hem.	x	y	z	T
Heschls Gy.	R	43	−22	8	7.65	Insula	L	−44	0	−2	6.52
Heschls Gy.	L	−37	−27	10	6.48	Postcentral Gy.	R	23	−37	72	7.11
$\lvert\delta_q\rvert$						$\lvert\epsilon_2\rvert$					
Region	Hem.	x	y	z	T	Region	Hem.	x	y	z	T
Insula	L	−32	26	5	8.29	Precentral Gy.	R	46	6	28	7.89
Mid. Temporal Gy.	L	−50	−60	2	6.93						
Precentral Gy.	R	46	6	28	7.89	Inf. Parietal Lob.	R	46	−50	50	7.31
Inf. Parietal Lob.	L	−40	−80	22	7.54	Inf. Parietal Lob.	L	−50	−47	55	7.02
Insula	L	−32	26	5	5.81	Sup. Parietal Lob.	R	6	−67	48	6.48
Inf. Frontal Gy.	L	−47	16	30	5.20	Caudate Nucleus	R	13	0	20	5.25
ϵ_{choice}						$P(\theta = 1)$					
Region	Hem.	x	y	z	T	Region	Hem.	x	y	z	T
Inf. Occipital Gy.	L	−42	−77	−10	5.25	Sup. Occipital Gy.	R	20	−97	28	4.21
		-				Posterior-medial frontal Gy.	L	−12	18	60	4.55

$$\mu_p(t) = y_{perception}(t-1) \tag{5}$$

The mean of the prior distribution "sensory memory" (sensory memory $\sim \mathcal{N}(\mu_s, \pi_s^{-1})$) in trial t was defined by the visual percept in the preceding ambiguous trial t_a:

$$\mu_s(t) = y_{preception}(t_a) \tag{6}$$

In addition to these prior distributions, we defined the disambiguation (i.e., the presence of motion streaks along the trajectory of tilt) by means of the likelihood weight "disambiguation" (disambiguation $\sim \mathcal{N}(\mu_{dis}, \pi_{dis}^{-1})$) in trial t:

$$\mu_{dis}(t)\begin{cases} 1: & CW & (disambiguation) \\ 0.5: & CW/CCW & (ambiguous) \\ 0: & CCW & (disambiguation) \end{cases} \tag{7}$$

To predict the perceptual outcomes, we derived the posterior distribution with respect to CW or CCW tilt from the model. This distribution results from a weighting of a bimodal likelihood distribution by a combination of prior distributions such as "associative learning", "priming", "sensory memory", as well as the likelihood weight "disambiguation".

For a specific combination of these prior distributions, a joint prior distribution with mean μ_m and variance π_m can be calculated by adding up the means of influencing factors relative to their respective precision:

$$\mu_m(t) = \frac{\pi_a\mu_a(t) + \pi_p\mu_p(t) + \pi_s\mu_s(t)}{\pi_m} \tag{8}$$

$$\pi_m = \pi_a + \pi_p + \pi_s \tag{9}$$

This joint prior distribution (described by μ_m and π_m) as well as the disambiguation (defined by μ_{dis} and π_{dis}) is used to adjust the density ratio of the posterior for the two peak locations $\theta_0 = 0$ and $\theta_1 = 1$:

$$r(t) = \frac{P(\theta_1(t))}{P(\theta_0(t))} = exp\left(\frac{\left(\theta_1 - \frac{\pi_m\mu_m(t) + \pi_{dis}\mu_{dis}(t)}{\pi_m + \pi_{dis}}\right)^2 - \left(\theta_0 - \frac{\pi_m\mu_m(t) + \pi_{dis}\mu_{dis}(t)}{\pi_m + \pi_{dis}}\right)^2}{2 * (\pi_m + \pi_{dis})^{-2}}\right) \tag{10}$$

$$P(\theta_1) = \frac{1}{r(t) + 1} \tag{11}$$

$P(\theta_1)$ denotes the posterior probability of CW tilt. Therefore, $1 - P(\theta_1)$ represents the posterior probability of CCW tilt. For simplicity, we refer to $P(\theta_1)$ and $P(\theta_0)$ as "low-level" posteriors."

The model prediction $\hat{y}_{Perception}$ on the participants percept is given by applying a unit sigmoid function with inverse decision temperature $\zeta = 1$ to $P(\theta_1)$:

$$\hat{y}_{Perception} = \frac{P(\theta_1)^\zeta}{P(\theta_1)^\zeta + (1 - P(\theta_1))^\zeta} \tag{12}$$

From here, we extracted a "perceptual prediction error", which was given by:

$$\delta_q = P(\theta_1) - y_{perception} \tag{13}$$

In addition, we defined a "choice prediction error", which was obtained by subtracting the inferred conditional probability of CW tilt given the tone (i.e., μ_a) from the actual perceptual outcome $y_{\text{perception}}$:

$$\epsilon_{choice} = \mu_a - y_{perception} \tag{14}$$

Contingency model. To extract the inferred trial-by-trial prediction $\widehat{\mu}_1(t)$, we used a version of the three-level hierarchical Gaussian filter (Mathys et al., 2011). The input to the HGF modeling the inferred contingency between auditory and visual stimuli was defined by the following:

$$Input(t) = |P(\theta_1(t)) - \beta(t)| \tag{15}$$

Please note that, due to the lack of a stereodisparity cue in ambiguous trials, $P(\theta_1(t))$ is closer to 0 or 1 on unambiguous trials. Therefore, updates in the inferred contingency are smaller in ambiguous cases and the HGF implemented here specifically takes differences in perceptual certainty between ambiguous and unambiguous trials into account.

Likewise, the participants' prediction was defined as follows:

$$y_{prediction}(t) = \begin{cases} |1 - \beta(t)|: & CW \quad tilt \\ |0 - \beta(t)|: & CCW \quad tilt \end{cases} \tag{16}$$

The posterior of the first level $\mu_1(t)$ is set to be equal to $Input(t)$:

$$\mu_1(t) = Input(t) \tag{17}$$

The second-level prediction of the HGF models the tendency of the first level toward $\mu_1(t) = 1$ and is given by the following:

$$\mu_2(t) = \widehat{\mu}_2(t) + \frac{1}{\pi_2(t)} * \delta_1(t) \tag{18}$$

$$\widehat{\mu}_2(t) = \mu_2(t-1) \tag{19}$$

Please note that we refer to the strength of the second-level prediction $|\widehat{\mu}_2(t)|$ as the crossmodal or "high-level" prediction. The precision of the second-level prediction evolves according to the following:

$$\pi_2(t) = \widehat{\pi}_2(t) + \frac{1}{\widehat{\pi}_1(t)} \tag{20}$$

The first-level prediction $\widehat{\mu}_1$ is defined by a logistic sigmoid transform of the second-level prediction μ_2 as follows:

$$\widehat{\mu}_1(t) = s(\mu_2(t-1)) \tag{21}$$

The difference between the first level prediction $\mu_1(t)$ and first-level posterior $\mu_1(t)$ yields a prediction error $\delta_1(t)$ as follows:

$$\delta_1(t) = \mu(t) - \hat{\mu}_1(t) \tag{22}$$

Crucially, $\delta_1(t)$ is combined with the second level precision Π_2, yielding the precision-weighted "high-level" prediction error $\varepsilon_2(t)$, which updates second-level prediction $\hat{\mu}_2(t)$ as follows:

$$\epsilon_2(t) = \frac{1}{\pi_2} * \delta_1(t) \tag{23}$$

The precision of the prediction on the first and second level evolve according to the following:

$$\widehat{\pi}_1(t) = \frac{1}{\widehat{\mu}_1(t) * (1 - \widehat{\mu}_1(t))} \tag{24}$$

$$\widehat{\pi}_2(t) = \frac{1}{\sigma_2(t) + \exp(\kappa_2 * \mu_3(t-1) + \omega_2)} \tag{25}$$

The volatility prediction error δ_2 governs the update to the third level of the HGF and is given by the following:

$$\delta_2(t) = \left(\frac{1}{\pi_2(t)} + (\mu_2(t) - \widehat{\mu}_2(t))^2 \right) * \widehat{\pi}_2(t) - 1; \tag{26}$$

The third-level prediction $\widehat{\mu}_3(t)$ and its precision $\widehat{\pi}_3(t)$ are defined by the following:

$$\widehat{\mu}_3(t) = \mu_3(t-1); \tag{27}$$

$$\widehat{\pi}_3(t) = \frac{1}{\sigma_3(t-1) + \omega_3}; \tag{28}$$

Finally, the third level posterior $\mu_3(t)$ and its precision $\pi_3(t)$ are given by the following:

$$w_2(t) = \widehat{\pi}_2(t) * exp(\kappa_2 * \mu_3(t-1) * \omega_2) \tag{29}$$

$$\pi_3(t) = \widehat{\pi}_3(t) + 0.5 * \kappa_2^2 * w_2(t) * (w_2(t) + (2 * w_2(t) - 1) * \delta_2(t)) \tag{30}$$

The model prediction $\hat{y}_{Prediction}$ on the participants' predicted tilting direction of the upcoming visual stimulus is given by applying a unit sigmoid function with inverse decision temperature $\zeta = 1$ to $\widehat{\mu}_1$ as follows:

$$\hat{y}_{Prediction} = \frac{\widehat{\mu}_1{}^{\zeta}}{\widehat{\mu}_1{}^{\zeta} + (1 - \widehat{\mu}_1)^{\zeta}} \tag{31}$$

Finally, combining the two log-likelihoods of $\hat{y}_{Prediction}$ and $\hat{y}_{Perception}$ given the actual responses $y_{\text{Prediction}}$ and $y_{\text{Perception}}$ yields the modeling cost. From here, the precision of the prior distributions can be optimized via the minimization of free energy (which represents a lower bound on the log-likelihood) with regard to the predicted responses.

As an optimization algorithm, we chose the quasi-Newton Broyden-Fletcher-Goldfarb-Shanno minimization (as implemented in the HGF 4.0 toolbox). To assess the evidence for existence of the prior distributions "associative learning", "priming" and "sensory memory", their precisions were either estimated as free parameters in the perceptual model or fixed to zero (thereby effectively removing a prior distribution from the model). The precision of the prior distribution "disambiguation" was always estimated as a fixed parameter; therefore, this yielded $2^3 = 8$ models.

The prior distributions for π_a, π_p, and π_s had a mean of 0.5 and a variance of 1 when the corresponding parameter was estimated and π_a, π_p, and π_s were set to 0 when they were not estimated. π_{dis} was fixed to 1.5. Parameters from the HGF were defined as follows: μ_2, 0 = 0; μ_3, 0 = 1; σ_2, 1 = log(4.6413); σ_3, 1 = log(4); κ_2, 0 = 1; ω_2, 1 = −1.28; ω_3, 1 = −6.14; ζ, 0 = 1. Indices denote the level of the HGF.

Model inversion was performed separately for each run of the experiment and estimated models were compared using Bayesian model selection (fixed effects on the subject level and random effects on the group level) as implemented in SPM12. From the winning model, we extracted posterior parameters and averaged across runs and participants.

fMRI

Acquisition and preprocessing. We recorded BOLD images by T2-weighted gradient-echo echoplanar imaging (TR 2500 ms, TE 25 ms, voxel size 2.5 × 2.5 × 2.5 mm) on a 3T MRI scanner (Tim Trio; Siemens). The number of volumes amounted to ~1330 volumes for the main experiment and 220 volumes for the localizers. We used a T1-weighted MPRAGE sequence (voxel size 1 × 1 × 1 mm) to acquire anatomical images. Image preprocessing (slice timing with reference to the middle slice, standard realignment, coregistration, normalization to MNI stereotactic space using unified segmentation, spatial smoothing with 8 mm full-width at half-maximum isotropic Gaussian kernel) was performed with SPM12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12).

General linear models (GLMs)

Whole-brain analysis. To probe the potential neural correlates of predictive processes in the main experiment, we conducted a model-based fMRI approach using model quantities from the inverted behavioral model. Here, we aimed at disentangling the neural representation of the crossmodal "high-level" prediction $|\widehat{\mu}_2|$ from the "low-level" prediction μ_a. In addition, we considered a number of model-based regressors of no interest.

The "high-level" prediction $|\widehat{\mu}_2|$ describes the individual participants' estimate in the predictive strength of the auditory cue with regard to the visual target on a trial-by-trial basis. Due to changes in the contingencies between auditory cues and visual targets at time points unknown to the participants, such estimates in the predictive strength varied during the course of the experiment. Importantly, this quantity is orthogonal to the specific direction being predicted at a given trial.

In turn, the "low-level" prediction μ_a describes the inferred conditional probability of CW tilt given the tone, which ranges from 0 to 1. Its computation is contingent on the participants current estimate for the "high-level" prediction, whereas the two entities $|\widehat{\mu}_2|$ and μ_a are orthog-

onal to each other. This is because the conditional probabilities are defined on a stimulus-level (with regard to CW and CCW tilt), whereas the "high-level" prediction describe the strength of the overall contingencies. Importantly, because the conditional probabilities sum up to 1, the conditional probability of CCW tilt is given by $1 - \mu_a$.

Next to these quantities of interest, we considered a number of regressors to account for additional variance of no interest. Here, we included $\widehat{\pi}_2$, which represents the precision of the "high-level" prediction $\widehat{\mu}_2$ and describes how persistent a participant's belief in the audiovisual contingency is over time as well as in the light of potentially contradictory evidence. Furthermore, we took the absolute precision-weighted prediction error $|\epsilon_2|$ into account, which describes the update in the "high-level" prediction $\widehat{\mu}_2$. It is larger for unexpected visual stimuli and for situations in which the participant has an imprecise belief about the current cue–target contingency. On the level of the visual stimuli, in turn, we considered the "low-level" prediction error ε_{choice}, which is given by the difference between the actual visual outcome and the conditional probability of CW tilt. We also considered the "low-level" posterior probability of CW tilt $P(\theta_1)$, which results from the integration of the visual stimulation and the prior predictions (i.e., "associative learning", "priming", "sensory memory"). This entity predicts visual perception on a trial-by-trial basis. Last, we considered the remaining evidence for the alternative visual percept in the posterior distribution as the perceptual prediction error δ_q (for an in-depth discussion of the quantity, see also Weilnhammer et al., 2017).

The GLM contained the regressors tone and tilt, which were represented by stick functions and temporally aligned to the presentation of the auditory cue and to the onset of the tilting movement (regardless of direction or ambiguity).

Furthermore, the tone regressor was parametrically modulated by our two model quantities of interest: the crossmodal "high-level" prediction $|\widehat{\mu}_2|$ as well as the "low-level" prediction μ_a (i.e., the inferred conditional probability of CW tilt given the tone). To account for additional variance, we included the precision of the "high-level" prediction $\widehat{\pi}_2$ as a further regressor.

The tilt regressor, in turn, accounted for additional variance and was modulated by the "high-level" prediction error $|\epsilon_2|$ as well as the "low-level" perceptual posterior $P(\theta_1(t))$, the "low-level" choice prediction error ε_{choice}, and the absolute perceptual prediction error $|\delta_q|$. All model trajectories were extracted separately for each experimental run from the winning model of our Bayesian model comparison.

Regressors were convolved with the canonical hemodynamic response function as implemented in SPM12. Please note that the regressors of interest $\widehat{\mu}_2$ and μ_a were placed at the last positions of the design matrix. To ensure that our design was able to segregate between regressors of interest and regressors of no interest, we computed the colinearity between the SPM regressors and averaged across participants. The highest values of colinearity for the cue-related regressor of interest $\widehat{\mu}_2$ (i.e., the "high-level" prediction) with target-related regressors were 0.50 ± 0.02 for the "high-level" prediction error ε_2 and 0.4414 ± 0.03 for the perceptual prediction error δ_q. The highest values of collinearity for the cue-related regressor of interest μ_a (i.e., the "low-level" prediction) with target-related regressors were 0.57 ± 0.02 for the "low-level" prediction error ε_{choice} and 0.46 ± 0.04 for the perceptual posterior $P(\theta_1)$.

We added six rigid-body realignment parameters as nuisance covariates and applied high-pass filtering at 1/128 Hz. We estimated single-participant statistical parametric maps and created contrast images which were entered into voxelwise one-sample *t* tests at the group level. Anatomic labeling of cluster peaks was performed using the SPM Anatomy Toolbox Version 1.7b. We assessed our data across the whole brain reporting voxels surviving FWE correction at $p < 0.05$.

ROI analysis. We hypothesized that the "low-level" conditional stimulus probabilities would correlate with BOLD activity in retinotopic representations of the motion trajectories during CW and CCW tilt in primary visual cortex across all trials. To test this idea, we defined the correlates of the trajectories of CW and CCW tilt (which are highlighted in red and blue in Fig. 1*A*) by intersecting contrast images obtained from both the localizer and the main experiment:

From the localizer experiment, we estimated single-participant GLMs that contained box-car regressors representing the presentation of checkerboards over the upper-right and lower-left trajectories for CW tilt and lower-right and upper-right quadrant for CCW tilt and computed statistical parametric maps as well as contrast images for "CW tilt > CCW tilt" (and vice versa), thresholded at $p < 0.05$, uncorrected.

To only select voxels that were highly specific for CW and CCW tilt in the main experiment, we estimated a second set of single-subject GLMs from the main experiment, containing CW and CCW tilt for ambiguous and unambiguous trials separately, and computed single subject parametric maps as well as contrast images for "CW tilt > CCW tilt" (and vice versa) for unambiguous trials only, thresholded at $p < 0.05$, uncorrected. Please note that these contrasts are orthogonal to all predictive factors and are thus apt for the definition of functional ROIs (see also Friston et al., 2010).

ROIs were then defined by intersecting the respective contrast images for "CW tilt > CCW tilt" and "CCW tilt > CW tilt". Parameter estimation was performed using MARSBAR (marsbar.sourceforge.net/) with a design identical to whole-brain analyses. Specifically, we investigated the correlation of activity in retinotopic representations of the motion trajectories on all trials with the "low-level" prediction of tilt direction μ_a (i.e., the conditional probability of CW tilt given the tone) and $1 - \mu_a$ (i.e., the conditional probability of CCW tilt given the tone). Please note that the design matrix contained information about the posterior $P(\theta_1)$ and thus the actual sensory information. Therefore, any correlation between the BOLD signal and μ_a and $1 - \mu_a$ will be due to variance that is explained by "low-level" predictions independently of the sensory stimulation per se.

Results

Behavioral analysis

Conventional analysis

Assessing the potential effect of crossmodal predictions on perceptual decisions under ambiguity, we found that $82.61 \pm 3.87\%$ of all ambiguous trials were perceived according to the currently prevalent hidden contingency ($p < 10^{-5}$, $T = 11.4494$, one-sided test). The effect of priming on ambiguous trials ($53.62 \pm 0.98\%$; $p = 0.0013$, $T = 3.6858$) was substantially smaller, whereas conventional analyses discarded a significant impact of sensory memory on perceptual responses under ambiguity ($52.90 \pm 2.59\%$; $p = 0.2757$, $T = 1.1178$, Fig. 3*B*). As expected, $97.46 \pm 0.62\%$ of all unambiguous trials were perceived according to the disambiguation.

We proceeded by evaluating a potential mediating role of perceptual uncertainty for the influence of associative learning on perceptual decisions under ambiguity. Here, our rating experiment indicated that the majority of unambiguous trials ($67.20 \pm 5.09\%$) elicited very clear motion percepts. Such very clear motion percepts were less frequent for ambiguous trials ($31.59 \pm 6.00\%$; Fig. 3*A*). There was no significant across-subject correlation between the average perceptual certainty at ambiguous trials and the proportion of ambiguous trials perceived according to the currently prevalent hidden contingency ($\rho = 0.1238$, $p = 0.5735$).

In brief, conventional analyses indicated that the crossmodal associations significantly affected perceptual decisions under ambiguity, whereas we could not observe a relation between the strength of this effect and perceptual uncertainty.

Bayesian modeling

To infer the participants' trial-by-trial prediction about the crossmodal association and to quantify its impact for perceptual decisions under ambiguity, we conducted a Bayesian modeling approach. First, we used Bayesian model selection between models incorporating all combinations of the factors "associative learning", "prim-

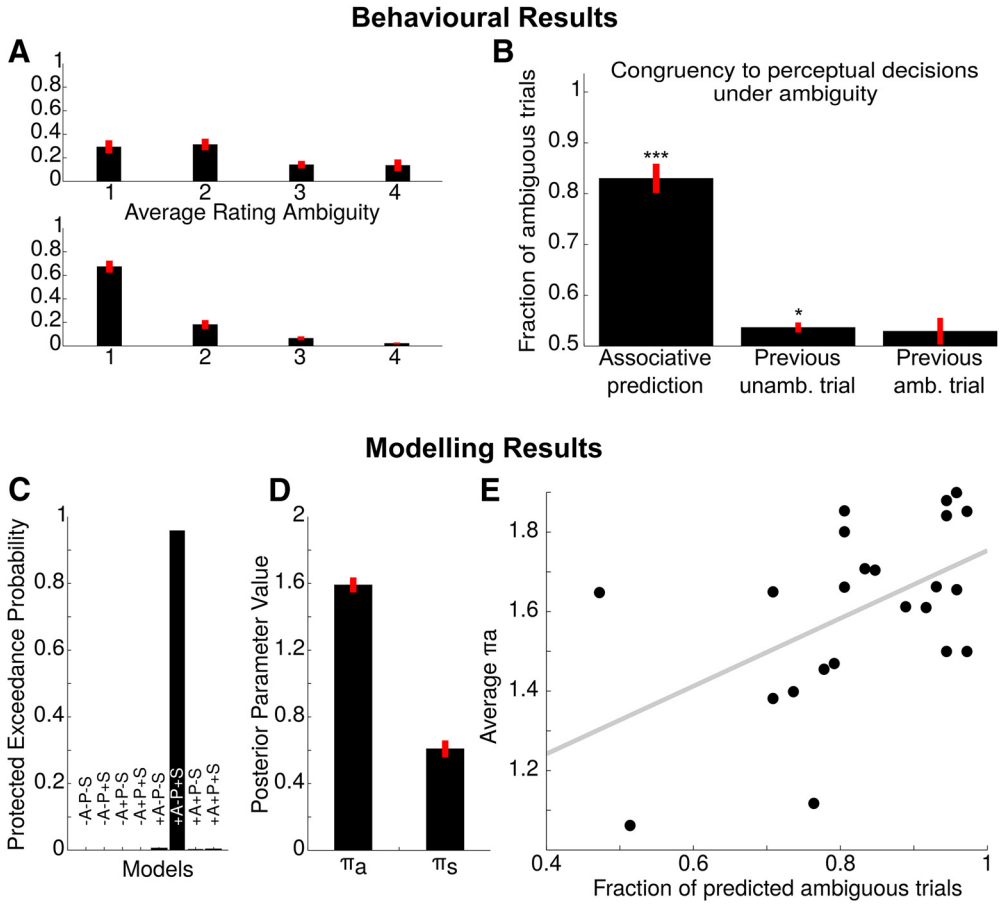


Figure 3. Behavioral analysis. ***A***, Perceptual rating. Participants tended to report a higher perceptual certainty at disambiguated trials (bottom) compared with ambiguous trials (top). ***B***, Conventional analyses. Here, we show the proportion of ambiguous trials perceived according to the current hidden contingency (associative learning, $p < 10^{-5}$, $T = 11.4494$, one-sided test), the preceding unambiguous trial (priming, $p = 0.0013$, $T = 3.6858$, one-sided ttest) and preceding ambiguous trial (sensory memory, $p = 0.2757$, $T = 1.1178$). Overall, the current cue–target association most strongly affected perceptual decisions under ambiguity, whereas the effect of priming was much smaller and conventional statistics discarded a significant impact of sensory memory. ***C***, Bayesian model comparison. Random effects Bayesian model selection indicated that the model incorporating the factors "associative learning" (+A) and "sensory memory" (+S) best explained the behavioral data collected in this experiment at a protected exceedance probability of 97.77%. This is reflected by the Bayesian model family comparison shown in the inset (A+ 99.99%, P + 2.93%, S+ 94.86% exceedance probabilities). ***D***, Posterior model parameters extracted from the winning model of our Bayesian model comparison. In analogy to conventional analysis of the contributing factors, we found a stronger influence of "associative learning" (as expressed by π_a) than for "sensory memory" (π_s). "Priming" (π_p) is not displayed because it was not part of the winning model. ***E***, Correlation between conventional metrics and inverted model quantities. The fraction of ambiguous trials perceived according to the currently prevalent hidden contingency was highly correlated with π_a ($\rho = 0.5208$, $p < 0.0108$, Pearson correlation), indicating successful model inversion. *$p < 0.05$, ***$p < 0.001$.

ing", and "sensory memory" to establish which factors were likely to affect visual perception.

Random effects Bayesian model comparison indicated evidence for an influence of the factors "associative learning" and "sensory memory" by identifying model 6 as a clear winning model at an protected exceedance probability of 97.77% (Fig. 3*C*). This is also reflected by model family comparison, which yielded clear evidence for a contribution of the factors "associative learning" (exceedance probability for associative learning models: 99.99%) and "sensory memory" (exceedance probability for sensory memory models: 94.86%), while rejecting a significant influence of priming on perceptual decisions (exceedance probability for priming models: 2.93%).

To assess the winning model on a parameter level, we extracted posterior model parameters from the perceptual model and averaged across runs and participants (Fig. 3*D*). Consistent with conventional analyses of the contributing factors, the effect (i.e., precision) of associative learning on visual outcomes (1.5862 ± 0.0607) was enhanced compared with sensory memory (0.6612 ± 0.0708; Fig. 3*D*).

Bayesian model comparison and posterior parameter estimates paralleled the results from conventional analysis by showing that the learned crossmodal association was most influential in biasing perceptual decisions at ambiguous trials, whereas the effects of perceptual history (sensory memory and priming) were estimated to be much smaller or negligible.

As an indication of successful inversion of our Bayesian model, π_a (as the metric for the strength of the impact of crossmodal associations on ambiguous trials) was highly correlated with the proportion of ambiguous trials perceived according to the currently prevalent hidden contingency ($\rho = 0.5208$, $p < 0.0108$; Fig. 3*E*). In analogy to conventional analyses, we did not observe a significant correlation between posterior HGF parameters describing the strength of the influence of associative learning on perceptual outcomes (i.e., π_a) with perceptual certainty at ambiguous trials as indicated by the independent perceptual rating experiment ($\rho = -0.0184$, $p = 0.9338$). With this, we corroborated a significant impact of predictions on perceptual decisions regardless of perceptual uncertainty and ensured successful inversion of our model.

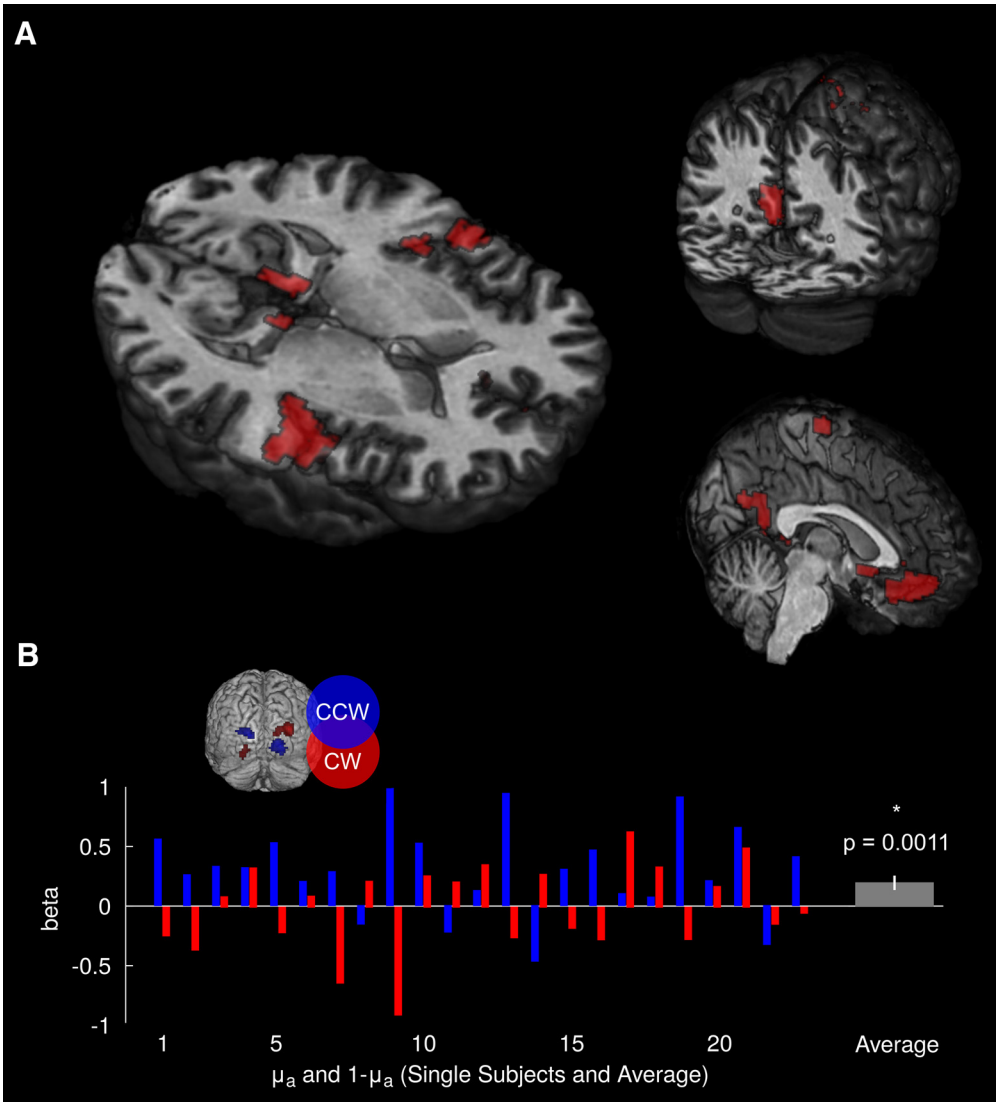


Figure 4. Imaging analyses. ***A***, Whole-brain results. The time course of the "high-level" prediction about the cue–target contingency correlated with activity in bilateral medial orbital gyrus, posterior hippocampus at the intersection to the precuneus, precuneus, postcentral gyrus, as well as right insula, precentral gyrus, posterior medial frontal gyrus, and left caudate nucleus ($p < 0.05$, FWE). ***B***, ROI-based analysis. Activity in retinotopic representation of the visual targets (i.e., the tilt trajectories for CW an CCW tilt) was related to the inferred conditional probability of CW (μ_a) and CCW ($1 - \mu_a$) at the time of cue onset ($p = 0.0011$, $t_{(23)} = 3.7585$, one-sided t test).

In brief, our behavioral analysis indicate a major influence of predictions driven by crossmodal associative learning next to a minor influence by predictions derived from perceptual history such as sensory memory and priming.

fMRI

GLMs

Whole-brain analysis. Having identified the optimal behavioral model, we aimed at identifying the neural correlates of "high-level" versus "low-level" predictions while considering additional model quantities as regressors of no interest. Because these entities served as parametric modulators for the onsets of the auditory cue and the visual target, respectively, we first mapped the contrasts "tone > baseline" and "tilt > baseline". As expected, "tone > baseline" yielded significant clusters in bilateral superior temporal gyrus, cerebellum, and thalamus as well as bilateral posterior medial frontal gyrus, left precentral gyrus, and insula, whereas the contrast "tilt > baseline" showed significant activations in bilateral inferior and middle occipital gyrus, right inferior temporal gyrus (V5/hMT+), bilateral putamen, and right lingual gyrus, as well as bilateral fusiform gyrus (FWE, $p < 0.05$).

For the main focus of whole-brain analysis, we found that the "high-level" cross modal prediction $|\widehat{\mu_2}|$ correlated with activity in supramodal brain regions such as bilateral middle orbital gyrus, bilateral rolandic operculum, bilateral Heschl's gyrus, right superior medial frontal gyrus, left caudate nucleus, bilateral postcentral gyrus, right precentral gyrus and right insula. Moreover, $|\widehat{\mu_2}|$ was also associated with activity in bilateral posterior hippocampus at the intersection to the precuneus and bilateral precuneus (Fig. 4*A*, $p < 0.05$ FWE).

In contrast, "low-level" predictions (i.e., μ_a) were not significantly related to activity in any region of the brain when applying the same rigorous threshold ($p < 0.05$, FWE). However, consistent with the results of the ROI analyses described below, "low-level" predictions as expressed by μ_a correlated with activity in occipital cortex at a more liberal statistical threshold ($p < 0.001$, uncorrected).

The remaining parametric regressors ($\hat{\pi}_2$, $|\epsilon_2|$, $\varepsilon_{\text{choice}}$, $|\delta_q|$ and $P(\theta_1)$) were added to the GLM to account for additional variance in the BOLD signal and corroborated previous neuroimaging results. The stability of the "high-level" prediction $\widehat{\pi}_2$, which determines how stable a given "high-level" prediction is over time, correlated with activity right Heschl's gyrus as well as left insula and right postcentral gyrus ($p < 0.05$, FWE). Further explorative analyses indicated that the "high-level" precision-weighted contingency prediction errors ($|\epsilon_2|$) correlated with activity in posterior medial frontal cortex, right middle frontal gyrus, inferior parietal lobulus, left insula, and right caudate nucleus ($p < 0.05$, FWE), which overlaps with results from Iglesias et al. (2013).

In turn, "low-level" perceptual prediction errors ($|\delta_q|$) were associated with BOLD activity in areas such as left insula, right precentral gyrus, and left middle temporal gyrus ($p < 0.05$, FWE), which is consistent with results from Weilnhammer et al. (2017). As expected, "low-level" choice prediction errors ($\varepsilon_{\text{choice}}$) and the posterior probability of CW tilt $P(\theta_1)$ were associated with activity in occipital cortex.

ROI-based analysis. We furthermore examined how BOLD responses in retinotopic representations of motion trajectories of CW and CCW tilt across all trials would relate to conditional probabilities of the visual targets as defined by the inverted behavioral model. To account for interindividual variability in the retinotopic organization of visual cortex, our approach was based on ROIs that were functionally defined for each individual. As predicted, we found that the "low-level" prediction parametrized by the conditional probabilities of CW tilt μ_a and CCW tilt $1 - \mu_a$ were significantly correlated with BOLD time courses in voxels corresponding to the respective trajectories of CW and CCW tilt ($p = 0.0011$, $t_{(23)} = 3.7585$, one-sided one-sample t test).

The "high-level" prediction $|\widehat{\mu}_2|$ was not related to BOLD time courses in retinotopic stimulus representations. In an explorative analysis, we found that the posterior probabilities of CW tilt $P(\theta_1)$ and CCW tilt $P(\theta_0) = 1 - P(\theta_1)$ were related to activity in voxels corresponding to the respective trajectories of CW and CCW tilt ($p < 10^{-7}$, $t_{(14)} = 9.0292$, two-sided one-sample t test). This result is expected given that this posterior also contains information of the sensory stimulation per se (CW tilt or CCW tilt). When assessing the remaining parameters of our GLM as a negative control, we did not find any significant correlation to retinotopic BOLD data for the choice prediction error $\varepsilon_{\text{choice}}$, the absolute perceptual prediction error $|\delta_q|$, or the absolute "high-level" prediction error $|\epsilon_2|$.

In sum, ROI-based analyses indicated that primary visual cortex implements "low-level" predictions encoding conditional visual stimulus probabilities as opposed to "high-level" predictions encoding crossmodal cue–stimulus associations.

Discussion

In this work, we studied the neural correlates of dynamically updated prior predictions and their effect on perceptual decisions in a crossmodal associative learning experiment. Crucially, this task required participants to engage in hierarchical learning to represent both the dynamically changing strength of cue–target associations as well as conditional target probabilities given a specific cue. Due to the existence of covertly interspersed ambiguous trials, our paradigm enabled us to study processes involved in perceptual inference with regard to the combination of sensory information with conditional target probabilities and prior influences from perceptual history such as priming and sensory memory. Thereby, our paradigm afforded the dissociation between "high-level" predictions about the strength of cue–target associations and "low-level" predictions about both the conditional probability of the binary visual outcome.

Conventional and model-based behavioral analyses indicated that participants successfully engaged in hierarchical associative learning. Here, perceptual decisions under ambiguity were strongly biased by changing cue–target associations. This is consistent with our previous results from an analogous behavioral experiment using ambiguous structure-from-motion spheres (Schmack et al., 2016). Both in the current and in the previous study, individual perceptual uncertainty ratings of ambiguous stimuli were not correlated to the size of the impact introduced by crossmodal associative learning. To our minds, this is most likely because the ambiguous trials elicited bistable perception while participants did not have metacognitive access to the ambiguity of the visual stimuli.

However, there is an ongoing debate about the interaction of bistable perception and perceptual uncertainty (Knapen et al., 2011). Strikingly, in the current version of the experiment using ambiguous apparent motion stimuli, the impact of associative learning was substantially greater than in our previous study using ambiguous spheres (Schmack et al., 2016). This intended difference might arise because the stimulus interpretations induced by apparent motion in our present experiment were characterized by lower perceptual certainty compared with ambiguous spheres and might thus be more susceptible to prior predictions. We believe that future studies are needed to investigate how perceptual decisions under ambiguity and their modulation by prior predictions might interact with differing levels of perceptual uncertainty.

Although conventional statistics did not show evidence for a significant contribution of sensory memory to perceptual decisions under ambiguity, the winning model from Bayesian model comparison statistics incorporated a minor impact of the factor sensory memory. This discrepancy is most likely to be caused by differences in the statistical approaches. In Bayesian analysis, the factor sensory memory is embedded within a generative model and evaluated in terms of protected exceedance probability, whereas conventional statistics look at all factors in isolation.

Importantly, our model-based fMRI results indicate that "high-level" predictions are related to activity in supra-modal brain areas such as middle orbital gyrus, insula, posterior medial frontal gyrus, postcentral gyrus, as well as the posterior hippocampus extending into the precuneus. These findings suggest that activity in such regions tracks an individual participant's trial-by-trial belief in the strength of the cue–target association. In the context of the present experiment, our results suggest that activity in these brain areas may determine the stability over time of learned associations between auditory cues and visual targets.

Therefore, increased activity in these areas reflects a currently strong "high-level" prediction. In this case, the participant strongly relies on past experiences for the prediction of future outcomes. Furthermore, an unexpected visual outcome is rather attributed to the inherent stochasticity of the experiment, that is, expected uncertainty, and has therefore relatively little effect on the currently assumed cue–outcome contingency. In contrast, decreased activity in these brain areas reflects a currently weak "high-level" prediction. In this case, the participant is unsure about the prevalent cue–outcome contingency and therefore only weakly relies on past experiences for the prediction of future outcomes. Furthermore, unexpected visual outcomes have a relatively strong affect the assumed cue–target contingency.

In more general terms, our results suggest that activity in regions such as middle orbital gyrus, insula, posterior medial frontal gyrus, postcentral gyrus, and posterior hippocampus encode

the strength of an agents belief in the statistical dependencies within the environment. With regard to the example of a badminton game, this would translate to how strongly the current wind condition is believed to be stable and therefore taken into account when estimating the trajectory of the shuttlecock.

The encoding of the "high-level" prediction in these regions is consistent with results from closely related experiments on unexpected and expected uncertainty (Payzan-LeNestour et al., 2013). Here, the probability of a change in the statistical properties of the experimental environment (i.e., the negative "high-level" prediction) was negatively correlated with activity in left insula, bilateral postcentral gyrus, left hippocampus, as well as posterior cingulate cortex and left middle temporal gyrus. Furthermore, placebo experiments related activity in orbitofrontal cortex to the build-up and maintenance of predictions regarding sensory outcomes (Petrovic et al., 2002; Wager et al., 2004). Finally, a recent experiment using behavioral modeling and muscimol inactivation in rats has revealed a potential implication of both the orbitofrontal cortex as well as the dorsal hippocampus in model-based planning (Miller et al., 2017). This is interesting because behavior associated with model-based planning relates to relying on a "high-level" prediction about the statistical properties of the environment.

Another important functional aspect of brain areas coding for the "high-level" prediction could be the instantiation of the effect of predictions on sensory processing through feedback processes. Consistent with our results, regions in the orbitofrontal cortex have repeatedly been discussed as mediators for the effect of predictions on sensory processing (Bar et al., 2006; Kveraga et al., 2007; Summerfield and Koechlin, 2008). Moreover, studies on the role of predictions for perceptual inference in healthy participants and patients with paranoid schizophrenia have highlighted the impact of feedback processes from orbitofrontal cortex to sensory areas on the modulation of perceptual decisions under ambiguity by prior knowledge (Schmack et al., 2013, 2017).

In contrast to "high-level" predictions about the strength of the association between cue and target, "low-level" predictions about the conditional probabilities of binary perceptual outcomes at the time of cue presentation were reflected by retinotopic representations of the visual stimulus. This finding provides a potential neural correlate for the influence of predictions on perceptual decisions. One might speculate that this phenomenon is mediated by similar feedback mechanisms as those involved in spatial- or feature-based attention, which are known to modulate brain activity in primary visual cortex (Gandhi et al., 1999; Posner and Gilbert, 1999).

In relation to work by Iglesias et al. (2013), who focused on hierarchical precision-weighted prediction errors, our study extends these findings by looking more closely at the neural correlates of hierarchical predictions, which are key elements of hierarchical predictive coding schemes. The computation of conditional target probabilities represented in primary visual cortex is contingent on the inferred cue–target association reflected by activity in regions such as the orbitofrontal cortex, hippocampus, and precuneus. This suggests an interplay between "high-level" and "low-level" regions in human cortex via feedback connections, which might mediate the influence of prior knowledge on perceptual decisions. Therefore, the aforementioned regions and the effective connectivity between them will be interesting targets for the investigation of aberrant predictive processes in neuropsychiatric disorders such as schizophrenia (Adams et al., 2013; Powers et al., 2017).

Together, our results suggest that observers flexibly use dynamic predictions derived from hierarchical associative learning adapted to a volatile environment to perform perceptual inference. Our imaging analyses indicate that "high-level" predictions about cue–target associations are represented in supramodal brain regions such as orbitofrontal cortex and hippocampus, whereas "low-level" conditional target probabilities are associated with activity in primary visual areas, providing a potential neural correlate for the influence of prior knowledge on perceptual decisions.

References

Adams RA, Stephan KE, Brown HR, Frith CD, Friston KJ (2013) The computational anatomy of psychosis. Front Psychiatry 4:47. CrossRef Medline

Bar M, Kassam KS, Ghuman AS, Boshyan J, Schmid AM, Schmidt AM, Dale AM, Hämäläinen MS, Marinkovic K, Schacter DL, Rosen BR, Halgren E (2006) Top-down facilitation of visual recognition. Proc Natl Acad Sci U S A 103:449–454. CrossRef Medline

Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. Nat Neurosci 10:1214–1221. CrossRef Medline

Clark A (2013) Whatever next? predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci 36:181–204. CrossRef Medline

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360:815–836. CrossRef Medline

Friston KJ, Rotshtein P, Geng JJ, Sterzer P, Henson RN (2010) A critique of functional localizers. In: Foundational issues in human brain mapping (Hanson SJ and Bunzl M, eds), pp 3–24. Cambridge, MA: MIT.

Gandhi SP, Heeger DJ, Boynton GM (1999) Spatial attention affects brain activity in human primary visual cortex. Proc Natl Acad Sci U S A 96:3314–3319. CrossRef Medline

Hohwy J, Roepstorff A, Friston K (2008) Predictive coding explains binocular rivalry: an epistemological review. Cognition 108:687–701. CrossRef Medline

Iglesias S, Mathys C, Brodersen KH, Kasper L, Piccirelli M, den Ouden HE, Stephan KE (2013) Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. Neuron 80:519–530. CrossRef Medline

Knapen T, Brascamp J, Pearson J, van Ee R, Blake R (2011) The role of frontal and parietal brain areas in bistable perception. J Neurosci 31:10293–10301. CrossRef Medline

Knill DC, Pouget A (2004) The {Bayesian} brain: the role of uncertainty in neural coding and computation. Trends Neurosci 27:712–719. CrossRef Medline

Kveraga K, Ghuman AS, Bar M (2007) Top-down predictions in the cognitive brain. Brain Cogn 65:145–168. CrossRef Medline

Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A Opt Image Sci Vis 20:1434–1448. CrossRef Medline

Mathys CD, Lomakina EI, Daunizeau J, Iglesias S, Brodersen KH, Friston KJ, Stephan KE (2014a) Uncertainty in perception and the hierarchical Gaussian filter. Front Hum Neurosci 8:825. CrossRef Medline

Mathys CD, Lomakina EI, Daunizeau J, Iglesias S, Brodersen KH, Friston KJ, Stephan KE (2014b) Uncertainty in perception and the hierarchical Gaussian filter. Front Hum Neurosci 8:825. CrossRef

Mathys C, Daunizeau J, Friston KJ, Stephan KE (2011) A Bayesian foundation for individual learning under uncertainty. Front Hum Neurosci 5:39. CrossRef Medline

Miller KJ, Botvinick MM, Brody CD (2017) Dorsal hippocampus contributes to model-based planning. Nat Neurosci 20:1269–1276. CrossRef Medline

Muckli L, Kohler A, Kriegeskorte N, Singer W (2005) Primary visual cortex activity along the apparent-motion trace reflects illusory perception. PLoS Biol 3:e265. CrossRef Medline

Nassar MR, Wilson RC, Heasly B, Gold JI (2010) An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. J Neurosci 30:12366–12378. CrossRef Medline

Payzan-LeNestour E, Dunne S, Bossaerts P, O'Doherty JP (2013) The neural representation of unexpected uncertainty during value-based decision making. Neuron 79:191–201. CrossRef Medline

Pearson J, Brascamp J (2008) Sensory memory for ambiguous vision. Trends Cogn Sci 12:334–341. CrossRef Medline

Petrovic P, Kalso E, Petersson KM, Ingvar M (2002) Placebo and opioid

analgesia: imaging a shared neuronal network. Science 295:1737–1740. CrossRef Medline

Posner MI, Gilbert CD (1999) Attention and primary visual cortex. Proc Natl Acad Sci U S A 96:2585–2587. CrossRef Medline

Powers AR, Mathys C, Corlett PR (2017) Pavlovian conditioninginduced hallucinations result from overweighting of perceptual priors. Science 357:596–600. CrossRef Medline

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87. CrossRef Medline

Schmack K, Gòmez-Carrillo de Castro A, Rothkirch M, Sekutowicz M, Rössler H, Haynes JD, Heinz A, Petrovic P, Sterzer P (2013) Delusions and the role of beliefs in perceptual inference. J Neurosci 33:13701–13712. CrossRef Medline

Schmack K, Weilnhammer V, Heinzle J, Stephan KE, Sterzer P (2016) Learning what to see in a changing world. Front Hum Neurosci 10:263. CrossRef Medline

Schmack K, Rothkirch M, Priller J, Sterzer P (2017) Enhanced predictive signalling in schizophrenia. Hum Brain Mapp 38:1767–1779. CrossRef Medline

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. Neuroimage 46:1004–1017. CrossRef Medline

Sterzer P, Kleinschmidt A (2007) A neural basis for inference in perceptual ambiguity. Proc Natl Acad Sci U S A 104:323–328. CrossRef Medline

Sterzer P, Haynes JD, Rees G (2006) Primary visual cortex activation on the path of apparent motion is mediated by feedback from hMT+/V5. Neuroimage 32:1308–1316. CrossRef Medline

Summerfield C, Koechlin E (2008) A neural representation of prior information during perceptual inference. Neuron 59:336–347. CrossRef Medline

Wager TD, Rilling JK, Smith EE, Sokolik A, Casey KL, Davidson RJ, Kosslyn SM, Rose RM, Cohen JD (2004) Placebo-induced changes in fMRI in the anticipation and experience of pain. Science 303:1162–1167. CrossRef Medline

Weilnhammer V, Stuke H, Hesselmann G, Sterzer P, Schmack K (2017) A predictive coding account of bistable perception: a model-based fMRI study. PLOS Comput Biol 13:e1005536. CrossRef Medline

Yu AJ, Dayan P (2005) Uncertainty, neuromodulation, and attention. Neuron 46:681–692. CrossRef Medline

2.3 Psychotic experiences in schizophrenia and sensitivity to sensory evidence

Weilnhammer VA, Röd L, Eckert A, Stuke H, Sterzer. Schizophrenia bulletin 46, 927–936 (2020). DOI: https://doi.org/10.1093/schbul/sbaa003

The above publications characterize bistable perception in terms of the interplay between internal predictions that are generated by learning[50] and prediction errors that are driven by external sensory information[80]. Importantly, previous research has proposed that imbalances between predictions and prediction errors may contribute to the experience of psychotic symptoms[40–42]. In this study, patients suffering from paranoid schizophrenia and matched healthy controls indicated their conscious experience at varying level of signal-to-ambiguity in partially disambiguated Lissajous figures (Figure 2C). Relative to controls, patients were more sensitive to prediction errors driven by external sensory information. The sensitivity toward prediction errors correlated with the severity of hallucinatory experiences in patients. These results therefore argues in favor of the *weak prior* account hallucinations (Figure 1C).

The following text corresponds to the abstract of the article[43]:

"Perceptual inference depends on an optimal integration of current sensory evidence with prior beliefs about the environment. Alterations of this process have been related to the emergence of positive symptoms in schizophrenia. However, it has remained unclear whether delusions and hallucinations arise from an increased or decreased weighting of prior beliefs relative to sensory evidence. To investigate the relation of this prior-to-likelihood ratio to positive symptoms in schizophrenia, we devised a novel experimental paradigm which gradually manipulates perceptually ambiguous visual stimuli by disambiguating stimulus information. As a proxy for likelihood precision, we assessed the sensitivity of individual participants to sensory evidence. As a surrogate for the precision of prior beliefs in perceptual stability, we measured phase duration in ambiguity. Relative to healthy controls, patients with schizophrenia showed a stronger increment in congruent perceptual states for increasing levels of disambiguating stimulus evidence. Sensitivity to sensory evidence correlated positively with the individual patients' severity of perceptual anomalies and hallucinations. Moreover, the severity of such experiences correlated negatively with phase duration. Our results indicate that perceptual anomalies and hallucinations are associated with a shift of perceptual inference toward sensory evidence and away from prior beliefs. This reduced prior-to-likelihood ratio in sensory processing may contribute to the phenomenon of aberrant salience, which has been suggested to give rise to the false inferences underlying psychotic experiences."

Schizophrenia Bulletin vol. 46 no. 4 pp. 927–936, 2020
doi:10.1093/schbul/sbaa003
Advance Access publication 24 February 2020

Psychotic Experiences in Schizophrenia and Sensitivity to Sensory Evidence

Veith Weilnhammer*[,1], Lukas Röd[2], Anna-Lena Eckert[1,3,4], Heiner Stuke[1], Andreas Heinz[1,3,4], and Philipp Sterzer[1–4]

[1]Department of Psychiatry, Charité Universitätsmedizin Berlin, 10117 Berlin, Germany; [2]Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, 10099 Berlin, Germany; [3]Bernstein Center for Computational Neuroscience, Charité Universitätsmedizin Berlin, 10117 Berlin, Germany; [4]Einstein Center for Neurosciences Berlin, Charite Universitatsmedizin Berlin, 10117 Berlin, Germany

*To whom correspondence should be addressed; Department of Psychiatry, Charité Campus Mitte, Charitéplatz 1, 10117 Berlin, tel: 0049-(0)30-450-517-317, fax: 0049-(0)30-450-517-944, e-mail: veith-andreas.weilnhammer@charite.de

Perceptual inference depends on an optimal integration of current sensory evidence with prior beliefs about the environment. Alterations of this process have been related to the emergence of positive symptoms in schizophrenia. However, it has remained unclear whether delusions and hallucinations arise from an increased or decreased weighting of prior beliefs relative to sensory evidence. To investigate the relation of this prior-to-likelihood ratio to positive symptoms in schizophrenia, we devised a novel experimental paradigm which gradually manipulates perceptually ambiguous visual stimuli by disambiguating stimulus information. As a proxy for likelihood precision, we assessed the sensitivity of individual participants to sensory evidence. As a surrogate for the precision of prior beliefs in perceptual stability, we measured phase duration in ambiguity. Relative to healthy controls, patients with schizophrenia showed a stronger increment in congruent perceptual states for increasing levels of disambiguating stimulus evidence. Sensitivity to sensory evidence correlated positively with the individual patients' severity of perceptual anomalies and hallucinations. Moreover, the severity of such experiences correlated negatively with phase duration. Our results indicate that perceptual anomalies and hallucinations are associated with a shift of perceptual inference toward sensory evidence and away from prior beliefs. This reduced prior-to-likelihood ratio in sensory processing may contribute to the phenomenon of aberrant salience, which has been suggested to give rise to the false inferences underlying psychotic experiences.

Key words: psychosis/Bayesian perceptual inference/predictive coding/bistable perception

Introduction

When perceiving our surroundings, we are confined to inherently noisy and ambiguous sensory representations of the environment. However, conscious experience usually provides us with an unequivocal impression of our world. According to Bayesian theories,[1–3] our brain bridges this gap by actively employing beliefs to interpret sensory information and forms a hypothesis (or *posterior* probability distribution, figure 1A) about the cause of current sensory data.[4] Along this line of thought, conscious experience represents a *controlled hallucination*, that is concurrently being shaped by internally generated beliefs (*prior* distributions) and constrained by external sensory information (the *likelihood* distribution).[5]

Alterations in the relative weighting (or *precision*[6]) of prior and likelihood may lead to false (or dysfunctional) inferences[7–9]: If prior precision is overestimated relative to the likelihood (increased prior-to-likelihood ratio, figure 1B), inference will be driven too strongly by prior beliefs and violations of prior beliefs by sensory data (ie, *prediction errors*) will be overly attenuated. In contrast, a decreased prior-to-likelihood ratio (figure 2C) will lead to a stronger weighting of the sensory data, thus instigating aberrant prediction errors.

Previous work has discussed both increases and decreases of the prior-to-likelihood ratio in relation to cognitive and perceptual anomalies in psychosis-prone individuals and patients with schizophrenia (Scz, for review, see[10] and[11]). Interestingly, delusions have often been related to a decreased prior-to-likelihood ratio,[8,12–16] whereas studies on hallucinations have pointed to an increased prior-to-likelihood ratio.[17–22] As it seems unlikely that delusions and hallucinations, 2 frequently co-occurring symptom domains, should be due to opposing alterations in inference, it was recently proposed that these apparently contradictory findings may be reconciled within the framework of hierarchical predictive coding[1,2,23]: The prior-to-likelihood ratio may indeed be generally reduced at low levels, eg, in early sensory areas, leading to aberrant salience of sensory stimuli and the emergence of delusions.[24,25]

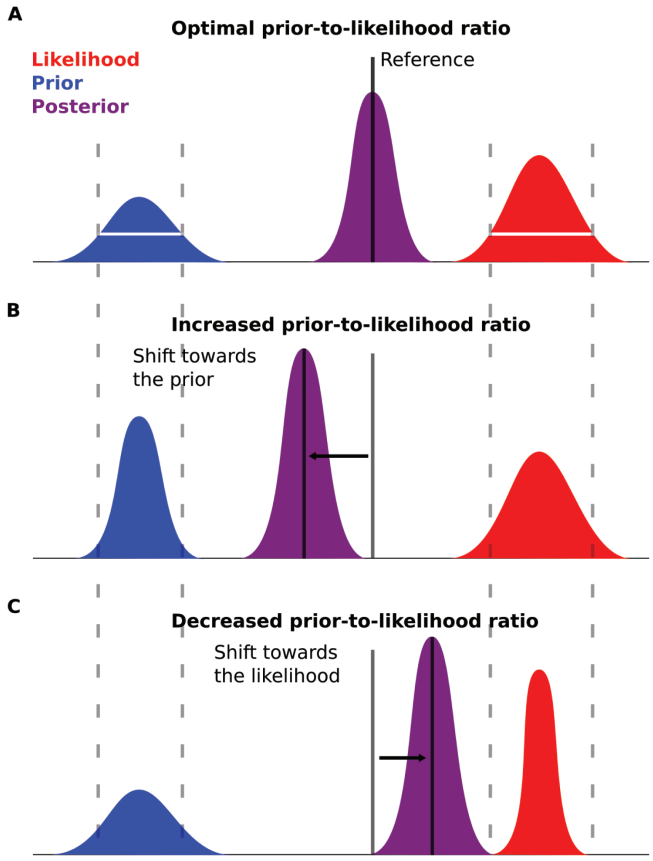


Fig. 1. The prior-to-likelihood ratio in Bayesian perceptual inference. Perceptual inference depends on the ratio of prior and likelihood precision. (**A**) Here, we depict a reference scenario with optimal precision estimates (Gaussian distributions, variance in white, mean of the posterior in black). (**B**) Changes in these estimates of precision may lead to alterations in perception. In case of an overestimation of prior precision and/or underestimation of likelihood precision, the posterior is shifted toward the prior. (**C**) By analogy, an overestimation of likelihood precision and/or underestimation of prior precision is associated with a shift of the posterior toward the likelihood.

In contrast, higher-level priors may become overly precise in an attempt to compensate for aberrant salience and contribute to the emergence of hallucinations.[10,11,26]

In the present study, we tested the hypothesis that psychotic experiences in Scz are related to a decreased prior-to-likelihood ratio at low hierarchical levels. We asked whether the precision of the likelihood mapping between the causes of sensations and the sensory consequences was elevated in Scz relative to healthy controls. This precision is often referred to as sensory precision, where an elevated precision is sometimes attributed to a failure of sensory attenuation. Moreover, we tested whether such a stronger weighting of sensory evidence is associated with the experience of delusions, hallucinations, or both.

We developed a novel experimental paradigm based on bistable perception, ie, the spontaneous alternation between 2 perceptual states that occurs when sensory information is ambiguous.[27] Predictive coding posits that the dynamics of bistability reflect the 2 components of the prior-to-likelihood ratio[28,29]: The current perceptual state represents the best hypothesis (ie, the prior) about the cause of sensory information (ie, the likelihood). Due to ambiguity, neither of the 2 mutually exclusive perceptual hypotheses can fully account for the sensory data. Hence, a prediction error accumulates and eventually leads to a perceptual transition.

Here, we induced the phenomenon of *graded ambiguity* by parametrically manipulating the available sensory evidence for the 2 alternative perceptual hypotheses of an ambiguous Lissajous figure (see figure 2A and Supplementary Video 1). When a perceptual hypothesis is congruent to disambiguating stimulus evidence, prediction errors should be reduced and perceptual transitions to the incongruent perceptual states less likely. Incongruence, in turn, should lead to enhanced prediction errors and increased probability of a transition to the congruent perceptual state. In sum, the probability of perceptual states congruent with disambiguating stimulus evidence should vary with the individual participants' sensitivity to sensory evidence. Thus, it serves as a proxy for the prior-to-likelihood ratio.

We studied the sensitivity to disambiguating stimulus evidence in patients with paranoid Scz and a matched control group. Under the assumption of a decreased prior-to-likelihood ratio in psychosis, we expected an increased sensitivity to disambiguating stimulus evidence in patients with Scz. We furthermore hypothesized a positive correlation of sensitivity to disambiguating stimulus evidence with the severity of delusions and hallucinations.

Methods

Participants

We excluded 1 control due to impaired stereovision, 3 controls due to elevated scores for Cardiff Anomalous Perception Scale (CAPS) and Peters Delusion Inventory (PDI) (threshold/scores ≥ 3 SDs above the group's mean), 1 control due to reduced frequency of congruent perceptual states (frequency ≤ 3 SDs below the mean computed across groups in any of the conditions D1–D7), and 1 patient who did not complete the experiment. The final sample was matched for gender, age, and handedness (see table 1) and consisted of 23 patients (International Classification of Diseases 10: F20.0, 18 male, age = 37.13 ± 2.42) recruited from in- and out-patient services at Charité Universitätsmedizin Berlin and 23 control participants (17 male, age = 33.57 ± 1.74 y). All participants had (corrected-to-)normal vision, were naive to the purpose of the study, and gave informed, written consent prior to the experiment authorized by the Charité Ethics Committee.

Questionnaires and Clinical Rating

Participants completed the 40-item PDI[30] to quantify delusional ideation[13,14,17,31–33] and the 32-item CAPS[34] to

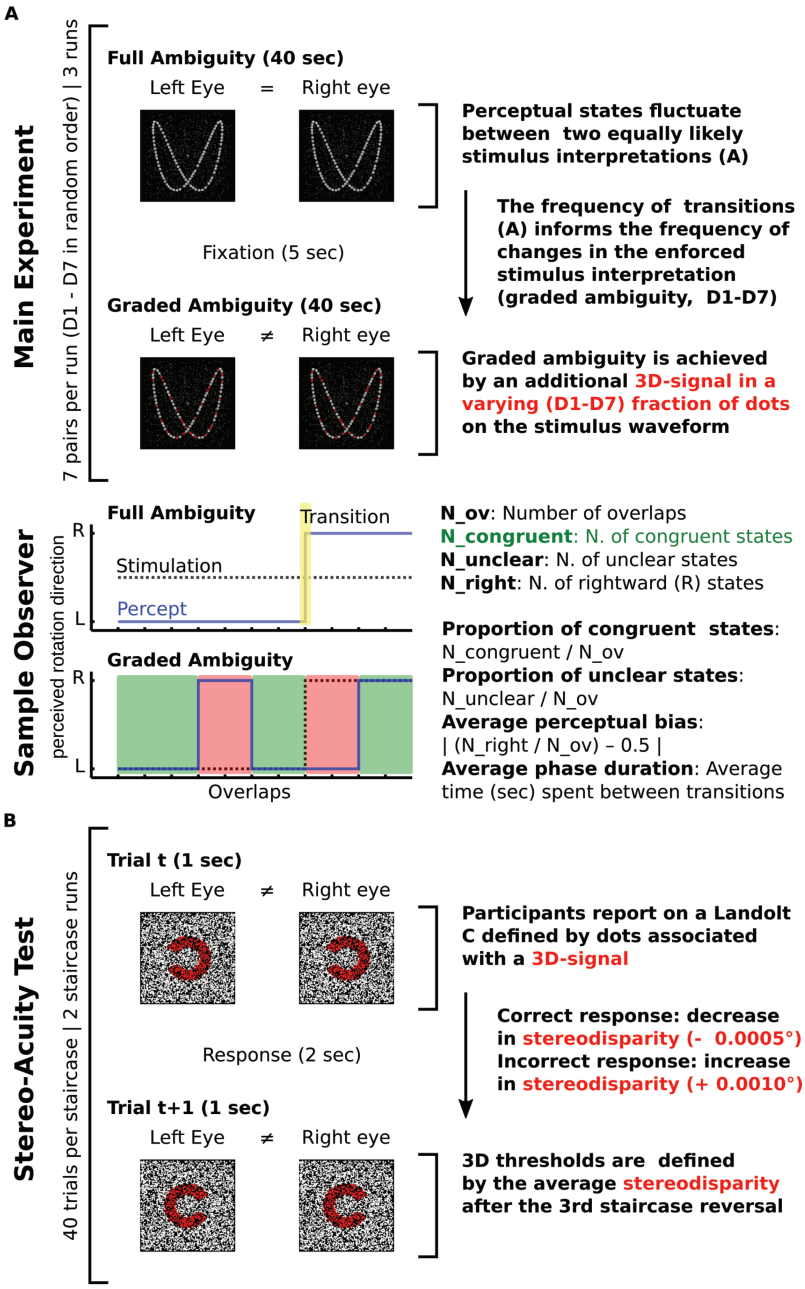


Fig. 2. Behavioral experiment. **(A)** In the main experiment, we measured the individual participants' sensitivity to disambiguating stimulus evidence as a proxy for the prior-to-likelihood ratio. To visualize relevant variables, the lower panel displays typical perceptual responses in an ambiguous block and the corresponding partially disambiguated block. (B) To probe potential differences in stereovision, we determined individual stereo-disparity thresholds in an independent stereoacuity test.

Table 1. Sample Characteristics

Group	*N*	Female	Smoking	Stat	Age	ED	CAPS	PDI	PANSS: P	*N*	G	DOI	CPZe
				Mean	33.6	77	6.7	22	NA	NA	NA	NA	NA
Controls	23	6	10	SD	8.4	40	9.2	28	NA	NA	NA	NA	NA
				Mean	37.1	75	65.0	139	18.4	19.4	33	15	190
Patients	23	5	15	SD	11.6	44	50.1	80	6.3	8.2	10	12	172

Note: Patients with Scz scored higher than controls on the PDI (patients: 138.83 ± 16.64 SEM, controls: 21.87 ± 5.75, Welch 2-sample *t*-test: $T(27) = 6.64$, $P = 3.81 \times 10^{-7}$) and CAPS (patients: 64.96 ± 10.45, controls: CAPS of 6.65 ± 1.91, $T(23) = 5.49$, $P = 1.32 \times 10^{-5}$). One patient received a typical antipsychotic, 18 patients were prescribed an atypical antipsychotic, and 4 were without medication.

measure perceptual anomalies. Reported scores reflect sums over questionnaire subscales. We assessed clinical symptom severity using the Positive and Negative Syndrome Scale (PANSS).[35]

Behavioral Experiments

Apparatus. We presented all stimuli using a mirror stereoscope placed in front of a 98PDF-CRT-Monitor (60 Hz, 1042 × 768 pixels, 59.50 cm viewing distance, 30.28 pixels per degree visual angle; °) using Psychtoolbox 3[36] and Matlab R2007b (MathWorks).

Main Experiment. The main experiment (figure 2A) assessed the modulation of perceptual states by levels of disambiguating stimulus evidence. In 3 runs (10.52 min each), participants viewed 7 pairs of ambiguous and partially disambiguated versions of a rotating discontinuous Lissajous figure (see Supplementary Video 1) presented in blocks of 40.08 s each, separated by 5 s of fixation. We randomly placed 300 dots (0.05°) on the stimulus waveform (2.05° × 2.05°) defined by the perpendicular intersection of 2 sinusoids [$x(t) = sin(A * t)$ and $y(t) = cos(B * t + \delta)$ with $A = 3$, $B = 6$, and δ increasing from 0 to 2π at 6.80 s per revolution and 6 revolutions per block]. We relocated the dots at a probability of 0.02 per frame. Stimuli were surrounded by rectangular fusion frames and presented on the background of random-dot noise (700 dots of 0.05°, 1.98°/s speed, changes in motion direction at 1 Hz). We displayed a fixation cross in the center of the visible screen (0.10°).

During ambiguous blocks, we presented identical Lissajous figures to the 2 eyes. Participants indicated changes in the perceived direction of rotation by pressing the left (rotation of the front surface to the left, right index finger), right (rotation to the right, right ring finger), or down (unclear direction of rotation, right middle finger) arrow key on a standard USB keyboard.

The indicated direction of rotation in an ambiguous block determined the time-points of changes in sensory evidence in the upcoming disambiguated block. To add additional sensory evidence (graded disambiguation) to the Lissajous figure, we shifted a proportion of the stimulus dots by a δ of 0.02π in the corresponding direction between monocular channels. Crucially, we varied the amount of disambiguating stimulus evidence across 7 conditions (D1: 1.25%, D2: 3.75%, D3: 8.75%, D4: 16.25%, D5: 26.25%, D6: 50.00%, and D7: 100.00% of dots disambiguated). Each condition appeared once per run and in random order. Participants reported changes in the perceived direction of rotation as well as unclear perceptual states.

Stereoacuity. We assessed stereo-disparity thresholds in an independent stereoacuity test (similar to[37], figure 2B). To this end, we presented a number of 5000 dots (each at 0.15°) within a square of 11 × 11°. We attached a stereo-disparity signal to dots lying on a Landolt C, ie, a circle (1.37° radius, 2.06° width) with a 90° gap located at the left, top, right, or bottom. Following 5 s of fixation and 1 s of stimulus presentation, participants reported the location of the gap in the Landolt C by pressing the up-, down-, left-, or right-arrow key (response interval = 2 s). Fixation crosses (0.10°) were presented in the center of visible screen.

Participants performed 2 runs of 40 trials each. At each trial, we determined the amount of presented stereo disparity based on the response from the previous trial by a 2-up-1-down staircase procedure (correct response: decrease in the available stereo disparity by 1 step; incorrect response: increase by 2 steps, initial step size: 0.001°, reduction to 0.0005° after first reversal). The initial stereo disparity was 0.0045° in run 1 and 0.0005° in run 2.

Analyses

Main Experiment. For the main experiment, we based our analyses on perceptual transitions reported by the participants. Because perceptual transitions occur at overlapping configurations of the Lissajous figure,[29,38–41] we corrected the timing of each perceptual transition to the time of the overlap preceding the corresponding button press. This decomposed the perceptual time course into a sequence of discrete perceptual states (leftward, rightward, and unclear rotation of the front surface, 3.40 s inter-overlap interval).

As variable-of-interest (see figure 2A), we computed the proportion of congruent perceptual states (ie, perceptual states perceived in congruence with the disambiguating stimulus evidence) for all parametric levels of disambiguation (D1–D7). This variable served as a proxy for the prior-to-likelihood balance during graded ambiguity. In addition, we determined individual perceptual stability in terms of average phase duration (ie, time spent between 2 perceptual transitions). As potential confounds, we computed the probability of unclear perceptual states for all conditions (ambiguity and D1–D7) separately and absolute perceptual bias[42] (ie, the absolute difference between the probability of both perceptual states and chance level) in ambiguous blocks. Within participants, we averaged all dependent variables across runs.

We performed group-level statistics using mixed ANOVA (within-subject factor: levels of disambiguating stimulus evidence D1–D7; between-subject factor: diagnostic group). Given heteroscedasticity between groups for congruent perceptual states (Levene test: $P = .043$), we used a linear mixed-effects (nlme R-package) model. The diagnostic group and disambiguating stimulus evidence defined fixed effects. Individual participants defined random effects. Weights were adjusted to account for unequal variance between groups.

We further fitted a set of functions [linear: $y = a + b * x$; exponential: $y = c * exp(g * x)$; sigmoid: $y = 0.5 + (0.5 - l)/(1 + exp(-(x - m)/n)]$ to the proportion of congruent perceptual states across conditions D1–D7. After identifying the exponential fit by means of the highest adjusted R^2, we compared individual growth rates as surrogates for the sensitivity to sensory evidence between groups. Because the number of free parameters (ie, complexity) in these models was fixed, the measure of accuracy can be treated as model evidence (ie, we performed a simple form of model comparison). Due to non-normality (Kolmogorov-Smirnov test: $P < .0001$), we used bootstrapping (R-dabestr[43]) to estimate confidence intervals (CI) for between-group differences in growth rates (see Supplementary Materials 1 for analyses of the linear fit) and perceptual bias.

In Supplementary Materials 2, we provide post hoc simulation analyses to illustrate the relation of our psychophysical approach to the predictive coding model of bistable perception.[29]

Stereo Disparity. We determined stereo-disparity thresholds by computing the average of presented stereo disparity at trials following the third reversal of each run and averaged across runs. Due to non-normality (Kolmogorov-Smirnov test: $P < .0001$), we probed a potential between-group difference by bootstrapping CIs.

Correlative Analyses. Finally, we asked whether individual questionnaire scores (PDI and CAPS; Bonferroni-corrected) correlated with the sensitivity to sensory evidence and average phase duration. In addition, we tested correlations with the PANSS subitems P1 (delusions) and P3 (hallucinations). Control analyses probed potential correlations to perceptual bias, unclear perceptual states, stereoacuity, as well as negative and general PANSS subscales (see Supplementary Materials 1 for median split analyses of CAPS/P3 and complete correlograms). Due to non-normality (Kolmogorov-Smirnov tests $P < .0001$ for all variables), we computed standard Spearman correlations. To correct for potential confounds that may influence performance in the Lissajous task and/or the severity of psychotic experiences, we assessed partial correlation coefficients. Such factors comprised stereoacuity (due to its potential influence on graded ambiguity, see above), the participants' age (due to its impact on bistable perception[44]), as well as the duration of illness and chlorpromazine equivalents as measures of disease severity. To ascertain specificity for the dimensions of psychotic experience, we also included scores on the alternative questionnaire (for correlations with PDI/CAPS), the respective alternative PANSS subitems (for correlations with P1/P3) and PANSS subscales (general and negative).

Results

Main Experiment

The nlme R-package model indicated a main effect of disambiguating stimulus evidence on the fraction of congruent perceptual states [$F(6) = 15.16$, $P = 6.44 \times 10^{-15}$], but no main effect of group [$F(1) = 0.02$, $P = .88$]. Importantly, we observed a significant interaction between diagnostic group and disambiguating stimulus evidence [$F(6) = 2.52$, $P = .02$, see figure 3A]. Mixed ANOVA yielded qualitatively identical results.

The change in the fraction of congruent perceptual states across D1–D7 was best fit by an exponential function (adjusted $R^2 = 0.39 \pm 0.10$, best fit in 70% of Scz patients and 65% of controls) as compared with linear (adjusted $R^2 = 0.38 \pm 0.10$) and sigmoid (adjusted $R^2 = 0.10 \pm 0.10$) functions. Sensitivity to additional sensory evidence as expressed by the growth rate of the exponential function was equal to 0.06 ± 0.01 in patients and 0.02 ± 0.02 in controls. Bootstrapping revealed a borderline significant difference between patients and controls (95% CI = 0.004 to −0.08, see figure 3B). Analysis of the linear fit yielded qualitatively identical results (see Supplementary Materials 1).

Mixed ANOVA did not yield a main effect of group or disambiguating stimulus evidence nor a between-factor interaction for the proportion of unclear perceptual states (patients: 0.01 ± 0.001; controls: 0.004 ± 0.001) or phase duration (patients: 21.25 ± 0.35 s; controls: 21.56 ± 0.36 s; see Supplementary Materials 1). Furthermore, we did not observe a significant between-group difference with regard to perceptual biases in ambiguity (patients: 0.09 ± 0.02, controls: 0.10 ± 0.02, 95% CI = −0.06 to 0.04).

Stereoacuity

Stereo-disparity thresholds amounted to $0.003 \pm 0.001^{\circ}$ in patients and $0.003 \pm 0.001^{\circ}$ in controls with no significant between-group difference (95% CI = –0.002 to 0.001).

Correlative Analyses

Within patients, sensitivity to disambiguating stimulus evidence correlated positively with the CAPS ($R = 0.51$, $P = .02$; figure 4). This was corroborated by the respective partial correlation ($R = 0.55$, $P = .03$, see above). Similarly, there was a significant correlation of the sensitivity parameter to PANSS subitem P3 (standard correlation: $R = 0.52$, $P = .01$; partial correlation: $R = 0.52$, $P = .04$). We did not observe a significant association between sensitivity to disambiguating stimulus evidence and PDI (standard correlation: $R = 0.36$, $P = .19$; partial correlation: $R = -0.35$, $P = .19$) or P1 (standard correlation: $R = 0.35$, $P = .11$; partial correlation: $R = 0.07$, $P = .78$). Analyses of the linear fit yielded qualitatively identical results.

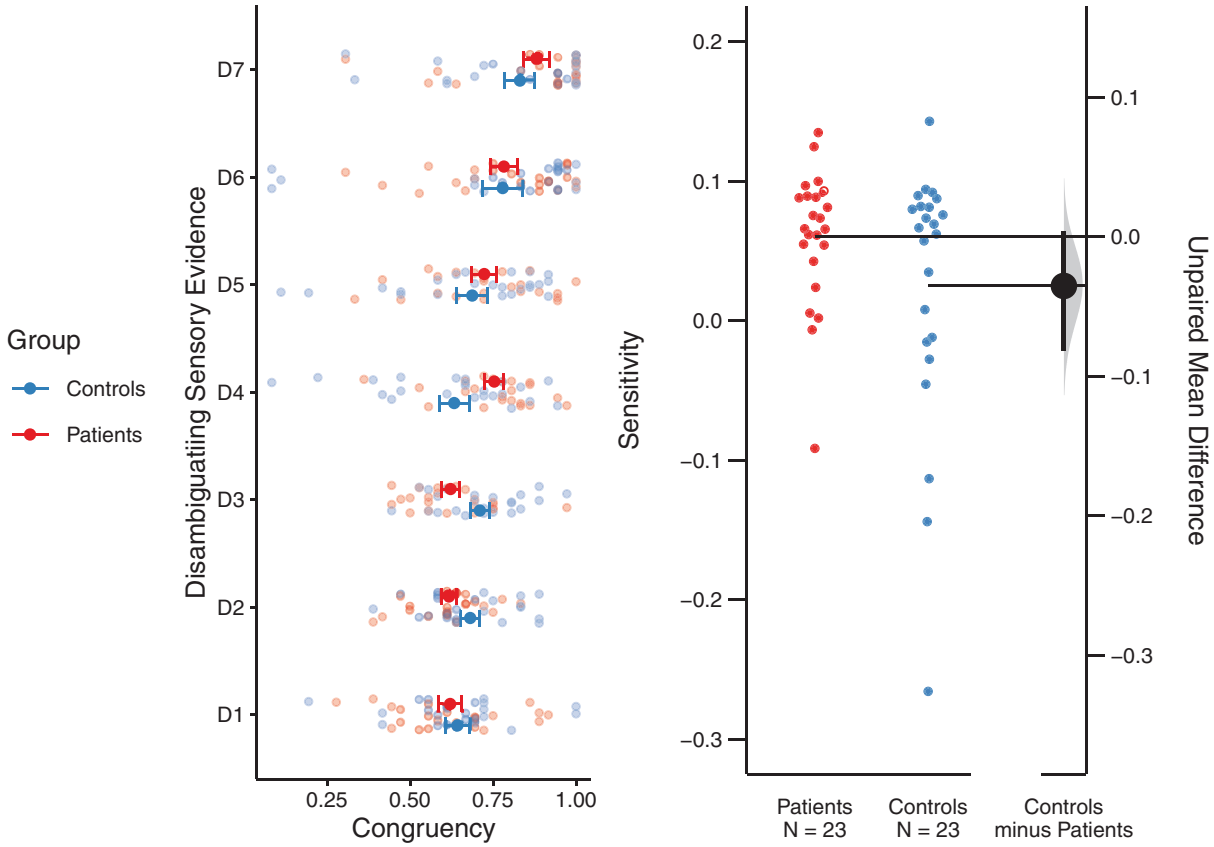


Fig. 3. Sensitivity to disambiguating stimulus evidence. We depict the fraction of congruency between perceptual states and sensory evidence across the levels of disambiguating stimulus evidence (D1–D7, left panel). Error bars represent the respective standard error of the mean. The nlme model yielded a main effect of disambiguating stimulus evidence [$F(6) = 15.16$, $P = 6.44 \times 10^{-15}$], and a significant interaction between the diagnostic group and the disambiguating stimulus evidence [$F(6) = 2.52$, $P = .02$]. The left panel shows the implicit interaction between levels of disambiguating stimulus evidence and diagnostic group: At low levels of disambiguation (D1–D3), controls exhibit a marginally higher proportion of congruent perceptual states. This is reversed for higher levels of disambiguating stimulus evidence (D4–D7), where patients show a greater proportion of congruency. We used the growth rate of individual exponential fits to the fraction of congruent perceptual states to express the individual sensitivities to disambiguating stimulus evidence during graded ambiguity (right panel; horizontal lines point to sample means; vertical line spans over the 95% CI). Bootstrapping revealed a borderline-significant between-group difference (estimated 95% CI = 0.004 to −0.08).

Furthermore, we observed a significant negative correlation of average perceptual phase duration with the CAPS (standard correlation: $R = -0.54$, $P = .01$; partial correlation: $R = -0.64$, $P = .01$) and a trendwise correlation to P3 (standard correlation: $R = -0.39$, $P = .07$; partial correlation: $R = -0.46$, $P = .07$). We did not find a significant association of phase duration to PDI or P1 in standard (PDI: $R = -0.21$, $P = .68$; P1: $R = -0.26$, $P = .23$) or partial correlations (PDI: $R = -0.35$, $P = .19$; P1: $R = -0.21$, $P = .44$).

Confirmatory analyses indicated a significant positive correlation of the sensitivity parameter to the positive and general PANSS subscale ("Positive": $R = 0.5$, $P = .02$; "General": $R = 0.52$, $P = .01$; "Negative": $R = 0.11$, $P = .61$). Interestingly, there were no significant correlations between sensory precision and negative symptoms or signs. CAPS and PDI were highly correlated in patients ($R = 0.76$, $P = 2.81 \times 10^{-5}$) and showed a trend for controls ($R = 0.35$, $P = .1$).

Neither of the 2 questionnaire scores (PDI/CAPS) and PANSS subitems (P1/P3) correlated with perceptual biases, fraction of unclear perceptual states, stereo-disparity thresholds, duration of illness, or chlorpromazine equivalents. Within controls, we did not find any significant correlation between questionnaire scores and the aforementioned variables (see Supplementary Materials 1 for additional correlation analyses and correlograms).

Discussion

In this study, we asked whether the experience of psychotic symptoms is associated with an increased impact of sensory evidence on perceptual inference relative to prior predictions (ie, a reduced prior-to-likelihood ratio at sensory processing levels).

Firstly, Scz patients showed an increased proportion of disambiguation-congruent perceptual states at high levels of stimulus information (D4–D7). At low levels (D1–D3),

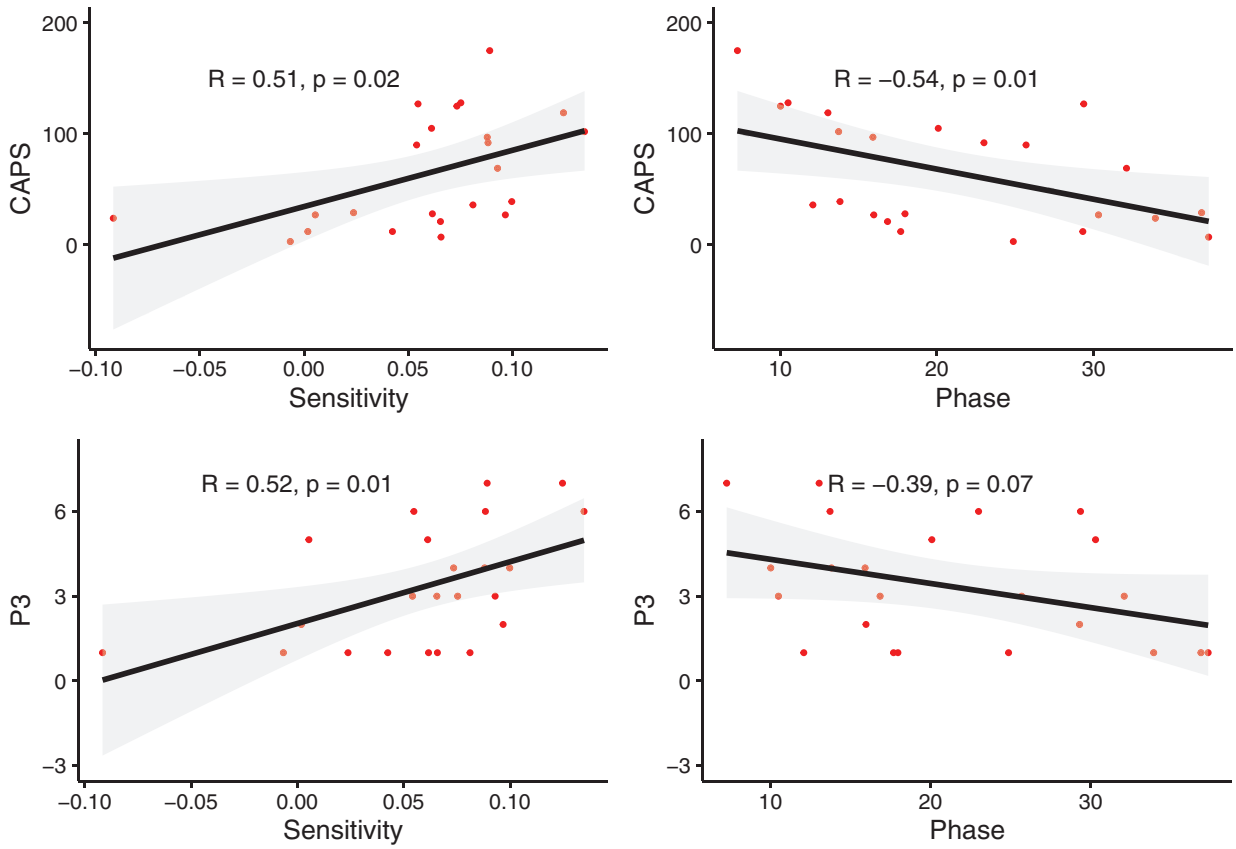


Fig. 4. Individual symptom severity. Here, we depict the individual patients' symptom severity with regard to perceptual anomalies (CAPS, top) and hallucination (P3, bottom) against the sensitivity to stimulus evidence (left) and phase duration (right) alongside regression lines (black) and 95% CI (light gray).

this proportion was similar between groups or even appeared to be reduced in patients (D3). This interaction thus speaks against a global increase in sensitivity to sensory evidence in Scz. Rather, it may suggest that patients show a greater benefit (or *gain*) at increasing levels of stimulus information. Indeed, due to this nonlinearity, these findings defy a simple explanation. Supplementary Materials 2 provides post hoc simulations of this interaction from a predictive coding model of bistable perception.[28,29]

Secondly, we found that the severity of perceptual anomalies and hallucinations correlated *positively* with the sensitivity to disambiguating stimulus evidence and *negatively* with average phase duration in Scz. Predictive coding models of bistable perception[28,29] relate enhanced sensory sensitivity to a shift of precision estimates *toward stimulus representations* (ie, the likelihood). In turn, such models assume that shorter phase durations signal a shift of precision estimates *away from implicit predictions* about perceptual stability (see[29] and Supplementary Materials 2). Through this lens, the two behavioral results, therefore, suggest that hallucinations are related to a decreased prior-to-likelihood ratio at sensory processing levels. At the same time, they contradict the hypothesis that a global shift toward prior precision (ie, an increased prior-to-likelihood ratio) underlies the experience of hallucinations.

These findings align with the "canonical" predictive coding account of Scz,[10] which assumes that psychotic symptoms arise due to a relative shift of inference away from priors and toward sensory evidence.[8] Along these lines, our results reverberate with the association of Scz to a reduced susceptibility to visual illusions,[16] impaired smooth pursuit,[45] and reduced sensory attenuation during force matching.[15,46] While our findings speak for a decrease as opposed to an increase in the prior-to-likelihood ratio, they cannot distinguish between a decrease in prior precision alone, an increase in likelihood precision alone or a combination of the two. Moreover, our results are compatible with alternative algorithms of dynamic belief updating such as circular inference[47,60] and alternative implementational frameworks of bistable perception such as mutual inhibition and adaption models.[48] In this context, differences in the excitation-inhibition balance[49] may lead to weaker inhibition between competing neuronal populations, which could explain why hallucinations correlated with individual characteristics of bistable perception.

Importantly, our results seem to contradict the association of hallucinations to overly precise priors.[19,21,22] However, this apparent discrepancy may be resolved by a differential modulation of the prior-to-likelihood ratio

across levels of the predictive coding hierarchy: Our paradigm targeted the interaction of prior and likelihood at sensory levels. A reduced prior-to-likelihood ratio may elicit the aberrant salience of sensory events.[24,25] This may drive higher levels into an overly strong weighting of priors and entail enhanced top-down influences on perception.[11] Finally, such a compensatory mechanism may trigger hallucinations,[21] thereby *explaining away*[5] aberrant salience at sensory levels.

Albeit strongly correlated with perceptual anomalies and hallucinations, our current findings did not reveal an association of delusional ideation to either sensitivity to sensory evidence or perceptual stability. This discrepancy to previous work[14] may result from differences between the experimental paradigms (Schmack et al.[14] stabilized perceptual states through intermittent presentation,[50] while we used a continuous stimulus). Speculatively, intermittent paradigms may boost perceptual priors and thus be more sensitive toward the relation of perceptual stability and delusions. In turn, manipulating sensory evidence through graded ambiguity may be more apt to detect associations to perceptual abnormalities. To resolve this discrepancy, future work should combine the novel paradigm of graded ambiguity with both intermittent presentation of bistable stimuli[13,14] and manipulations of higher-level beliefs.[33,51–53]

In contrast to our findings, previous research has revealed deficits in binocular depth perception in Scz.[54–57] Our stereoacuity assessment was analogous to the established *Random-Dot test*,[37,55] but estimated perceptual thresholds in a psychophysical staircase. This yielded values in the range commonly reported for stereoacuity.[55] In addition, our study did not show a global reduction in perceptual performance in Scz patients relative to controls. It thus seems less likely that low-level deficits (eg, reduced stereoacuity, contrast sensitivity,[55] or motion intergration[58]) can account for the current findings. Finally, perceptual biases (eg, when perceiving facial expressions[59]) are frequently reported in Scz. In the context of bistable perception, global differences in the probabilities of perceptual alternatives are a common phenomenon.[42] Importantly, this study did not reveal any significant effect of bias, which is thus unlikely to contribute to our results.

In sum, this study associates the experience of psychotic symptoms with an altered integration of prior beliefs and sensory evidence. Our results relate perceptual anomalies and hallucinations to a reduction of the prior-to-likelihood ratio in perception. This provides empirical evidence for the view that predictive processing deficits contribute to the emergence of psychotic symptoms and will enable novel approaches to the pathophysiological mechanisms of psychosis.

Supplementary Material

Supplementary data are available at *Schizophrenia Bulletin* online.

Acknowledgment

We thank Dr Rick Adams for a very valuable discussion on a previous version of the manuscript. The authors have declared that there are no conflicts of interest in relation to the subject of this study.

Funding

V.W. and H.S. were supported by the Berlin Institute of Health (BIH). A.H. and P.S. were supported by Deutsche Forschergemeinschaft (DFG; Funding Identification: HE 2597/19-1 and STE 1430/8-1). A.E. was supported by the Einstein Center for Neurosciences (ECN) Berlin.

References

1. Lee TS, Mumford D. Hierarchical Bayesian inference in the visual cortex. *J Opt Soc Am A Opt Image Sci Vis.* 2003;20(7):1434–1448.
2. Friston K. A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci.* 2005;360(1456):815–836.
3. Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci.* 2013;36(3):181–204.
4. Hohwy J. Attention and conscious perception in the hypothesis testing brain. *Front Psychol.* 2012;3:96.
5. Clark A. *Surfing Uncertainty: Prediction, Action and The Embodied Mind*. Oxford, UK: Oxford University Press; 2016:424.
6. Mathys C, Daunizeau J, Friston KJ, Stephan KE. A Bayesian foundation for individual learning under uncertainty. *Front Hum Neurosci.* 2011;5:39.
7. Fletcher PC, Frith CD. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat Rev Neurosci.* 2009;10(1):48–58.
8. Adams RA, Stephan KE, Brown HR, Frith CD, Friston KJ. The computational anatomy of psychosis. *Front Psychiatry.* 2013;4:47.
9. Stephan KE, Mathys C. Computational approaches to psychiatry. *Curr Opin Neurobiol.* 2014;25:85–92.
10. Sterzer P, Adams RA, Fletcher P, et al. The predictive coding account of psychosis. *Biol Psychiatry.* 2018;84(9):634–643.
11. Corlett PR, Horga G, Fletcher PC, Alderson-Day B, Schmack K, Powers AR 3rd. Hallucinations and Strong Priors. *Trends Cogn Sci.* 2019;23(2):114–127.
12. Jardri R, Hugdahl K, Hughes M, et al. Are hallucinations due to an imbalance between excitatory and inhibitory influences on the brain? *Schizophr Bull.* 2016;42(5):1124–1134.
13. Schmack K, Gòmez-Carrillo de Castro A, Rothkirch M, et al. Delusions and the role of beliefs in perceptual inference. *J Neurosci.* 2013;33(34):13701–13712.
14. Schmack K, Schnack A, Priller J, Sterzer P. Perceptual instability in schizophrenia: probing predictive coding accounts of delusions with ambiguous stimuli. *Schizophr Res Cogn.* 2015;2(2):72–77.
15. Teufel C, Kingdon A, Ingram JN, Wolpert DM, Fletcher PC. Deficits in sensory prediction are related to delusional ideation in healthy individuals. *Neuropsychologia.* 2010;48(14):4169–4172.
16. Notredame CE, Pins D, Deneve S, Jardri R. What visual illusions teach us about schizophrenia. *Front Integr Neurosci.* 2014;8:63.

17. Teufel C, Subramaniam N, Dobler V, et al. Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. *Proc Natl Acad Sci U S A.* 2015;112(43):13401–13406.
18. Davies DJ, Teufel C, Fletcher PC. Anomalous perceptions and beliefs are associated with shifts toward different types of prior knowledge in perceptual inference. *Schizophr Bull.* 2018;44(6):1245–1253.
19. Powers AR III, Kelley M, Corlett PR. Hallucinations as top-down effects on perception. *Biol Psychiatry Cogn Neurosci Neuroimaging.* 2016;1(5):393–400.
20. Alderson-Day B, Lima CF, Evans S, et al. Distinct processing of ambiguous speech in people with non-clinical auditory verbal hallucinations. *Brain.* 2017;140(9):2475–2489.
21. Powers AR, Mathys C, Corlett PR. Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. *Science.* 2017;357(6351):596–600.
22. Cassidy CM, Balsam PD, Weinstein JJ, et al. A perceptual inference mechanism for hallucinations linked to striatal dopamine. *Curr Biol.* 2018;28(4):503–514.e4.
23. Rao RP, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci.* 1999;2(1):79–87.
24. Heinz, A. Dopaminergic dysfunction in alcoholism and schizophrenia–psychopathological and behavioral correlates. *Eur Psychiatry.* 2002;17:9–16.
25. Kapur S. Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *Am J Psychiatry.* 2003;160(1):13–23.
26. Heinz, A. et al. Towards a unifying cognitive, neurophysiological, and computational neuroscience account of schizophrenia. *Schizophr Bull.* 2018;45(5):1092–1100.
27. Brascamp J, Sterzer P, Blake R, Knapen T. Multistable perception and the role of the frontoparietal cortex in perceptual inference. *Annu Rev Psychol.* 2018;69:77–103.
28. Hohwy J, Roepstorff A, Friston K. Predictive coding explains binocular rivalry: an epistemological review. *Cognition.* 2008;108(3):687–701.
29. Weilnhammer V, Stuke H, Hesselmann G, Sterzer P, Schmack K. A predictive coding account of bistable perception – a model-based fMRI study. *PLoS Comput Biol.* 2017;13(5):e1005536.
30. Peters ER, Joseph SA, Garety PA. Measurement of delusional ideation in the normal population: introducing the PDI (Peters et al. Delusions Inventory). *Schizophr Bull.* 1999;25(3):553–576.
31. Stuke H, Stuke H, Weilnhammer VA, Schmack K. Correction: psychotic experiences and overhasty inferences are related to maladaptive learning. *PLoS Comput Biol.* 2017;13(2):e1005393.
32. Preti A, Rocchi MB, Sisti D, et al. The psychometric discriminative properties of the Peters *et al*. Delusions Inventory: a receiver operating characteristic curve analysis. *Compr Psychiatry.* 2007;48(1):62–69.
33. Schmack K, Rothkirch M, Priller J, Sterzer P. Enhanced predictive signalling in schizophrenia. *Hum Brain Mapp.* 2017;38(4):1767–1779.
34. Bell V, Halligan PW, Ellis HD. The Cardiff Anomalous Perceptions Scale (CAPS): a new validated measure of anomalous perceptual experience. *Schizophr Bull.* 2006;32(2):366–377.
35. Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophr Bull.* 1987;13(2):261–276.
36. Brainard DH. The Psychophysics Toolbox. *Spat Vis.* 1997;10(4):433–436.
37. Tittes J, Baldwin AS, Hess RF, et al. Assessment of stereovision with digital testing in adults and children with normal and impaired binocularity. *Vision Res.* 2019;164:69–82.
38. Pastukhov A, Vonau V, Braun J. Believable change: bistable reversals are governed by physical plausibility. *J Vis.* 2012;12(1):17, 1–16.
39. Weilnhammer VA, Ludwig K, Hesselmann G, Sterzer P. Frontoparietal cortex mediates perceptual transitions in bistable perception. *J Neurosci.* 2013;33(40):16009–16015.
40. Weilnhammer VA, Ludwig K, Sterzer P, Hesselmann G. Revisiting the Lissajous figure as a tool to study bistable perception. *Vision Res.* 2014;98:107–112.
41. Weilnhammer VA, Sterzer P, Hesselmann G. Perceptual stability of the Lissajous figure is modulated by the speed of illusory rotation. *PLoS One.* 2016;11(8):e0160772.
42. Zhang X, Xu Q, Jiang Y, Wang Y. The interaction of perceptual biases in bistable perception. *Sci Rep.* 2017;7:42018.
43. Ho J, Tumkaya T, Aryal S, Choi H, Claridge-Chang A. Moving beyond *P* values: everyday data analysis with estimation plots. *Nat Methods.* 2019;16(7):565–566.
44. Díaz-Santos M, Mauro S, Cao B, Yazdanbakhsh A, Neargarder S, Cronin-Golomb A. Bistable perception in normal aging: perceptual reversibility and its relation to cognition. *Neuropsychol Dev Cognition B Aging Neuropsychol Cogn.* 2017;24;115–134.
45. Thakkar KN, Diwadkar VA, Rolfs M. Oculomotor prediction: a window into the psychotic mind. *Trends Cogn Sci.* 2017;21(5):344–356.
46. Shergill SS, Samson G, Bays PM, Frith CD, Wolpert DM. Evidence for sensory prediction deficits in schizophrenia. *Am J Psychiatry.* 2005;162(12):2384–2386.
47. Jardri R, Duverne S, Litvinova AS, Denève S. Experimental evidence for circular inference in schizophrenia. *Nat Commun.* 2017;8:14218.
48. Wilson HR. Minimal physiological conditions for binocular rivalry and rivalry memory. *Vision Res.* 2007;47(21):2741–2750.
49. Lisman J. Excitation, inhibition, local oscillations, or large-scale loops: what causes the symptoms of schizophrenia? *Curr Opin Neurobiol.* 2012;22(3):537–544.
50. Pearson J, Brascamp J. Sensory memory for ambiguous vision. *Trends Cogn Sci.* 2008;12(9):334–341.
51. Schmack K, Weilnhammer V, Heinzle J, Stephan KE, Sterzer P. Learning what to see in a changing world. *Front Hum Neurosci.* 2016;10:263.
52. Kornmeier J, Wörner R, Riedel A, Tebartz van Elst L. A different view on the Necker cube-differences in multistable perception dynamics between Asperger and non-Asperger observers. *PLoS One.* 2017;12(12):e0189197.
53. Weilnhammer VA, Stuke H, Sterzer P, Schmack K. The neural correlates of hierarchical predictions for perceptual decisions. *J Neurosci.* 2018;38(21):5008–5021.
54. Schechter I, Butler PD, Jalbrzikowski M, Pasternak R, Saperstein AM, Javitt DC. A new dimension of sensory dysfunction: stereopsis deficits in schizophrenia. *Biol Psychiatry.* 2006;60(11):1282–1284.
55. Kantrowitz JT, Butler PD, Schecter I, Silipo G, Javitt DC. Seeing the world dimly: the impact of early visual deficits on visual experience in schizophrenia. *Schizophr Bull.* 2009;35(6):1085–1094.

56. Hui L. Xia HS, Tang AS, et al. Stereopsis deficits in patients with schizophrenia in a Han Chinese population. *Sci. Rep*. 2017;7:45988.
57. Wang Z. Yu Z, Pan Z, et al. Impaired binocular depth perception in first-episode drug-naive patients with schizophrenia. *Front Psychol.* 2018;9:850.
58. Carter O, Bennett D, Nash T, et al. Sensory integration deficits support a dimensional view of psychosis and are not limited to schizophrenia. *Transl Psychiatry.* 2017;7(5):e1118.
59. Huang J, Chan RC, Gollan JK, et al. Perceptual bias of patients with schizophrenia in morphed facial expression. *Psychiatry Res.* 2011;185(1–2):60–65.
60. Leptourgos P, Notredame C-E, Eck M, Jardri R, Denève S. Circular inference in bistable perception. *bioRxiv.* 2019;521195.

2.4 An active role of inferior frontal cortex in conscious experience

Weilnhammer VA, Fritsch M, Chikermane M, Eckert AL, Kanthak K, Stuke H, Kaminski J, Sterzer P. Current Biology 31, R853-R856 (2021). DOI: https://doi.org/10.1016/j.cub.2021.04.043

The above publications highlight how bistable perception can be used to investigate the computational, behavioral and neural correlates of conscious experience in health and disease[43,50,73,80]. Specifically, they provide correlative evidence pointing to the IFC as a key region in perceptual inference[50,73,80] that may provide a target for non-invasive brain stimulation in patients suffering from psychotic symptoms such as hallucinations[43]. In this study, we used neuro-navigated theta-burst transcranial magnetic stimulation[82] (TMS) to probe whether virtual lesions in the IFC modulate conscious experiences during bistable perception. Using computational modeling and fMRI in two independent groups of healthy participants, we replicated our previous findings regarding the representation of prediction errors in IFC[80]. Crucially, TMS-induced virtual lesions in IFC reduced the frequency of transitions in conscious experience during bistable perception. This provides causal evidence for an active role of IFC in the resolution of sensory ambiguity and suggested the IFC as a potential target for the therapeutic modulation of hallucinatory experiences.

The following text corresponds to the abstract of the article[52]:

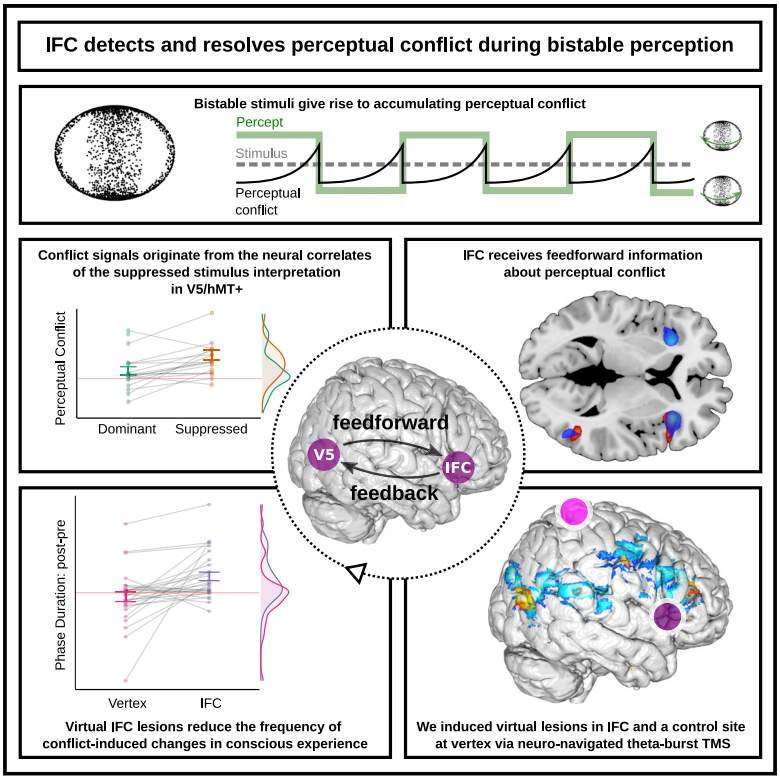
"In the search for the neural correlates of consciousness, it has remained controversial whether prefrontal cortex determines what is consciously experienced or, alternatively, serves only complementary functions, such as introspection or action. Here, we provide converging evidence from computational modeling and two functional magnetic resonance imaging experiments that indicated a key role of inferior frontal cortex in detecting perceptual conflicts caused by ambiguous sensory information. Crucially, the detection of perceptual conflicts by prefrontal cortex turned out to be critical in the process of transforming ambiguous sensory information into unambiguous conscious experiences: in a third experiment, disruption of neural activity in inferior frontal cortex through transcranial magnetic stimulation slowed down the updating of conscious experience that occurs in response to perceptual conflicts. These findings show that inferior frontal cortex actively contributes to the resolution of perceptual ambiguities. Prefrontal cortex is thus causally involved in determining the contents of conscious experience."

Article

Current Biology

An active role of inferior frontal cortex in conscious experience

Graphical abstract



Highlights

- V5/hMT+ detects conflicting sensory information during bistable perception
- Signals of perceptual conflicts are fed forward to inferior frontal cortex (IFC)
- Feedback from IFC to V5/hMT+ resolves perceptual conflicts
- IFC regulates how conflicting sensory information enters into conscious experience

Authors

Veith Weilnhammer, Merve Fritsch, Meera Chikermane, ..., Heiner Stuke, Jakob Kaminski, Philipp Sterzer

Correspondence

veith-andreas.weilnhammer@charite.de

In brief

Weilnhammer et al. use computational modeling, functional magnetic resonance imaging, and transcranial magnetic stimulation to show that inferior frontal cortex detects and resolves perceptual conflicts during bistable perception.

Weilnhammer et al., 2021, Current Biology *31*, 2868–2880
July 12, 2021
https://doi.org/10.1016/j.cub.2021.04.043

CellPress

CellPress

Current Biology

Article

An active role of inferior frontal cortex in conscious experience

Veith Weilnhammer,[1,2,6,7,8,*] Merve Fritsch,[1,6] Meera Chikermane,[1] Anna-Lena Eckert,[1,3,5] Katharina Kanthak,[1] Heiner Stuke,[1,2] Jakob Kaminski,[1,2] and Philipp Sterzer[1,2,3,4,5]

[1]Department of Psychiatry, Charité-Universitätsmedizin Berlin, 10117 Berlin, Germany
[2]Berlin Institute of Health, Charité-Universitätsmedizin Berlin and Max Delbrück Center, 10178 Berlin, Germany
[3]Bernstein Center for Computational Neuroscience, Charité-Universitätsmedizin Berlin, 10117 Berlin, Germany
[4]Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, 10099 Berlin, Germany
[5]Einstein Center for Neurosciences Berlin, Charité-Universitätsmedizin Berlin, 10117 Berlin, Germany
[6]These authors contributed equally
[7]Twitter: @weilnhammer
[8]Lead contact
*Correspondence: veith-andreas.weilnhammer@charite.de
https://doi.org/10.1016/j.cub.2021.04.043

SUMMARY

In the search for the neural correlates of consciousness, it has remained controversial whether prefrontal cortex determines what is consciously experienced or, alternatively, serves only complementary functions, such as introspection or action. Here, we provide converging evidence from computational modeling and two functional magnetic resonance imaging experiments that indicated a key role of inferior frontal cortex in detecting perceptual conflicts caused by ambiguous sensory information. Crucially, the detection of perceptual conflicts by prefrontal cortex turned out to be critical in the process of transforming ambiguous sensory information into unambiguous conscious experiences: in a third experiment, disruption of neural activity in inferior frontal cortex through transcranial magnetic stimulation slowed down the updating of conscious experience that occurs in response to perceptual conflicts. These findings show that inferior frontal cortex actively contributes to the resolution of perceptual ambiguities. Prefrontal cortex is thus causally involved in determining the contents of conscious experience.

INTRODUCTION

The neural basis of conscious experience is one of today's greatest mysteries.[1] Its unraveling will have important implications for how we approach patients who remain unresponsive after brain damage or who suffer from hallucinatory distortions of perception.[2] Likewise, such progress may expand our ability to detect the presence of conscious experience in organic and artificial life beyond the human mind.[3] Solutions to these challenges will require identifying not only the neural correlates of consciousness[4,5] but also the computational function of specific brain regions for conscious experience.[6]

Bistable perception has been a key experimental approach in the search for a neuro-computational understanding of consciousness for more than two decades.[7] In this phenomenon, stimuli that are compatible with two interpretations give rise to perceptual conflict.[8] Faced with this conflict, observers perceive periodic changes in conscious experience, oscillating between two mutually exclusive perceptual states.[9] Thereby, bistable perception provides a unique window into a fundamental functional aspect of consciousness: the transformation of ambiguous sensory information into unambiguous conscious experience.[10,11]

Functional neuroimaging studies in humans have identified the right inferior frontal cortex (IFC) (a subregion of prefrontal cortex; Figure 1A) as a key region in bistable perception.[7] When compared to stimulus-driven changes in perception, spontaneous perceptual changes during bistability were consistently associated with increased activity in IFC,[7] suggesting that prefrontal cortex actively contributes to conscious experience.[9,11–13] Along this line of thought, IFC may resolve perceptual conflict by triggering perceptual changes through feedback signaling to sensory areas (Figure 1A).[9,11]

However, this *feedback* account has been questioned by work that related IFC to cognitive phenomena that occur in the wake of conscious experience, such as the processing of perceptual uncertainty,[14] motor behavior,[15] or, more broadly, the engagement of executive functions in response to changes in perception.[16,17] Perceptual conflict may thus rather be resolved within visual cortex and elicit activity in IFC through a *feedforward* mechanism. Accordingly, IFC activity may not reflect the cause but the consequence of changes in conscious experience.

To settle the ongoing controversy between the feedforward and feedback account of bistable perception will be a critical step in elucidating the computational role of prefrontal cortex for

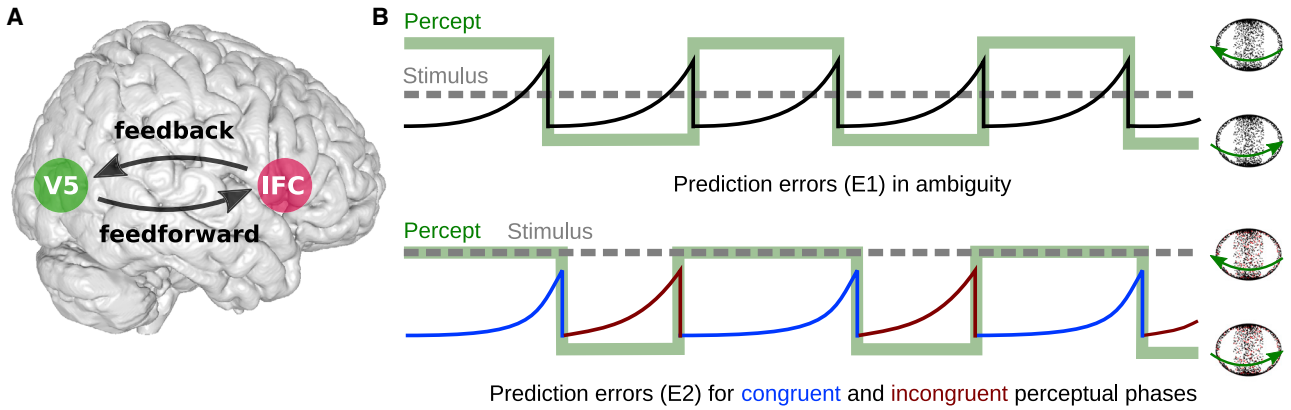


Figure 1. Concept

(A) The role of IFC (inferior frontal cortex; schematic overlay in pink) in conscious experience is controversial: according to one view, feedback from IFC may modulate perceptual processing in visual cortex (motion-sensitive visual cortex V5/hMT+; highlighted in green). This may reflect an active contribution of prefrontal brain activity to conscious experience. The opposing view links neural activity in IFC to the subjective uncertainty, report, or reportability of perceptual events. This suggests that conscious experience may be realized within visual cortex and may drive activity in IFC by means of feedforward processing.

(B) Here, we depict conscious experience and associated changes in accumulating perceptual conflict for bistable perception induced by a random dot structure-from-motion stimulus (RDK). Perceived direction of rotation (green line) alternates between left- and rightward motion of the front surface (icons on the right). In the absence of disambiguating sensory evidence (upper panel; gray dotted line), prediction errors (black solid line) accumulate while perception remains constant, until the perceptual conflict is reduced by a change in conscious experience. When faced with additional stimulus information (lower panel), conscious experience fluctuates between congruent or incongruent perceptual states. If an observer adopts a percept that is congruent with the disambiguating stimulus information, prediction errors are reduced (blue line). Accordingly, conflict-driven changes in conscious experience toward the alternative stimulus interpretation become less likely (vice versa for incongruent perceptual states; red line).

See also Figure S2 and Video S1.

consciousness. In this work, we conjectured that these seemingly contradictory views may be reconciled within an explanatory framework that incorporates both feedforward and feedback processing. To this end, we drew on a closely related line of research into the role of parietal cortex in bistable perception[18–22] that supports the idea that spontaneous changes in conscious experience may be best explained by hierarchical models of perceptual inference.[10,23] Specifically, results from these studies suggest that subregions within intraparietal sulcus may have complementary roles in perceptual inference, with an anterior subregion providing perceptual hypotheses via feedback to sensory areas and a posterior subregion signaling conflicts between the current hypothesis and the available sensory data in a feedforward manner.[7,21]

Here, we reasoned that the apparent discrepancy between feedforward and feedback accounts of prefrontal cortex function in bistable perception may be resolved along similar lines. First, we hypothesized that IFC may detect perceptual conflicts that arise between the contents of conscious experience and the available sensory data through a feedforward mechanism. To test this hypothesis, we used functional magnetic resonance imaging (fMRI) in conjunction with computational modeling in a Bayesian framework. Second, we reasoned that the detection of perceptual conflict by IFC may in turn trigger changes in conscious experience via feedback signaling to sensory areas. This latter hypothesis was tested by disrupting IFC activity through neuronavigated transcranial magnetic stimulation (TMS).

RESULTS

In a series of three experiments (E1–E3; STAR Methods; Figure S1), human observers reported changes in their perception of a rotating sphere (rightward, leftward, or unclear direction of rotation). In this structure-from-motion stimulus, random dots distributed on two intersecting rings induce illusory 3D motion (Video S1). Due to the perceptual conflict inherent in the ambiguous visual input, participants perceived spontaneous changes between left- and rightward rotation that occurred every 25.08 ± 2.57 s (phase duration, i.e., the average time spent between two consecutive changes in conscious experience; Figures S2A and S2B).

With this type of stimulus, unclear perceptual states[14] are rare (0.6% ± 0.18%). Moreover, changes in perceived direction of rotation occur only when fore- and background of the illusory sphere overlap (Figures S3A and S3B).[24] These perceptual features ensured the temporal precision of our approach and allowed us to compute response times (average response time [*RT*] = 0.81 ± 0.05 s) as a control variable for processes associated with behavioral reports.[15,16]

IFC detects accumulating perceptual conflict

Bayesian perceptual inference[25] provides a plausible computational explanation for the effects of conflicting stimulus information on perception. In this framework, conscious experience is understood as an iterative process, constantly generating and testing hypotheses about the most likely cause of sensory data.[26] In bistable perception, the two alternating stimulus interpretations reflect mutually exclusive hypotheses that are both compatible with but cannot fully account for the ambiguous sensory data. This results in perceptual conflict.[7,8,11]

As a quantitative representation of such conflict, the residual evidence for the alternative to the currently dominant stimulus interpretation can be conceived of as a perceptual prediction

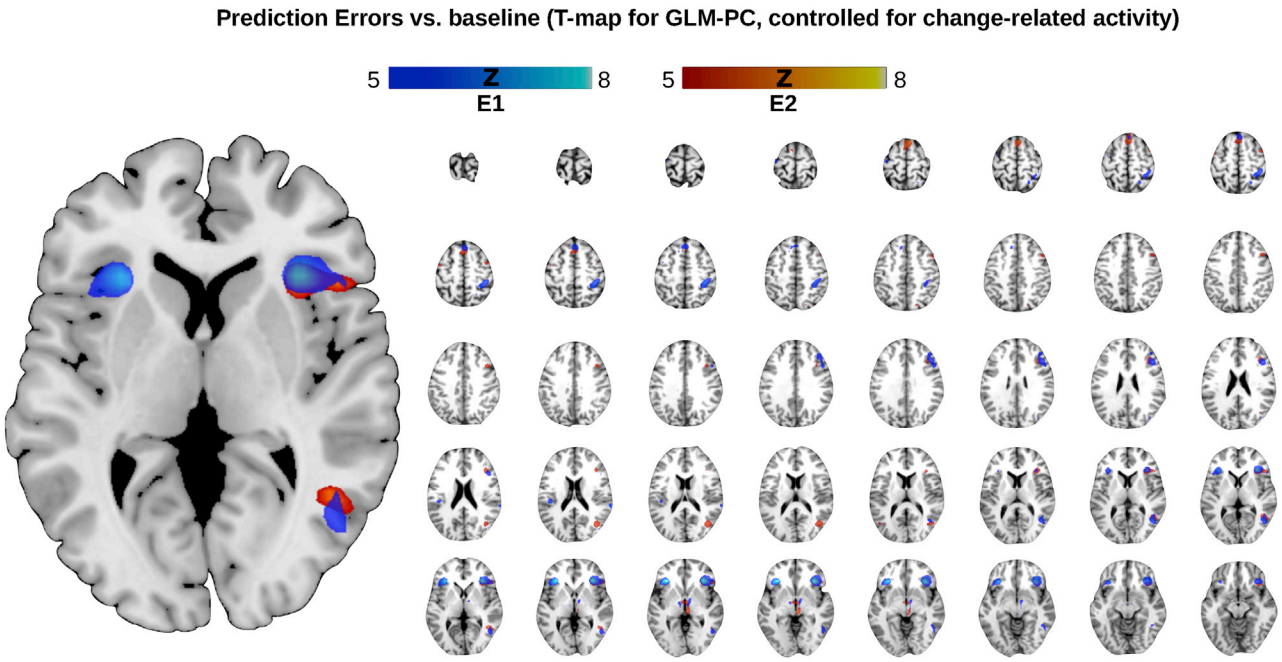


Figure 2. The neural correlates of accumulating perceptual conflict

Converging evidence from two fMRI experiments (E1: blue heatmap; E2: red heatmap; both displayed for T > 5) indicated that perceptual prediction errors correlate with neural activity in right IFC (anterior insula and inferior frontal gyrus) and V5/hMT+. Additional activations were located in left insula, right posterior-medial frontal gyrus, and right inferior parietal lobulus (all $p_{FWE} < 0.05$; see corresponding Table S1). Please note that these analyses controlled for change-related BOLD responses. See also Figure S4 and Table S1.

error.[10] This unexplained error induces a progressive shift in the balance between the two perceptual hypotheses.[13] Over time, prediction errors therefore accumulate until the increasing perceptual conflict is reduced by a change in conscious experience from the dominant to the alternative stimulus interpretation (Figure 1B). A recent proof-of-concept study has linked this process to neural activity in prefrontal cortex:[13] during ambiguous visual stimulation, blood-oxygen-level-dependent (BOLD) signals in IFC gradually increased while perception remained constant, peaking at the time of a conflict-induced change in conscious experience.

In experiment E1, we sought to (1) confirm the previously suggested role of IFC in detecting perceptual conflict and to (2) test the hypothesis that such perceptual conflict originates from visual cortex.

To identify the neural representation of perceptual conflict, we acquired fMRI data during bistable perception and searched for correlations of BOLD activity with the dynamics of perceptual prediction errors. To this end, we inverted an established computational model of bistable perception (STAR Methods)[10,13] based on the individual participants' behavioral reports on perceptual changes. This model estimates dynamic perceptual prediction errors to explain the sequence of conscious experiences during bistable perception. In model-based fMRI, we searched for the neural correlates of these prediction errors while controlling for BOLD activity associated with the timing and report of perceptual changes. In line with previous results,[13] we found that perceptual prediction errors correlated with BOLD activity in right-hemispheric IFC (anterior insula, inferior frontal gyrus;[7] Bonferroni corrected for family-wise error $p_{FWE} < 0.05$; Figure 2; Table S1).

Of note, previous studies have predominantly linked IFC to event-related processes associated specifically with perceptual changes.[7] We reasoned that this often-reported finding of change-related activity in IFC may actually correspond to the peak of accumulating prediction errors (Figures 1B and S4). In our data, such change-related IFC activity was only detectable if the analysis did not control for prediction errors (Figure 3A). Indeed, a direct comparison based on posterior probability maps[27] confirmed that BOLD activity in right-hemispheric IFC was better explained by prediction errors that gradually accumulated in each perceptual phase than by perceptual change events (Figure 3B). This suggests that, during bistable perception, IFC activity reflects the gradual accumulation of perceptual conflict[13] until it is temporarily reduced by a conflict-driven change in conscious experience (Figure 1B; see below for a replication of this finding in experiment E2).

Yet as a supra-modal brain region, IFC is unlikely to represent visual information independently of sensory brain regions. We therefore hypothesized that information about perceptual conflict may be fed forward to IFC from the representation of perceptual content in visual cortex.[28] Indeed, perceptual prediction errors also correlated with BOLD activity in the motion-sensitive extrastriate visual area V5/hMT+ (Figure 2).[29] Dynamic causal modeling[30] confirmed that these signals of accumulating perceptual conflict were most likely to originate from V5/hMT+, reaching IFC via feedforward effective connectivity (Figure S5).

Moreover, neural activity in V5/hMT+ also reflected the content of conscious experience, that is, whether participants experienced leftward or rightward rotation during bistable perception (Figures S6A and S6B): based on multi-voxel pattern analysis[31]

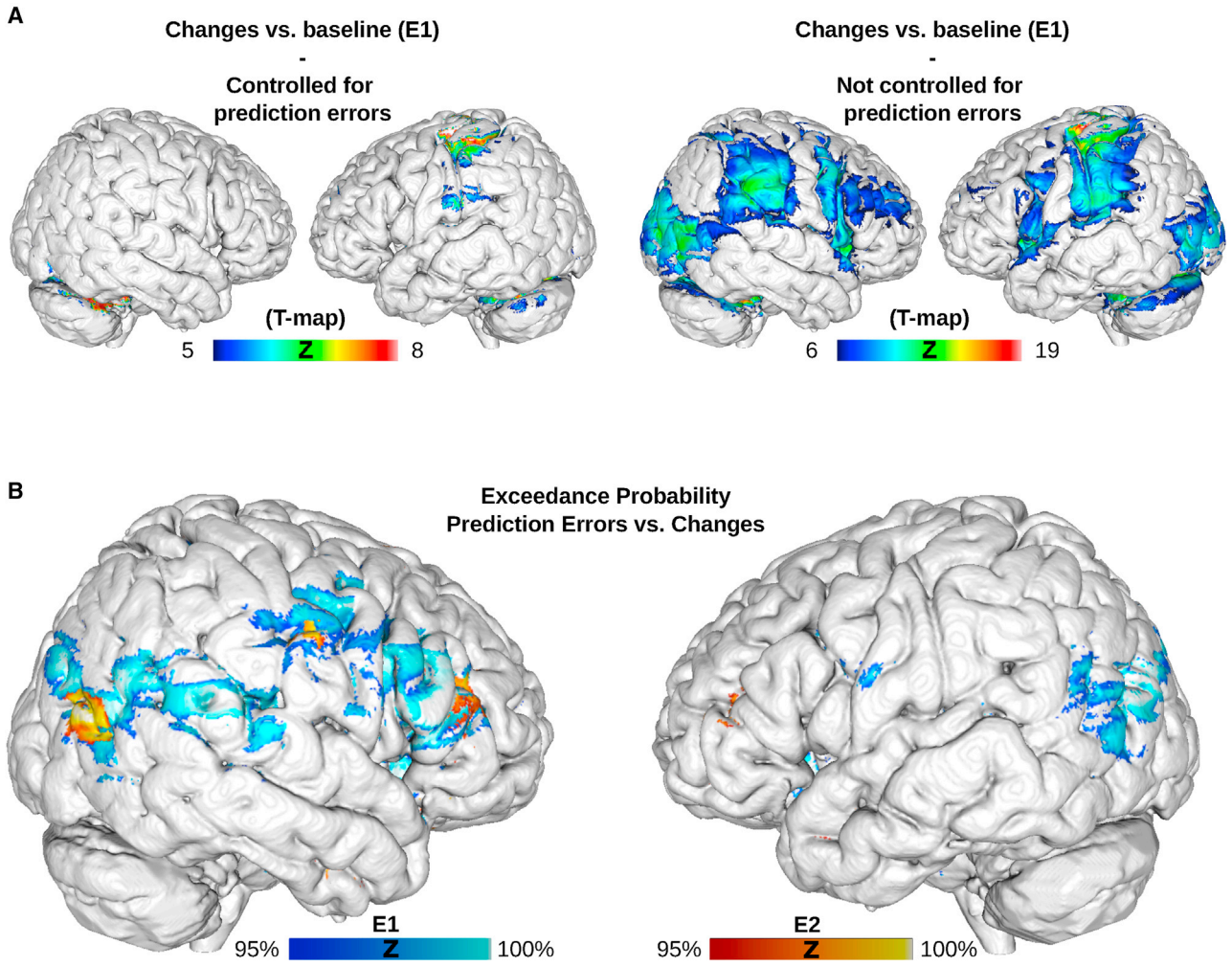


Figure 3. Conflict- versus change-related BOLD activity

(A) When analyzing the neural correlates of perceptual events while controlling for BOLD activity related to gradually accumulating perceptual prediction errors (generalized linear model [GLM]-PC, left panel), we found activations in bilateral cerebellum, left pre- and postcentral gyrus, bilateral midcingulate cortex and putamen, left insula, left IPL, as well as left medial frontal gyrus (p_{FWE} = 0.05). No significant clusters were observed in right-hemispheric IFC or V5/hMT+. Yet when assessing the neural correlates of perceptual events without controlling for perceptual prediction errors (i.e., by deleting the prediction-error regressor from GLM-PC, right panel), we observed highly significant change-related activity in bilateral insula, right inferior frontal gyrus, bilateral V5/hMT+, bilateral cerebellum, left pre- and postcentral gyrus, bilateral midcingulate cortex, bilateral inferior parietal lobulus, and left middle frontal gyrus ($p_{FWE} = 10^{-6}$). Hence, when studied in isolation of prediction errors, perceptual events did activate regions in right IFC.

(B) We applied a Bayesian posterior probability map approach to compare the explanatory power of gradually accumulating perceptual prediction errors against the explanatory power of event-related regressors that represent perceptual changes. Here, we display voxels where BOLD activity was better explained by gradually accumulating prediction errors at an exceedance probability above 95% (E1: blue heatmap; E2: red heatmap). Across both experiments, the posterior probability maps yielded converging evidence that neural signals from IFC and V5/hMT+ (as well as from additional parietal brain regions) were better explained by prediction-error-related activity as compared to change-related activity.

See also Figure S5.

of BOLD signals in V5/hMT+, we were able to decode which of the two stimulus interpretations was, at a given point in time, *dominant* or *suppressed* (see Figure S6B for region of interest [ROI]-based decoding from IFC, where we did not find conclusive evidence for or against the decodability of perceptual content).

We therefore asked whether the neural correlates of perceptual conflict originated from the voxels that represented the currently suppressed stimulus interpretation in visual cortex. As predicted, the BOLD signal in V5/hMT+ voxels that showed enhanced activity during perception of *leftward* rotation correlated more strongly with perceptual prediction errors when participants experienced *rightward* rotation ($BF_{10} = 2.24 \times 10^3$; Figure 4A) and vice versa ($BF_{10} = 2.57 \times 10^3$; see below for a replication of this finding in E2). This intriguing dissociation between the representation of perceptual content and perceptual conflict occurred only in voxels with strong biases toward one of the two stimulus interpretations (Figure S6C).

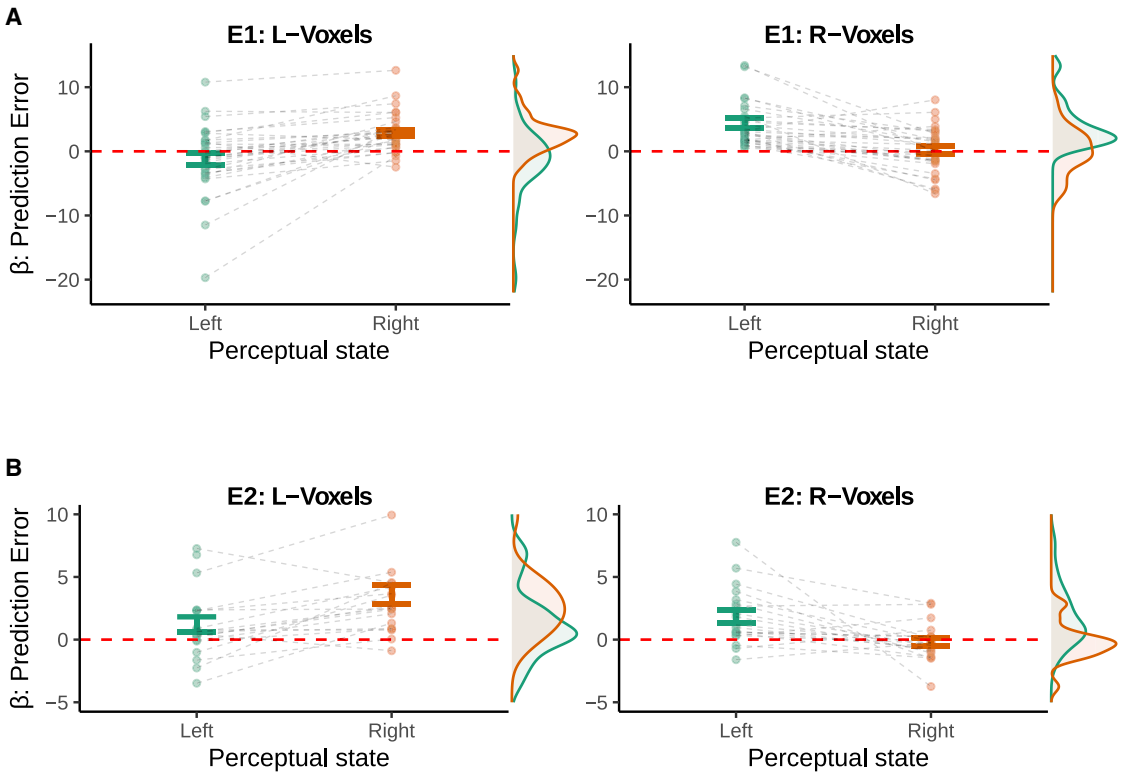


Figure 4. Neural correlates of perceptual conflict in V5/hMT+

(A) In experiment E1, we delineated V5/hMT+ based on the effects of visual stimulation (i.e., independently of our computational model of bistable perception) and identified *biased* voxel populations that showed elevated neural activity during either leftward (L) or rightward (R) illusory rotation (T value > 1; average number of voxels per population N_{pop} = 33.97 ± 1.78). While controlling for effects related to perceptual changes, we found that BOLD activity in L-voxels (left panel) correlated more strongly with perceptual prediction errors when participants consciously perceived rightward rotation (paired t test: T(32) = −5.26; p = 9.22×10^{-6}; $BF_{10} = 2.24\times10^{3}$). Conversely, R-voxels (right panel) correlated more strongly with perceptual prediction errors when participants consciously perceived leftward rotation (T(32) = 5.32; p = 7.94×10^{-6}; $BF_{10} = 2.57\times10^{3}$).

(B) Experiment E2 (N_{pop} = 32.92 ± 3.12) replicated these results: L-voxels (left panel) correlated more strongly with perceptual prediction errors during illusory rotation toward the right (paired t test: T(19) = −4.07; p = 6.49×10^{-4}; BF_{10} = 53.36). Inversely, R-voxels (right panel) correlated more strongly with perceptual prediction errors when the participants consciously perceived leftward rotation (T(19) = 3.11; p = 5.71×10^{-3}; BF_{10} = 8.2).

Error bars represent the SEM. See also Figure S6.

IFC is sensitive to graded changes in perceptual conflict

The results of E1 indicate that IFC receives feedforward information about perceptual conflict, emanating from the representations of ambiguous stimuli in visual cortex. Yet in everyday perception, fully ambiguous stimuli like those giving rise to bistable perception are rare. Rather, additional (i.e., disambiguating) stimulus information is usually available, albeit often incomplete.[32] In experiment E2, we sought to confirm the role of IFC in the signaling of perceptual conflict by measuring its responses to such disambiguating stimulus information.

To this end, we conducted an independent fMRI experiment based on the novel paradigm of graded ambiguity.[29,33] As in E1, participants reported changes in the perceived direction of rotation of a structure-from-motion stimulus. In contrast to E1, we introduced random changes in a disambiguating 3D signal attached to a fraction of the stimulus dots. The amount of disambiguating information was varied parametrically across six levels of signal-to-ambiguity ratio. As a consequence, conscious experience fluctuated to varying degrees between perceptual states that were congruent or incongruent with the disambiguating stimulus information (Figure 1B, lower panel). We assumed that, depending on the signal-to-ambiguity ratio, perceptual conflict should be greater during incongruent perceptual states, thus increasing the likelihood of conflict-driven changes toward the alternative stimulus interpretation.

As expected,[33] congruent perceptual states were indeed more frequent for increasing signal-to-ambiguity ratios ($BF_{10} = 2.91\times10^{22}$; Figure 5A). Both model simulation (Figures S7A–S7C) and computational modeling of the participants' behavior (Figure 5B) confirmed that prediction errors were enhanced during incongruent as compared to congruent perceptual states (main effect of congruency), with stronger effects at higher signal-to-ambiguity ratios (interaction between congruency and signal to ambiguity).

Crucially, this pattern was reflected by neural activity in IFC and V5/hMT+: while controlling for variations in BOLD signals associated with reported changes in conscious experience, we observed enhanced BOLD signals during incongruent perceptual states in right-hemispheric IFC and V5/hMT+ (main effect of congruency, $p_{FWE} < 0.05$; Figure 5C; see Table S2 for

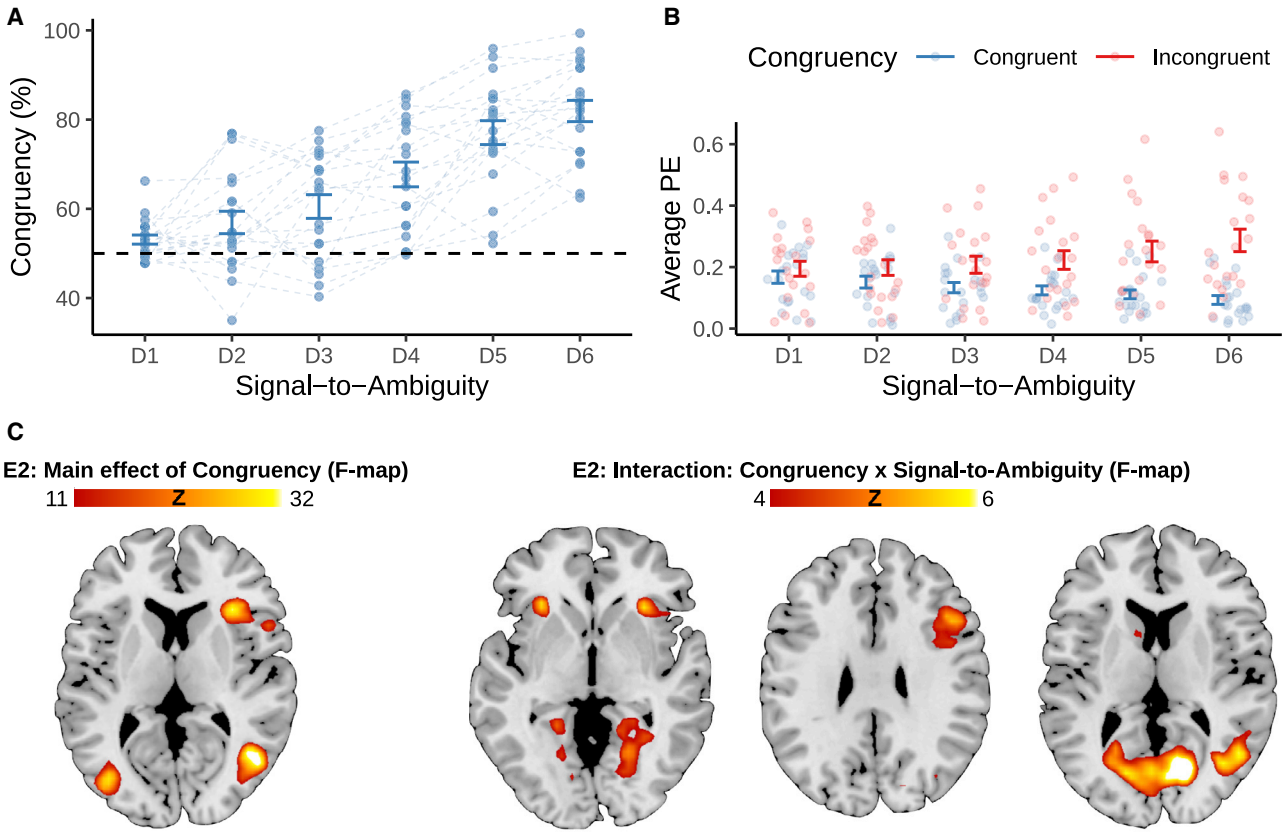


Figure 5. Perceptual conflict during graded ambiguity

(A) Conscious experience was biased toward perceptual states that were congruent with the disambiguating stimulus information ($T(19) = 8.45$; $p = 7.37 \times 10^{-8}$; $BF_{10} = 1.97 \times 10^{5}$). For increasing signal-to-ambiguity ratios (levels D1–D6), congruent perceptual states were more frequent ($F(95) = 51.14$; $p = 1.84 \times 10^{-25}$; $BF_{10} = 2.91 \times 10^{22}$).

(B) Computational modeling of behavior indicated that average prediction errors (PEs) were elevated during incongruent as compared to congruent perceptual states ($F(209) = 158.08$; $p = 2.29 \times 10^{-27}$; $BF_{10} = 3.09 \times 10^{20}$). The difference in PEs between congruent and incongruent perceptual states was enhanced for higher signal-to-ambiguity ratios ($F(209) = 10.41$; $p = 6.19 \times 10^{-9}$; $BF_{10} = 2.61 \times 10^{6}$). Overall, prediction errors did not vary across levels of signal to ambiguity ($F(209) = 0.54$; $p = 0.75$; $BF_{10} = 0.02$).

(C) We found enhanced BOLD responses during incongruent as opposed to congruent perceptual states in right-hemispherical IFC and V5/hMT+, alongside additional clusters in left precentral gyrus, right posterior-medial frontal gyrus (PMF), and right fusiform gyrus (left panel; $p_{FWE} < 0.05$; displayed for $F > 11$; see corresponding Table S2). Importantly, differences in BOLD activity between incongruent and congruent perceptual states were enhanced at higher levels of signal to ambiguity in right-hemispheric insula, inferior frontal gyrus, and V5/hMT+ (right panel; $p_{SCV} < 0.05$ within the main effect of congruency; displayed for $F > 4$).

Error bars represent the SEM. See also Figure S7 and Table S2.

additional activations). As predicted, both right-hemispheric IFC and V5/hMT+ showed larger differences between incongruent and congruent perceptual states at higher signal-to-ambiguity ratios (interaction between congruency and signal to ambiguity; small-volume correction at $p_{SVC} < 0.05$ within the main effect of congruency).

This factorial approach to the neural correlates of perceptual conflict was corroborated by model-based fMRI, which provided a complete replication of E1: while controlling for change-related activity, we found that accumulating perceptual prediction errors correlated with neural activity in right-hemispheric IFC and V5/hMT+ ($p_{FWE} < 0.05$; Figure 2; Table S1). In comparison to the analysis based on perceptual change events, the dynamic accumulation of perceptual conflict was better at explaining BOLD signals in right-hemispheric IFC (Figure 3B). Dynamic causal modeling indicated that signals of perceptual conflict were most likely to originate from V5/hMT+, reaching IFC via feedforward effective connectivity (Figure S5). Again, the BOLD signal in V5/hMT+ voxels that showed enhanced activity during perception of *leftward* rotation correlated more strongly with perceptual prediction errors when participants experienced *rightward* rotation ($BF_{10} = 53.36$) and vice versa ($BF_{10} = 8.2$; Figures 4B and S6D).

Together, E2 confirmed our hypothesis that IFC signals dynamic changes in perceptual conflict that are induced by disambiguating stimulus information. Additional control analyses (Figures S7D and S7E) ruled out variations in perceptual uncertainty and temporal imbalances between congruent and incongruent perceptual states as alternative explanations for our fMRI results.

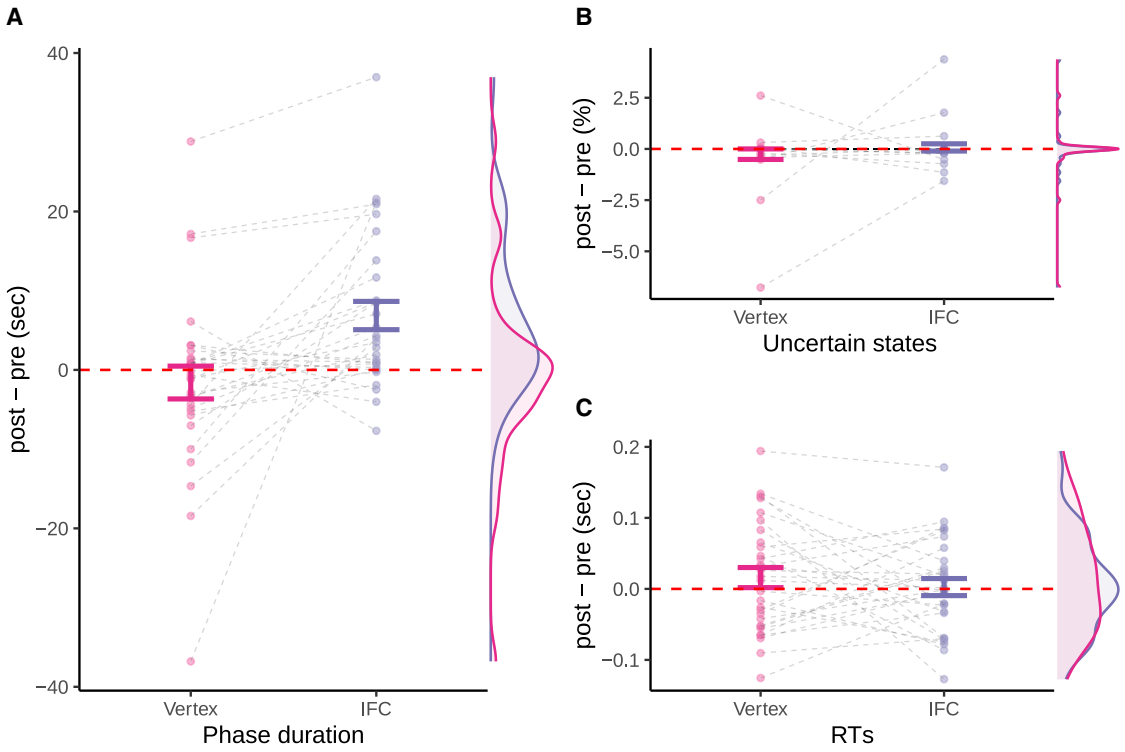


Figure 6. TMS effects on perception

(A) Virtual IFC lesions prolonged phase durations relative to the vertex condition (paired t test: $T(29) = -3.44$; $p = 1.77\times10^{-3}$; $BF_{10} = 20.05$) as well as against the baseline recorded prior to IFC stimulation (one-sample t test: $T(29) = 3.85$; $p = 6.08\times10^{-4}$; $BF_{10} = 51.47$).

(B) TMS to IFC did not alter the frequency of unclear perceptual states in comparison to the vertex condition ($T(29) = -1.04$; $p = 0.31$; $BF_{10} = 0.32$).

(C) Likewise, virtual IFC lesions did not affect RTs in comparison to control stimulation at vertex ($T(29) = 0.77$; $p = 0.45$; $BF_{10} = 0.26$).

Error bars represent the SEM. See also Figures S2 and S3.

Disruption of neural activity in IFC modulates the dynamics of conscious experience

The independent fMRI experiments E1 and E2 provide converging evidence that IFC detects the conflict inherent in sensory ambiguity. In a third experiment (E3), we asked whether this unconscious detection of perceptual conflict by IFC[34] is relevant for conscious experience. We reasoned that the signaling of perceptual conflict by IFC might facilitate changes in conscious experience during bistable perception. Consequently, disruption of IFC activity should reduce the frequency of such conflict-driven perceptual changes. To test this hypothesis, we used inhibitory TMS with a theta-burst stimulation protocol[35] to create virtual lesions in IFC.

In E3, we re-invited the participants from E1 for two TMS sessions scheduled on consecutive days. In each session, they first reported changes in conscious experience during two runs of ambiguous structure from motion. This was followed by 40 s of neuronavigated TMS to either IFC or a control location at the cranial vertex (see STAR Methods for details). Immediately afterward, participants again reported their perception during two runs of ambiguous structure from motion.

After neural activity in IFC was disrupted by TMS, changes in conscious experience occurred less frequently: for virtual lesions in IFC, we observed prolonged perceptual phase durations (post-pre: 6.86 ± 1.79 s) relative to the vertex condition (− 1.59 ± 2.07 s; paired t test: $BF_{10} = 20.05$; Figure 6A). This finding indicates that IFC not only detects gradually accumulating perceptual conflict but also has a causal role in triggering changes in conscious experience.

Two additional control analyses addressed alternative accounts of the observed TMS effect on perceptual phase durations. First, previous work has shown that activity in frontal brain regions is elevated at the time of unclear perceptual states during bistability.[14] Here, however, disruption of neural activity in IFC did not alter frequency of unclear perceptual states (post-pre: 0.07% ± 0.18%) in comparison to vertex stimulation (− 0.26% ± 0.26%; $BF_{10} = 0.32$; Figure 6B).

Second, when investigating frontal activity as a potential driver of changes in conscious experience during bistable perception, decision-related phenomena (such as task relevance) and output-related processes (such as motor preparation and button presses) represent potential confounds.[36] This issue has recently been addressed in "no-report" paradigms, which suggested that a subset of change-related activations in prefrontal cortex may represent the neural correlates of report rather than the mechanisms involved in conscious experience per se.[15,16] Here, we used RTs to ask whether inhibition of activity in IFC impaired the participants' ability to report on the contents of conscious experience. Changes in RTs did not differ between IFC (post-pre: 2.55×10^{-3} ± 0.01 s) and vertex stimulation (0.02 ± 0.01 s; $BF_{10} = 0.26$; Figure 6C).

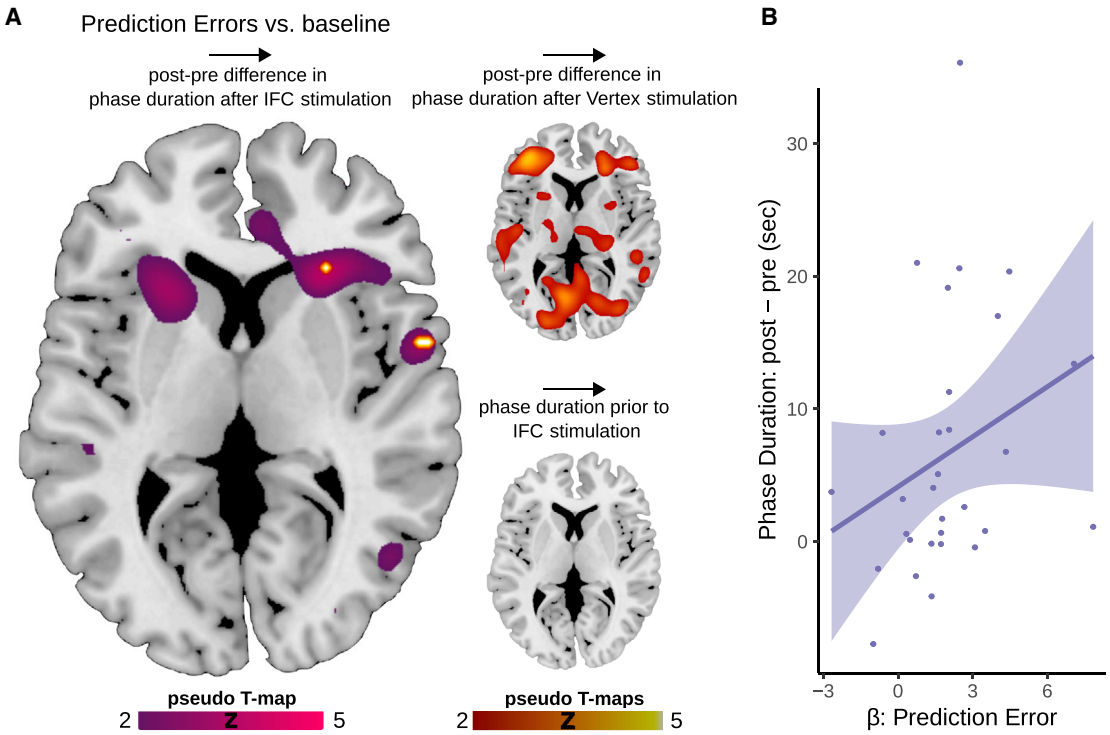


Figure 7. Brain-behavior associations

(A) Whole-brain searchlight decoding revealed that local fMRI activity patterns in IFC successfully predicted inter-individual differences in the effects of virtual IFC lesions on conflict-induced changes in conscious experience (support vector regression; voxels displayed for $T > 2$; $p_{FWE} < 0.05$ highlighted in yellow, left panel). Additional clusters were observed in bilateral temporal pole, left posterior-medial frontal gyrus, right superior medial gyrus, right IPL, and right V1 and left middle orbital gyrus. Voxels in hMT+/V5 did not survive FWE correction across the whole brain. Importantly, support vector regression (SVR) did not reveal any significant association between patterns of BOLD activity in IFC and individual post-pre differences in phase duration after control stimulation at vertex (upper right panel; no voxels surviving FWE correction) or phase duration prior to IFC stimulation (lower right panel; no voxels surviving FWE correction). On the level of behavior, we found that participants with longer pre-stimulation phase duration showed a larger post-pre difference in phase duration after stimulation at IFC ($\rho = 0.44$; $p = 0.02$), but not after control stimulation at vertex ($\rho = -0.1$; $p = 0.6$). This provided additional evidence against the possibility that differences in pre-stimulation baseline may have affected post-pre differences in phase duration irrespective of whether IFC activity was disrupted by TMS.

(B) Participants who represented perceptual conflict more strongly in IFC (correlation coefficient β of perceptual prediction errors to BOLD signals in individual IFC stimulation sites) showed an enhanced reduction of conflict-induced changes in conscious experience when neural activity in IFC was disrupted by TMS ($\rho = 0.42$; $p = 0.02$). Inter-individual differences in the neural representation of perceptual conflict thus provided a possible explanation for non-response to virtual IFC lesions, which was suggested to be present in 9 out of 30 participants by hierarchical agglomerative clustering.

In an additional set of control analyses (Figures S2 and S3), we replicated these findings in linear mixed effects modeling and ruled out exposure effects as well as regression toward the mean as alternative explanations of our TMS results.

In sum, disruption of activity in IFC reduced the frequency of changes in conscious experience during bistable perception. Importantly, we found no evidence for TMS effects on perceptual uncertainty or reporting behavior. These results support the hypothesis that IFC responds to conflicting sensory data by facilitating spontaneous changes in conscious experience, thereby temporarily resolving perceptual conflict.[10,13]

Individual differences in the representation of perceptual conflict predict the effect of virtual IFC lesions on conscious experience

Finally, we asked whether variability in the neural representation of perceptual conflict could predict inter-individual differences in the effect of virtual IFC lesions on conscious experience. We used support vector regression to test whether multi-voxel patterns[31] of conflict-related BOLD activity (E1) contained information on how conscious experience was altered when neural activity in IFC was disrupted (E3). Whole-brain searchlight decoding[37] revealed that localized multi-voxel BOLD activity in IFC, but not V5/hMT+, predicted the individual effects of virtual IFC lesions on phase duration (leave-one-out cross-validation with non-parametric permutation testing;[38] $p_{FWE} < 0.05$; Figure 7A).

In addition, we ensured that neural patterns of conflict representation in IFC selectively predicted the perceptual effects of IFC, but not vertex, stimulation and ruled out baseline differences in phase duration as an alternative explanation of the observed brain behavior association (Figure 7A). Univariate analyses confirmed that virtual IFC lesions reduced the frequency of changes in conscious experience to a greater extent in participants who represented perceptual prediction errors more reliably at IFC stimulation sites (Figure 7B).

At the level of IFC, inter-individual differences in detecting conflicting sensory information were thus directly linked to how

strongly prefrontal brain activity impacted on conscious experience, closing the loop between feedforward and feedback processing of perceptual conflicts.

DISCUSSION

In this work, we found compelling evidence for an active role of IFC in conscious experience: two independent fMRI experiments demonstrated that IFC signals the conflict that emerges between conscious experience and the underlying sensory data. Crucially, TMS-induced virtual lesions revealed that IFC facilitates changes in conscious experience that occur in response to accumulating perceptual conflict.

IFC detects and resolves perceptual conflict during bistable perception

At first glance, our results may seem at odds with the well-established dynamic system account of bistable perception.[39] This view proposes that, in the context of conflicting stimulus information, changes in conscious experience result from local mechanisms, such as inhibition, adaption, or noise.[7] Along these lines, neural activity occurring within sensory cortices may be sufficient to distill unambiguous conscious experiences from conflicting sensory data.

Indeed, our data verify that the contents of conscious experience can be decoded from BOLD activity at the level of V5/hMT+ (Figure S6B). Concurrently, we found that V5/hMT+ generates signals of accumulating perceptual conflict that originate from voxels coding for the currently suppressed stimulus interpretation (Figure 4). In the suppressed voxels, BOLD signals progressively increase prior to changes in conscious experience. In mechanistic terms, these escalating signals of perceptual conflict may be generated by neural populations that gradually escape from inhibition, as adaption reduces the activity in competing neural populations that represent the currently dominant stimulus interpretation. Our results therefore do not contradict the dynamic system account of bistable perception but suggest that the implementational concept of local adaption and inhibition[39] and the algorithmic hypothesis of dynamic conflict accumulation[10,13] are, in fact, complementary.[7]

Importantly, however, our results clearly indicate that the processing of perceptual conflicts does not end at the level of sensory brain regions but reaches prefrontal cortex through feedforward processing from V5/hMT+ to IFC (Figure 2). Crucially, we found that disrupting neural activity in IFC reduces the impact of perceptual conflict on conscious experience (Figures 6 and 7). This indicates that IFC activity is not just a downstream consequence of perceptual events that are realized within hMT+/V5 but actively contributes to the resolution of sensory ambiguity via feedback processes. Together, our findings thus reconcile the feedforward and feedback accounts of bistable perception,[7] suggesting a hybrid computational function of IFC in conscious experience: the detection and resolution of perceptual conflict.

Such a hybrid model[11] not only aligns with previous work suggesting a causal influence of prefrontal feedback on bistable perception[7] but also provides a plausible explanation for the absence of prefrontal activity when perceptual events remain invisible:[16,17] possibly, the capacity of IFC to detect perceptual conflict through feedforward processing may be limited to situations in which the competing states are perceptually distinguishable. When they are not, IFC may fail to read out conflicting stimulus representations,[28] leaving the resolution of perceptual conflict to sensory brain regions.[40] By analogy, our results account for the increase in neural activity observed during unclear or mixed conscious experience,[14] because such perceptual states represent instances of enhanced perceptual conflict and are typically linked to perceptual changes.

Beyond prefrontal cortex, hybrid models based on hierarchical perceptual inference[21,22] have been highly influential in interpreting the role of parietal cortex in bistable perception.[18–20] Here, we found that BOLD activity in parietal brain regions also reflects dynamic changes in perceptual conflict, most notably in the inferior parietal lobule (Figure 2; Table S1). Although pointing to a close connection between IFC and parietal cortex,[7] our results do not provide insight into whether prefrontal and parietal representations of perceptual conflict support redundant or distinct computational functions for bistable perception. Future experiments could resolve this important question by directly comparing the effects of virtual lesions in computationally defined subregions of prefrontal and parietal cortex.

Attention, response behavior, cognitive control, and subjective uncertainty as alternative accounts for IFC's role in bistable perception

IFC has been implicated in various domains of cognition, including attention,[41,42] response behavior,[16] and cognitive control.[43] IFC may therefore exert its influence on conscious experience through one of these alternative cognitive functions, rather than participating directly in the resolution of perceptual conflicts.

First, neural activity prefrontal cortex is known to support sustained attention.[41] During bistable perception, changes in conscious experience occur less frequently when attention is withdrawn.[44] One may therefore argue that virtual IFC lesions may have impaired the participants' ability to attend to the experimental task and, consequently, reduced the frequency of perceptual changes. Yet two observations argue against this proposition: first, we did not observe any effect of virtual IFC lesions on response times (Figure 6C), which closely link to levels of on-task attention.[45] Second, support vector regression revealed that the prefrontal impact on conscious experience is specifically predicted by how strongly IFC activity tracks the accumulation of perceptual conflict (Figure 7A). Sustained attention, in turn, is unlikely to increase systematically over the course of each perceptual phase. It is therefore improbable that the prolongation of perceptual phase durations following virtual IFC lesions can be explained solely on the ground of a global reduction in sustained attention. To directly test this caveat, future work could combine virtual IFC lesions with a parametric modulation of on-task attention during bistable perception.

Second, it has repeatedly been proposed that prefrontal cortex supports only the downstream report of changes of conscious experience that are realized at earlier processing stages.[15,16] Yet a selective impairment of motor behavior seems unable to explain why conflict-induced change in conscious experience is less likely to occur after virtual IFC lesions (Figure 6A), which left response times unaltered. In addition,

our fMRI analyses reveal consistent correlations between IFC activity and accumulating perceptual conflicts while explicitly controlling for the neural correlates of actively reported changes in conscious experience (Figure 2). Based on these findings, we conclude that the often-reported finding of change-related IFC activity is in fact likely to reflect the peak of accumulating perceptual conflict instead of the reported event per se (Figure 3).

Our results therefore align with previous work showing that change-related prefrontal BOLD activity seems to persist when bistable perception is investigated in the absence of active report.[46] Yet in the attempt to control for a range of post-perceptual cognitive phenomena, such as self-monitoring, introspection, cognitive control, or motor behavior,[36] no-report paradigms have produced mixed results with respect to the functional role of prefrontal cortex in conscious experience.[15,16,46–49] Thus, to further substantiate the view that IFC activity is not primarily linked to processes that are situated downstream of perception, future experiments should test whether the prefrontal representation of perceptual conflict and its causal effect on conscious experience are modulated by active report.[15,16,46]

Third, the gradual accumulation of IFC activity toward changes in conscious experience during bistable perception may alternatively be explained by processes related to cognitive control[43] and the anticipation of future events:[50] as the perceptual phase grows longer, participants may become increasingly prone to expect a change in perception. Conversely, they may be more relaxed once an event has occurred. Because average phase durations are quite consistent within individuals (Figure S2), participants may be capable of predicting the approximate timing of changes in conscious experience during bistable perception. Thus, phasic changes in the anticipation of upcoming events may indeed be compatible with the dynamic changes of BOLD observed in IFC.

It may be speculated that, when anticipating a perceptual event, participants could try to voluntarily increase the likelihood of a change in conscious experience.[51] Virtual lesions in dorsolateral prefrontal cortex (DLPFC) have been shown to impair the capacity to exert voluntary control over ambiguous structure-from-motion stimuli.[52] Under this assumption, the observed effect of virtual IFC lesions on conscious experience could be mediated via an impairment of cognitive control, rather than via a mechanism that resolves perceptual conflicts.

In our study, however, participants were naive to the ambiguity in the visual display. Moreover, they were explicitly instructed to passively view the display and report their conscious experience of the stimulus. In contrast to de Graaf et al.,[52] who found an effect of prefrontal TMS only on the voluntary control of bistable perception, we observed clear evidence for a prolongation of phase duration during passive viewing (Figure 6A). Next to differences in sample size (n = 30 versus n = 10) and stimulation protocol (theta-burst versus 1 Hz TMS), this discrepancy may also be due to the target region: while we stimulated IFC and defined stimulation sites based on the neural correlates of perceptual conflict in each participant individually, de Graaf et al.[52] stimulated DLPFC using standard 10/20 electroencephalography coordinates (F4). Yet to fully resolve the question whether anticipation induces prefrontal mechanisms of cognitive control that represent an additional driving factor for spontaneous perceptual changes, future work should use disambiguated stimuli to induce specific temporal expectations and test their effect on conscious experience during bistable perception.

Finally, it may be argued that, instead of coding directly for dynamic changes in perceptual conflict, BOLD activity in IFC may represent ongoing fluctuations in subjective uncertainty.[14] In this paradigm,[53] however, unclear perceptual experiences were extremely rare (Figures S3E and S3F). In addition, an offline rating experiment revealed that subjective uncertainty did not follow the modulation of perceptual conflict by external stimulus information (Figure S7D). Yet online assessments (such as gradual response mappings or secondary markers of confidence derived from eye tracking) could allow future experiments to clarify whether IFC signals ongoing fluctuations in subjective uncertainty beyond the representation of perceptual conflict.

TMS: Side effects and efficacy

On a related note, it may be argued that, due to co-stimulation of facial muscles and cutaneous nerves, prefrontal TMS may have had non-neural effects on cognition that were not controlled for by vertex stimulation. Thus, in addition to the control analyses outlined above, an improved matching of TMS-related side effects could help to rule out that changes in conscious experience associated with virtual IFC lesion may be confounded by global changes in cognitive functions, such as attention, alertness, introspection, response behavior, or cognitive control. Since contralateral stimulation seems suboptimal due to the bilateral representation of perceptual conflict (Table S1), future work could induce muscle contractions via electrodes placed at the IFC stimulation site during sham TMS.

A second TMS-related caveat concerns the general comparability of modulatory effects across regions. Although prefrontal theta-burst stimulation is known to be effective in modulating cognitive function,[54] responsivity has been shown to vary significantly between participants and across stimulation sites.[55,56] This may in part be due to structural differences, such as size, shape, or orientation of the stimulated regions.[55] In this study, however, we found that the efficacy of virtual IFC lesions was predicted by how strongly individual participants represented perceptual conflict in prefrontal cortex (Figure 7). Next to accounting for inter-individual differences in the efficacy of prefrontal theta-burst stimulation, this functional brain-behavior association provided a parsimonious explanation for why conscious experience was unaffected by the control stimulation at vertex, which was not located in the vicinity of any conflict-related brain region (Table S1).

IFC regulates the access of conflicting information into conscious experience

With respect to the role of prefrontal cortex in consciousness, our results speak against the notion that IFC activates merely as a consequence of perceptual events that are generated within sensory cortices.[14–16] As a significant extension, our work associates IFC with a specific computational function for conscious experience: in iterative feedback and feedforward interactions with sensory brain regions, IFC may determine how swiftly conscious experience is updated in response to perceptual conflict.[10,11,13] Intriguingly, this finding aligns with recent neural recordings in monkeys suggesting that prefrontal state

fluctuations precede changes in perception during no-report binocular rivalry.[49]

By controlling the entry of conflicting information into consciousness, IFC may ensure that perception is altered when discrepancies between conscious experience and sensory data have accumulated over time but remains stable when perceptual conflicts are transient. In mechanistic terms, feedback from IFC to sensory cortex could support this function by decreasing the mutual inhibition between competing neural populations,[57] by increasing the rate of adaption,[58] or by upregulating the level of noise in perceptual processing.[59] In these non-exclusive scenarios, feedforward-feedback loops between sensory and prefrontal cortex could benefit perception by facilitating changes in the content of conscious experience only in situations of escalating perceptual conflict.

Beyond the context of regulating the access of conflicting sensory information into conscious experience, IFC may play a similar adaptive role in orienting toward relevant stimuli,[41] in detecting change,[60] or in allocating object-based attention.[42] Altered states of consciousness, such as hallucinations,[61] could therefore relate directly to an impaired processing of sensory information in IFC. Indeed, previous research has associated sensitivity to perceptual conflict with the severity of hallucinations.[33] Correspondingly, functional imaging has repeatedly linked hallucinations to neural activity in IFC.[62,63] Non-invasive brain stimulation of IFC may thus represent a promising new approach in the search for the therapeutic modulation of altered states of consciousness.

In sum, our results demonstrate that prefrontal brain activity is relevant for transforming ambiguous sensory information into unambiguous conscious experiences. At the same time, the underlying dynamics of detecting perceptual conflicts do not seem to be consciously accessible.[10] Thus, although our findings strongly suggest that IFC is causally implicated in the selection of what is consciously perceived, they do not illuminate whether IFC is a necessary component of the neural processes that are jointly sufficient[4,5] or even constitutive[64] for conscious experience per se.

In the search for the neural correlates of consciousness, it is an important question whether the contents of conscious experience can be decoded from specific regions of cerebral cortex.[47,48,65] In line with previous results,[66,67] we found clear evidence for a representation of perceptual content in visual cortex, including V5/hMT+ (Figure S6). Decoding from IFC, in turn, failed to reach statistical significance across the whole brain but showed a trend toward above-chance classification in region-of-interest-based testing (Figure S6B). This difference between V5/hMT+ and IFC may be explained by factors such as mixed selectivity and weak spatial clustering, which may make decoding based on BOLD signals from prefrontal cortex harder than from visual cortex[68] and may become especially relevant in the light of limited statistical power. As an additional decoding-related caveat, our experimental approach may not have been ideal (and was not originally designed) for decoding conscious experience from brain activity, because we did not fully dissociate perceptual contents from behavioral reports.

Indeed, previous studies optimized for decoding have repeatedly shown that prefrontal cortex may indeed encode the contents of conscious experience[47,48,65] and may thus constitute a true neural correlate of consciousness.[47,48] Intersecting computational models of dynamic conflict accumulation[13] with no-report paradigms of bistable perception will enable future research to test whether the contents of conscious experience are represented[48] or multiplexed[69] within the neural correlates of perceptual conflict, creating exciting new opportunities to better understand the role of prefrontal cortex in consciousness.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Stimuli
  - Random dot kinematograms
  - Heterochromatic flicker photometry
  - 2D control stimuli
  - FMRI
  - TMS
  - Brain-behavior associations
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Conventional statistics
  - Computational modeling
  - Model description
  - Model inversion
  - Simulation

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2021.04.043.

ACKNOWLEDGMENTS

V.W., H.S., and J.K. are fellows of the Clinician Scientist Program funded by the Charité – Universitätsmedizin Berlin and the Berlin Institute of Health. This program was initiated and led by Prof. Dr. Duska Dragun to enable physicians to pursue a parallel career in academic research. With great sadness, we have received the news that Prof. Dragun passed away on December 28th of 2020. We dedicate this publication to her as a mentor, friend, role model, and stellar scientist. A.E. is a fellow of the Einstein Center for Neurosciences and the Bernstein Center for Computational Neurosciences Berlin. P.S. is funded by the German Research Foundation (STE 1430/8-1) and the German Ministry for Research and Education (ERA-NET NEURON program; 01EW2007A). The authors thank Andreas Kleinschmidt and Guido Hesselmann for helpful comments on an earlier version of the manuscript.

AUTHOR CONTRIBUTIONS

V.W. and P.S. conceptualized the study. V.W. designed the experiments. V.W., M.F., M.C., A.-L.E., K.K., H.S., and J.K. collected the data. V.W. and P.S. wrote the initial draft and edited the manuscript. All authors reviewed the manuscript.

DECLARATIONS OF INTEREST

The authors declare no competing interests.

Received: February 2, 2021
Revised: March 22, 2021
Accepted: April 19, 2021
Published: May 13, 2021

REFERENCES

1. Michel, M., Beck, D., Block, N., Blumenfeld, H., Brown, R., Carmel, D., Carrasco, M., Chirimuuta, M., Chun, M., Cleeremans, A., et al. (2019). Opportunities and challenges for a maturing science of consciousness. Nat. Hum. Behav. *3*, 104–107.
2. Sohn, E. (2019). Decoding the neuroscience of consciousness. Nature *571*, S2–S5.
3. Dehaene, S., Lau, H., and Kouider, S. (2017). What is consciousness, and could machines have it? Science *358*, 486–492.
4. Odegaard, B., Knight, R.T., and Lau, H. (2017). Should a few null findings falsify prefrontal theories of conscious perception? J. Neurosci. *37*, 9593–9602.
5. Boly, M., Massimini, M., Tsuchiya, N., Postle, B.R., Koch, C., and Tononi, G. (2017). Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? Clinical and neuroimaging evidence. J. Neurosci. *37*, 9603–9613.
6. Hohwy, J., and Seth, A. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness. Philos. Mind Sci. *1*, 3.
7. Brascamp, J., Sterzer, P., Blake, R., and Knapen, T. (2018). Multistable perception and the role of the frontoparietal cortex in perceptual inference. Annu. Rev. Psychol. *69*, 77–103.
8. Blake, R., and Logothetis, N. (2002). Visual competition. Nat. Rev. Neurosci. *3*, 13–21.
9. Leopold, D.A., and Logothetis, N.K. (1999). Multistable phenomena: changing views in perception. Trends Cogn. Sci. *3*, 254–264.
10. Hohwy, J., Roepstorff, A., and Friston, K. (2008). Predictive coding explains binocular rivalry: an epistemological review. Cognition *108*, 687–701.
11. Sterzer, P., Kleinschmidt, A., and Rees, G. (2009). The neural bases of multistable perception. Trends Cogn. Sci. *13*, 310–318.
12. Lumer, E.D., Friston, K.J., and Rees, G. (1998). Neural correlates of perceptual rivalry in the human brain. Science *280*, 1930–1934.
13. Weilnhammer, V., Stuke, H., Hesselmann, G., Sterzer, P., and Schmack, K. (2017). A predictive coding account of bistable perception - a model-based fMRI study. PLoS Comput. Biol. *13*, e1005536.
14. Knapen, T., Brascamp, J., Pearson, J., van Ee, R., and Blake, R. (2011). The role of frontal and parietal brain areas in bistable perception. J. Neurosci. *31*, 10293–10301.
15. Frässle, S., Sommer, J., Jansen, A., Naber, M., and Einhäuser, W. (2014). Binocular rivalry: frontal activity relates to introspection and action but not to perception. J. Neurosci. *34*, 1738–1747.
16. Brascamp, J., Blake, R., and Knapen, T. (2015). Negligible fronto-parietal BOLD activity accompanying unreportable switches in bistable perception. Nat. Neurosci. *18*, 1672–1678.
17. Zou, J., He, S., and Zhang, P. (2016). Binocular rivalry from invisible patterns. Proc. Natl. Acad. Sci. USA *113*, 8408–8413.
18. Carmel, D., Walsh, V., Lavie, N., and Rees, G. (2010). Right parietal TMS shortens dominance durations in binocular rivalry. Curr. Biol. *20*, R799–R800.
19. Kanai, R., Bahrami, B., and Rees, G. (2010). Human parietal cortex structure predicts individual differences in perceptual rivalry. Curr. Biol. *20*, 1626–1630.
20. Zaretskaya, N., Thielscher, A., Logothetis, N.K., and Bartels, A. (2010). Disrupting parietal function prolongs dominance durations in binocular rivalry. Curr. Biol. *20*, 2106–2111.
21. Kanai, R., Carmel, D., Bahrami, B., and Rees, G. (2011). Structural and functional fractionation of right superior parietal cortex in bistable perception. Curr. Biol. *21*, R106–R107.
22. Megumi, F., Bahrami, B., Kanai, R., and Rees, G. (2015). Brain activity dynamics in human parietal regions during spontaneous switches in bistable perception. Neuroimage *107*, 190–197.
23. Friston, K. (2005). A theory of cortical responses. Philos. Trans. R. Soc. Lond. B Biol. Sci. *360*, 815–836.
24. Pastukhov, A., Vonau, V., and Braun, J. (2012). Believable change: bistable reversals are governed by physical plausibility. J. Vis. *12*, 17.
25. Knill, D.C., and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci. *27*, 712–719.
26. Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. Front. Psychol. *3*, 96.
27. Rosa, M.J., Bestmann, S., Harrison, L., and Penny, W. (2010). Bayesian model selection maps for group studies. Neuroimage *49*, 217–224.
28. Heekeren, H.R., Marrett, S., Bandettini, P.A., and Ungerleider, L.G. (2004). A general mechanism for perceptual decision-making in the human brain. Nature *431*, 859–862.
29. Krug, K., Cicmil, N., Parker, A.J., and Cumming, B.G. (2013). A causal role for V5/MT neurons coding motion-disparity conjunctions in resolving perceptual ambiguity. Curr. Biol. *23*, 1454–1459.
30. Friston, K.J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. Neuroimage *19*, 1273–1302.
31. Haynes, J.D., and Rees, G. (2006). Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. *7*, 523–534.
32. Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as Bayesian inference. Annu. Rev. Psychol. *55*, 271–304.
33. Weilnhammer, V., Röd, L., Eckert, A.L., Stuke, H., Heinz, A., and Sterzer, P. (2020). Psychotic experiences in schizophrenia and sensitivity to sensory evidence. Schizophr. Bull. *46*, 927–936.
34. van Gaal, S., Ridderinkhof, K.R., Scholte, H.S., and Lamme, V.A. (2010). Unconscious activation of the prefrontal no-go network. J. Neurosci. *30*, 4143–4150.
35. Huang, Y.Z., Edwards, M.J., Rounis, E., Bhatia, K.P., and Rothwell, J.C. (2005). Theta burst stimulation of the human motor cortex. Neuron *45*, 201–206.
36. Tsuchiya, N., Wilke, M., Frässle, S., and Lamme, V.A.F. (2015). No-report paradigms: extracting the true neural correlates of consciousness. Trends Cogn. Sci. *19*, 757–770.
37. Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. Proc. Natl. Acad. Sci. USA *103*, 3863–3868.
38. Schmack, K., Burk, J., Haynes, J.D., and Sterzer, P. (2016). Predicting subjective affective salience from cortical responses to invisible object stimuli. Cereb. Cortex *26*, 3453–3460.
39. Wilson, H.R. (2007). Minimal physiological conditions for binocular rivalry and rivalry memory. Vision Res. *47*, 2741–2750.
40. Xu, H., Han, C., Chen, M., Li, P., Zhu, S., Fang, Y., Hu, J., Ma, H., and Lu, H.D. (2016). Rivalry-like neural activity in primary visual cortex in anesthetized monkeys. J. Neurosci. *36*, 3231–3242.
41. Corbetta, M., Patel, G., and Shulman, G.L. (2008). The reorienting system of the human brain: from environment to theory of mind. Neuron *58*, 306–324.
42. Baldauf, D., and Desimone, R. (2014). Neural mechanisms of object-based attention. Science *344*, 424–427.
43. Aron, A.R., Robbins, T.W., and Poldrack, R.A. (2014). Inhibition and the right inferior frontal cortex: one decade on. Trends Cogn. Sci. *18*, 177–185.
44. Alais, D., van Boxtel, J.J., Parker, A., and van Ee, R. (2010). Attending to auditory signals slows visual alternations in binocular rivalry. Vision Res. *50*, 929–935.
45. Prado, J., Carp, J., and Weissman, D.H. (2011). Variations of response time in a selective attention task are linked to variations of functional connectivity in the attentional network. Neuroimage *54*, 541–549.

46. Lumer, E.D., and Rees, G. (1999). Covariation of activity in visual and prefrontal cortex associated with subjective visual perception. Proc. Natl. Acad. Sci. USA *96*, 1669–1673.

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

48. Kapoor, V., Dwarakanath, A., Safavi, S., Werner, J., Besserve, M., Panagiotaropoulos, T.I., and Logothetis, N.K. (2020). Decoding the contents of consciousness from prefrontal ensembles. bioRxiv. https://doi.org/10.1101/2020.01.28.921841.

49. Dwarakanath, A., Kapoor, V., Werner, J., Safavi, S., Fedorov, L.A., Logothetis, N.K., and Panagiotaropoulos, T.I. (2020). Prefrontal state fluctuations control access to consciousness. bioRxiv. https://doi.org/10.1101/2020.01.29.924928.

50. Dürschmid, S., Reichert, C., Hinrichs, H., Heinze, H.J., Kirsch, H.E., Knight, R.T., and Deouell, L.Y. (2019). Direct evidence for prediction signals in frontal cortex independent of prediction error. Cereb. Cortex *29*, 4530–4538.

51. Klink, P.C., van Ee, R., Nijs, M.M., Brouwer, G.J., Noest, A.J., and van Wezel, R.J. (2008). Early interactions between neuronal adaptation and voluntary control determine perceptual choices in bistable vision. J. Vis. *8*, 16.1–18.

52. de Graaf, T.A., de Jong, M.C., Goebel, R., van Ee, R., and Sack, A.T. (2011). On the functional relevance of frontal cortex for passive and voluntarily controlled bistable vision. Cereb. Cortex *21*, 2322–2331.

53. Weilnhammer, V.A., Ludwig, K., Hesselmann, G., and Sterzer, P. (2013). Frontoparietal cortex mediates perceptual transitions in bistable perception. J. Neurosci. *33*, 16009–16015.

54. Lowe, C.J., and Hall, P.A. (2018). Reproducibility and sources of interindividual variability in the responsiveness to prefrontal continuous theta burst stimulation (cTBS). Neurosci. Lett. *687*, 280–284.

55. Huang, G., and Mouraux, A. (2015). MEP latencies predict the neuromodulatory effect of cTBS delivered to the ipsilateral and contralateral sensorimotor cortex. PLoS ONE *10*, e0133893.

56. Lowe, C.J., Manocchio, F., Safati, A.B., and Hall, P.A. (2018). The effects of theta burst stimulation (TBS) targeting the prefrontal cortex on executive functioning: a systematic review and meta-analysis. Neuropsychologia *111*, 344–359.

57. Stephan, K.E., Kasper, L., Harrison, L.M., Daunizeau, J., den Ouden, H.E., Breakspear, M., and Friston, K.J. (2008). Nonlinear dynamic causal models for fMRI. Neuroimage *42*, 649–662.

58. Toppino, T.C., and Long, G.M. (2015). Time for a change: what dominance durations reveal about adaptation effects in the perception of a bi-stable reversible figure. Atten. Percept. Psychophys. *77*, 867–882.

59. Moreno-Bote, R., Rinzel, J., and Rubin, N. (2007). Noise-induced alternations in an attractor network model of perceptual bistability. J. Neurophysiol. *98*, 1125–1139.

60. Garrido, M.I., Kilner, J.M., Stephan, K.E., and Friston, K.J. (2009). The mismatch negativity: a review of underlying mechanisms. Clin. Neurophysiol. *120*, 453–463.

61. Corlett, P.R., Horga, G., Fletcher, P.C., Alderson-Day, B., Schmack, K., and Powers, A.R., 3rd. (2019). Hallucinations and strong priors. Trends Cogn. Sci. *23*, 114–127.

62. Sommer, I.E.C., Diederen, K.M., Blom, J.D., Willems, A., Kushan, L., Slotema, K., Boks, M.P., Daalman, K., Hoek, H.W., Neggers, S.F., and Kahn, R.S. (2008). Auditory verbal hallucinations predominantly activate the right inferior frontal area. Brain *131*, 3169–3177.

63. Powers, A.R., Mathys, C., and Corlett, P.R. (2017). Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. Science *357*, 596–600.

64. Michel, M., and Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. Philos. Mind Sci. *1*, 8.

65. Wang, M., Arteaga, D., and He, B.J. (2013). Brain mechanisms for simple perception and bistable perception. Proc. Natl. Acad. Sci. USA *110*, E3350–E3359.

66. Brouwer, G.J., and van Ee, R. (2007). Visual cortex allows prediction of perceptual states during ambiguous structure-from-motion. J. Neurosci. *27*, 1015–1023.

67. Schmack, K., Gòmez-Carrillo de Castro, A., Rothkirch, M., Sekutowicz, M., Rössler, H., Haynes, J.D., Heinz, A., Petrovic, P., and Sterzer, P. (2013). Delusions and the role of beliefs in perceptual inference. J. Neurosci. *33*, 13701–13712.

68. Bhandari, A., Gagne, C., and Badre, D. (2018). Just above chance: is it harder to decode information from prefrontal cortex hemodynamic activity patterns? J. Cogn. Neurosci. *30*, 1473–1498.

69. Hesse, J.K., and Tsao, D.Y. (2020). A new no-report paradigm reveals that face cells encode both consciously perceived and suppressed stimuli. eLife *9*, e58360.

70. Brainard, D.H. (1997). The Psychophysics Toolbox. Spat. Vis. *10*, 433–436.

71. Morgan, M.J., and Thompson, P. (1975). Apparent motion and the Pulfrich effect. Perception *4*, 3–18.

72. O'Shea, J., and Walsh, V. (2007). Transcranial magnetic stimulation. Curr. Biol. *17*, R196–R199.

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

74. Rossini, P.M., Burke, D., Chen, R., Cohen, L.G., Daskalakis, Z., Di Iorio, R., Di Lazzaro, V., Ferreri, F., Fitzgerald, P.B., George, M.S., et al. (2015). Non-invasive electrical and magnetic stimulation of the brain, spinal cord, roots and peripheral nerves: Basic principles and procedures for routine clinical and research application. An updated report from an I.F.C.N. Committee. Clin. Neurophysiol. *126*, 1071–1107.

75. Schicktanz, N., Fastenrath, M., Milnik, A., Spalek, K., Auschra, B., Nyffeler, T., Papassotiropoulos, A., de Quervain, D.J., and Schwegler, K. (2015). Continuous theta burst stimulation over the left dorsolateral prefrontal cortex decreases medium load working memory performance in healthy humans. PLoS ONE *10*, e0120640.

76. Suppa, A., Ortu, E., Zafar, N., Deriu, F., Paulus, W., Berardelli, A., and Rothwell, J.C. (2008). Theta burst stimulation induces after-effects on contralateral primary motor cortex excitability in humans. J. Physiol. *586*, 4489–4500.

77. Valchev, N., Tidoni, E., Hamilton, A.F.C., Gazzola, V., and Avenanti, A. (2017). Primary somatosensory cortex necessary for the perception of weight from other people's action: A continuous theta-burst TMS experiment. Neuroimage *152*, 195–206.

78. Weilnhammer, V.A., Ludwig, K., Sterzer, P., and Hesselmann, G. (2014). Revisiting the Lissajous figure as a tool to study bistable perception. Vision Res. *98*, 107–112.

79. Friston, K.J., and Stephan, K.E. (2007). Free-energy and the brain. Synthese *159*, 417–458.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw and analyzed data	This paper	https://doi.org/10.17605/OSF.IO/YKM6X
Custom R markdown code	This paper	https://doi.org/10.17605/OSF.IO/YKM6X
Custom MATLAB code	This paper	https://doi.org/10.17605/OSF.IO/YKM6X
Software and algorithms		
MATLAB	https://www.mathworks.com/	RRID: SCR:001622
RStudio	https://www.rstudio.com/	RRID: SCR:000432
lme4	Rstudio	RRID: SCR:015654
afex	Rstudio	N/A
BayesFactor	Rstudio	N/A
lmBF	Rstudio	N/A
TAPAS toolbox	https://www.tnu.ethz.ch/en/software/tapas	N/A
SPM toolbox	https://www.fil.ion.ucl.ac.uk/spm/software/spm12/	RRID: SCR:007037
SPM anatomy toolbox	https://www.fz-juelich.de/portal/DE/Home/home_node.html	RRID: SCR:013273
MarsBaR region of interest toolbox for SPM	http://marsbar.sourceforge.net/	RRID: SCR:009605

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Veith Weilnhammer (veith-andreas.weilnhammer@charite.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All data and code associated with this study are available on OSF: https://osf.io/ykm6x/.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Experiment E1 consisted of a behavioral pre-test (Runs 1 and 2) and an fMRI-experiment (Runs 3 - 5). We recruited a total of 35 participants. Based on the behavioral pre-test, we excluded two participants who performed at chance-level when discriminating the direction of rotation of a fully disambiguated structure-from-motion stimulus. Thus, 33 participants took part in the fMRI-experiment (21 female, mean age: 27.3 ± 1.42 years). All participants agreed to be contacted later for a follow-up experiment using TMS (E3, see below).

Experiment E2 consisted of a behavioral pre-test and a fMRI-experiment. We recruited a total of 23 participants. We excluded three participants who performed at chance-level when discriminating the direction of rotation of a fully disambiguated structure-from-motion stimulus. The final sample thus consisted in 20 participants (11 female, mean age: 27.7 ± 0.98 years).

For experiment E3, we re-invited the participants from E1 to two TMS session scheduled on consecutive days. From this group, one participant could not be re-contacted at the time of the TMS-experiment. Two further participants did not tolerate the TMS-procedure. The final TMS-sample thus consisted of 30 participants (19 female, mean age: 27.33 ± 1.56 years).

All participants were right-handed, showed corrected-to-normal vision, had no prior neurological or psychiatric medical history and gave written informed consent prior to taking part in the study. All procedures were approved by the ethics committee at Charité Berlin.

METHOD DETAILS

Stimuli

Stimuli were presented using Psychtoolbox 3[70] and MATLAB R2007b (behavioral pre-test: CRT-Monitor at 60 Hz, 1042 × 768 pixels, 60 cm viewing distance and 30.28 pixels per degree visual angle; fMRI: LCD-Monitor at 60 Hz, 1280 × 1024 pixels, 160 cm viewing distance and 90.96 pixel per degree visual angle; TMS: LCD-Monitor at 60 Hz, 1280 × 1024 pixels, 60 cm viewing distance and 37.82 pixels per degree visual angle).

Random dot kinematograms

Throughout E1, E2 am E3, participants indicated their perception of a discontinuous random-dot kinematogram (RDK, see Video S1). In this stimulus, random dots distributed on two intersecting rings induce the perception of a spherical object rotating either left- or rightward around a vertical axis[24] (diameter: 15.86°, rotational speed: 12 s per rotation, rotations per block: 10, individual dot size: 0.12°). Each run consisted of six blocks of visual stimulation (120 s), separated by fixation intervals (behavioral pre-tests: 5 s; fMRI- and TMS-experiments: 10 s).

Depending on the experimental condition (Figure S1), the RDK could appear in three configurations: Complete ambiguity, levels of graded ambiguity and complete disambiguation. Complete ambiguity was achieved by presenting identical stimuli to the two eyes. This induced periodic changes in conscious experience (also dubbed *endogenous transitions*) between left- and rightward rotation (i.e., bistable perception).

For complete disambiguation, we used red-blue filter glasses (left eye: red channel, right eye: blue channel) to attach a stereo-disparity signal (1.8° visual angle) to all dots on the stimulus surface. By inverting the direction of rotation, we created stimulus-driven or *exogenous* changes in conscious experience.

During graded ambiguity,[33] we varied the proportion of disambiguated stimulus dots between 15%, 30%, 45%, 60%, 75% and 100% (conditions D1 to D6). This variation in the signal-to-ambiguity ratio parametrically modulated the perceptual conflict between conscious experience and visual stimulation: We predicted that perceptual conflict (and associated neural activity) should be enhanced during incongruent as compared to congruent perceptual states. Furthermore, this enhancement should increase at higher signal-to-ambiguity ratios. During runs with graded ambiguity, conditions D1 to D6 appeared in random order. Within each block, we introduced random changes in the direction of disambiguation (i.e., whether the parametric 3D cues enforced rightward or leftward rotation). The individual frequency of exogenous stimulus changes during graded ambiguity was determined based on the frequency of conflict-induced changes in conscious experience during full ambiguity.

Importantly, participants were naive to the potential ambiguity in the visual display and explicitly instructed to passively experience the stimulus, reporting their perception via button-presses (right index-finger: rotation of the front-surface to the left; right ring-finger: rotation to the right; right middle-finger: unclear direction of rotation) on a USB keyboard or a MRI-compatible button-box, respectively.

We based our behavioral analyses on perceptual events as reported by the participants. Since the RDK is not depth-symmetric over all rotational angles,[24,53] changes in conscious experience occurred only at overlapping configuration of the stimulus (Figures S3A and S3B). We thus corrected the timing of perceptual events to the last overlapping configuration of the stimulus preceding the button-press, representing the perceptual time-course as a discrete sequence of perceptual states (rotation of the front-surface to the right/left and unclear direction of rotation).

To describe the temporal dynamics of bistable perception, we computed *average phase durations* (the time spent between two changes in conscious experience, i.e., multiples of the 1.5 s inter-overlap-interval). The content of conscious experience was reflected by the dependent variables *directed bias* (the percentage of rightward perceptual states relative to the total number of perceptual states), *absolute bias* (the absolute difference between the absolute bias and chance level at 50%) and the percentage of *uncertain states*. To characterize processes involved in the behavioral report of perception, we computed *response times* by subtracting the timing of the last preceding overlapping configuration from the timing of the behavioral responses indicating a perceptual event. The impact of sensory data on perception was depicted by the dependent variable *perceptual congruency* (percentage of perceptual states congruent with the disambiguating 3D signal).

Heterochromatic flicker photometry

When using filter glasses (Experiment E2), the perceived direction rotation of RDKs can be biased by differences in the subjective luminance between red and blue (Pulfrich effect[71]). To estimate subjective equiluminance, we presented red and blue circles (diameter: 6.45°) alternating at a frequency of 15 Hz. Differences in subjective luminance of red and blue stimuli led to the experience of flicker. Participants reduced the flicker by adjusting the luminance of the red stimulus initially presented at a random luminance between 0% and 255% relative to the blue stimulus presented at a fixed luminance. Average equiluminance estimated across 10 such trials determined the monitor- and participant-specific luminance of the red- and blue-channels (average blue-channel luminance relative to red-channel: 2.02 ± 0.09).

2D control stimuli

At higher levels of signal-to-ambiguity, perceptual states were less likely to be incongruent with the disambiguating stimulus information. Hence, increments in the signal-to-ambiguity ratio increased the temporal imbalance between congruent and incongruent

perceptual states. To test for potential confounds introduced by temporal imbalance, we constructed a 2D control version (E2, Run 8) of the bistable RDK (identical stimulus diameter, number, size and speed of dots). Participants reported the direction of planar, horizontal 2D motion. For each participant, changes in the direction of planar dot motion were determined by both the temporal imbalance between congruent and incongruent perceptual phases (separately for conditions D1-D6) as well as the average frequency of changes in conscious experience observed in the main experiment (E2, Runs 5-7). We randomized the motion direction associated with reduced presentation-time.

FMRI

Acquisition and preprocessing

For E1 and E2, we recorded a T1-weighted MPRAGE sequence (voxel size 1 × 1 x 1 mm) for anatomical images and used T2-weighted gradient-echo planar imaging (TR 2000 ms, TE 25 ms, voxel-size 3 × 3 x 3 mm) to obtain a total of 400 BOLD images per run on a Siemens Prisma 3-Tesla-MRI-system (64-channel coil). Our pre-processing routine was carried out within SPM12 and consisted in slice time correction with reference to the middle slice, standard realignment, coregistration and normalization to MNI stereotactic space using unified segmentation. For standard analyses and support vector regression,[38] we applied spatial smoothing with 8 mm full-width at a half-maximum isotropic Gaussian kernel. For the analysis of voxel biases and support vector classification, we used unsmoothed data.

General linear models

To test for the neural correlates of perceptual conflict during sensory ambiguity in E1 and E2, we extracted dynamic perceptual prediction errors from the predictive-coding (PC) model of bistable perception,[13] which was inverted based on behavioral data. This *model-based* fMRI approach (GLM-PC) defined visual stimulation by stick-regressors aligned to the overlapping configurations of the structure-from-motion stimulus (*overlaps*). Relative to the *overlaps*, we defined two parametric regressors ordered as follows: (1) perceptual *changes* (binary; 0: no change, 1: change) and (2) absolute *prediction errors* (continuous, ranging from 0 to 1). For the analysis of direction-specific effects in voxel biases within V5/hMT+, the overlaps and the associated parametric modulators were modeled separately according to the current perceptual state (left- versus- rightward rotation). In addition to standard GLMs, we performed Bayesian second-level statistics[27] to compare the explanatory power between change-related models and prediction-error related models with regard to BOLD activity in IFC. To this end, we deleted one of the two parametric modulators in GLM-PC, creating Log-Evidence-Maps for the two degraded models (‘‘PE only’’ versus ‘‘Change only’’; z-scored parametric modulators).

In E2, we used an additional GLM (GLM-Congruency) to analyze BOLD activity during graded ambiguity independently of the assumptions inherent in the predictive-coding model of bistable perception. Next to perceptual changes (*T*, stick-function), this GLM represented perceptual states by box-car regressors defined according to two factors: First, perceptual states could be congruent (*C1*) or incongruent (*C2*) to conscious experience. Second, visual stimulation varied across six levels of signal-to-ambiguity (*D1-D6*). The GLM’s design matrix contained all combinations between the two factors ([C1D1 C1D2 (…) C1D6 C2D1 C2D2 (…) C2D6 T]). By analogy, we tested for a potential effect of temporal imbalances between congruent and incongruent perceptual phases in the fMRI control-experiment (run E2(8)). GLM-Control defined prolonged (*A1*) and shortened (*A2*) perceptual phases separately for all levels of temporal imbalance (Levels *I1* to *I6*; design-matrix = [A1I1 A1I2 (…) A1I6 A2I1 A2I2 (…) A2I6 T]).

In E3, we identified individual IFC stimulation sites based on the fMRI data acquired in E1. To delineate IFC independently of the assumptions inherent in the predictive-coding model of bistable perception (see GLM-PC), we adopted the conventional change-related fMRI approach to IFC,[14,16,53] representing endogenous perceptual changes as stick-functions and visual stimulation as a box-car regressor (GLM-Changes).

In all GLMs, we convolved the outlined regressors with the canonical hemodynamic response function (SPM12), added six rigid-body realignment parameters as nuisance covariates, applied high-pass filtering at 1/128 Hz and computed first-level one-sample t tests against baseline. On the second-level, the resulting images were submitted to second-level one-sample t tests (GLM-PC) or full factorial models (GLM-Congruency and GLM-Control). Second-level results were thresholded at $p < 0.05$ (FWE-corrected across the whole brain; SVC within orthogonal activation maps for GLM-Congruency). For Bayesian second-level statistics,[27] we display second-level results at an exceedance probability of 95% for ‘‘PE only.’’

Stimulation sites at IFC and vertex

In E3, we defined individual IFC coordinates for neuronavigated TMS based on BOLD activity associated with perceptual changes (data acquired during E1). Using GLM-Changes, we identified the peak voxel for ‘‘Changes vs. baseline’’ (first-level contrast at $p < 0.005$, uncorrected) within a literature-based IFC search-sphere (radius = 5 mm; center = [57 17 10]). This location was motivated by the neural correlates of conflict-driven as opposed to stimulus-driven changes in conscious experience in a closely related structure-from-motion stimulus.[13] Across participants, average stimulation sites were located at MNI = [55.6 ± 0.4 15.5 ± 0.39 10 ± 0.49].

By informing the TMS-intervention based on the conventional change-related approach to IFC,[14,16] we delineated the IFC stimulation site independently of our computational model of bistable perception.[13] As shown above, change-related activity coincided with the neural correlates of perceptual prediction errors (Figure 3A), which had more explanatory power with regard to BOLD signals in IFC[13] (Figure S3B). As expected, activity in the IFC stimulation site was thus highly correlated to perceptual prediction errors (average regression coefficients in spherical ROIs of 10 mm radius around individual coordinates: $\beta = 1.79 \pm 0.38$; $T(32) = 4.75$, $p = 4.15 \times 10^{-5}$, $BF_{10} = 565.3$).

The control stimulation site at vertex was determined by anatomical (T1) scans (MNI = [0 −25 85]). Given the spatial resolution of neuronavigated TMS,[72] vertex-TMS was extremely unlikely to exert local effects on any additional neural correlates of bistable perception (Table S1).

Regions-of-interest (ROI)

All ROIs were defined independently of the computational model of bistable perception[13] outlined in the STAR Methods section Computational modeling. With respect to IFC, we defined spherical ROIs (radius: 10 mm) around the individual IFC-TMS coordinates (see above). To delineate V5/hMT+, we constructed a search sphere (radius: 5 mm) around the peak-voxel for the second-level contrast "Visual Stimulation vs. baseline" (GLM-Changes, $p_{FWE} < 0.05$) within an anatomical mask for V5/hMT+.[73] Based on this search sphere, we constructed individual V5/hMT+ ROIs (radius: 10 mm) centered around the individual peak coordinates of the corresponding first-level contrast ($p < 0.005$, uncorrected). Within these ROIs, we defined voxels with biases for righward- and leftward perceptual states (L- and R-population) by thresholding the contrasts for "left vs. rightward perceptual states" and vice versa at a T-value of 1 (GLM-PC).

Functional ROI-based analyses including finite impulse response (FIR) models were carried out in MarsBaR (http://marsbar.sourceforge.net/). FIR models were estimated for a time window of −7.5 s until 14 s surrounding reported changes in conscious experience. The time points of changes in conscious experience (vertical dotted line in Figure S4) were defined by the last overlapping stimulus configuration that preceded the respective button press. Given a TR of 2 s and an inter-overlap interval of 1.5 s, we estimated the FIR models in time bins determined by the effective sampling rate of 0.5 s. Fits were computed using local polynomial regression fitting.

Anatomic Labeling

All anatomic labels were obtained from the Anatomy Toolbox.[73] The IFC was defined by the combination of anterior insula and inferior frontal gyrus (past triangularis and pars opercularis).

TMS

In E3, we used TMS with a theta-burst protocol to induce virtual lesions in the two stimulation sites (i.e., the target-region in right-hemispherical IFC and the control-region at the cranial vertex, see above). TMS was delivered in two separate TMS-sessions on two consecutive days. We counterbalanced the order of IFC- versus vertex-TMS across participants. Participants performed two runs of the experiment prior to TMS and two runs immediately after TMS.

We delivered TMS with a focal, figure-of-eight-shaped coil equipped with active cooling. Pulses where generated using a standard MagPro R30 stimulator (MagVenture Ltd, Farum, Denmark). Stimulation was guided by online neuro-navigation based on individual target regions projected onto the participants' T1 scans using the Localite TMS Navigator (Localite GmbH, Bonn, Germany) with an optical tracking camera PolarisVicra (Northern Digital, Ontario, Canada).

The coil was positioned tangentially to the subjects' head and adjusted such that the electric current in the center of the coil would run perpendicular to the course of the inferior frontal sulcus. Prior to each session, we identified individual resting motor thresholds (rMTs) for the right first dorsal musculus interosseous (FDI) by stimulation of left-hemispherical motor cortex.[74] The coil was held tangentially to the subject's skull at a 45° angle to the parasagittal line (4 cm lateral and 1 cm anterior to the vertex). The search for the hot-spot was additionally guided through the optical tracking system in order to locate the hand-knob. In order to find the rMT hot-spot, we started with 55% Maximum Simulator Output (MSO) and increased the intensity in 5% steps. If a motor evoked potential (MEPs) was elicited, adjustments were made in 1% steps. Pulses for MT search were delivered with a minimum of 5 s delay in order to avoid any change in excitability due to repeated stimulation. MEPs were recorded from the right FDI using self-adhesive gel electrodes in a standard belly-tendon fashion. RMTs were defined as the percentage of maximum stimulator output needed to evoke 50μV MEPs peak-to-peak in five out of ten consecutive trials (average rMT in vertex sessions: 41.67 ± 1.12% MSO; IFC sessions: 40.9 ± 1.15% MSO; paired t test: $T(29) = 0.89$, $p = 0.38$, $BF_{10} = 0.28$).

The theta-burst TMS-protocol consisted in a total 600 pulses applied within 40 s (50-Hz bursts with three pulses applied in intervals of 200 ms) at an intensity of 80% rMT. Stimulation parameters were in line with published safety guidelines and were chosen to produce a decrease in cortical excitability[35,75–77] throughout the 25 min test-phase following TMS.

Since stimulation intensities were determined relative to rMT, inter-individual differences in the surface-to-target distance between IFC (20.88 ± 0.55 mm) and motor cortex (24.06 ± 0.66 mm) may therefore have caused stronger prefrontal TMS-effects for participants in whom the IFC was relatively closer to the skull's surface (and vice versa). However, the absolute between-region difference in surface-to-target distance was relatively small (3.93 ± 0.5 mm). In comparison to motor cortex, individual IFC stimulation sites were closer to the skull's surface ($T(56.1) = -3.43$, $p = 1.15 \times 10^{-3}$, $BF_{10} = 28.06$, paired t test). Importantly, individual surface-to-target distances were positively correlated between IFC and motor cortex ($\rho = 0.37$, $p = 0.04$, Spearman correlation), arguing in favor of the notion that stimulation intensities were transferable between regions.

During IFC stimulation, we routinely observed co-stimulation of the temporal muscle, leading to involuntary up- and down-movements of the jaw. In some participants, we also observed co-stimulation of the orbicularis oculi muscle, leading to involuntary blinking of the right eye. To ameliorate the potential distress that may be caused by co-stimulation of facial muscles, participants were extensively briefed about this side-effect, including an explanation of its physiological mechanism, harmlessness and limitation to the time of stimulation. Immediately prior to stimulation, we instructed participants to relax their facial muscles, keeping their teeth apart and their mouth slightly open. No participant had to be excluded because of not tolerating the co-activation of facial muscles during TMS to IFC.

Co-stimulation of cutaneous nerves is a second potential side-effect of TMS, which can lead to painful sensations at the stimulation site. One participant had to be excluded because she experienced pain during both vertex- and IFC-stimulation, which led her to abort the latter. We excluded one additional participant who fainted during rMT estimation. In total, intolerance to our TMS procedures thus led to the exclusion of two participants.

Non-responders to the TMS intervention were identified using complete-linkage euclidian-distance hierarchical agglomerative clustering.[56] The criterion variable was defined by the prolongation of phase duration (sec) associated virtual IFC-lesions relative to the vertex condition (Figure 6A).

Brain-behavior associations

To relate inter-individual differences in the representation of perceptual conflict to the effects of virtual IFC-lesions on conscious experience, we assessed brain-behavior associations in both a *univariate* and a *multivariate* approach. In the *univariate* approach, we conducted a standard ROI-based analysis, extracting individual regression coefficients β of perceptual prediction errors to BOLD signals from individual IFC stimulation sites. We then used Spearman correlation to test whether individual β estimates predicted the behavioral effects of virtual lesions.

In *multivariate* pattern analysis, we predicted the effects of virtual IFC-lesions based on localized pattern of BOLD activity measured across the whole brain. Using searchlight decoding,[37] we extracted multidimensional pattern vectors from spherical clusters (8 mm radius) centered around each voxel within the individual participants' T-maps for *Perceptual prediction error versus baseline* (GLM-PC). These multidimensional vectors thus reflected how locally distributed patterns of fMRI activity represented perceptual prediction errors.

Based on these multidimensional vectors, we trained a support vector regression machine (SVR; linear kernel, constant regularization parameter of 1; implemented in LIBSVM, https://www.csie.ntu.edu.tw/~cjlin/libsvm) to predict the individual participants' post-pre difference in phase duration associated with virtual lesions in IFC. At each voxel, we performed 30 iterations of leave-one-out cross-validation, using the labeled data for 29 out of the 30 participants as the training-set and the remaining participant's data for testing. Prior to training, we normalized both the continuous labels and the multidimensional pattern vectors (i.e., $x_{norm} = (x - min(x))/(max(x) - min(x))$, with normalization parameters derived from the training set alone.[38]

In the test-set, we assessed predictive performance by calculating Pearson's correlation coefficients between the actual and the predicted difference in phase duration associated with virtual lesions in IFC. p values were computed at each voxel using nonparametric permutation testing. To create a null-distribution of correlation coefficients at each searchlight voxel, we repeatedly trained and tested the SVR with randomly permuted labels.

We considered prediction accuracy to be significant if permutation testing indicated that the probability of the true correlation occurred at $p_{FWE} < 0.05$. Prediction accuracy was thus assessed with Bonferroni-correction for multiple comparisons across all voxels in the whole volume of the brain. Therefore, the boundary p value surving FWE-correction was defined by $p < 0.05/n$, with n = 48 833 voxels inside the whole-brain volume. Permutation testing thus required up to 1/(0.05/n) = ~9770000 iterations at each voxel. We reduced the computational load by aborting permutation testing for a voxel where three values of the test statistic exceeded the true correlation coefficient.[38]

For visualization (Figure 7A), we computed *pseudo T-values* by drawing T-values corresponding to the nonparametric p values from an inverted student's T-distribution. We smoothed the resulting T-map with an 8 mm Gaussian kernel.

QUANTIFICATION AND STATISTICAL ANALYSIS

Conventional statistics

Summary statistics were carried out in *RMarkdown*. For linear mixed effects modeling, we used the R-packages *lme4* and *afex*. Bayes factors were computed using the R-package *BayesFactor*, using the function *ttestBF* for t tests (Cauchy prior; rscale = $\sqrt{2}/2$) and *lmBF* for linear mixed effects models (g-priors; fixed effects: rscale = $\sqrt{2}/2$; random effects = 1). To obtain Bayes factors for main effects and interactions, we estimated full and reduced models and divided the respective Bayes Factors.

Computational modeling

In this work, we investigated how neural activity in IFC related to the perceptual conflict inherent in ambiguous sensory information. Next to a *standard* assessment of perceptual conflict (see GLM-Congruency, E2), we applied an established *computational model* of bistable perception.[13,33,78] By inverting this model, we estimated perceptual prediction errors as a quantitative representation of perceptual conflict.

Here, we provide a mathematical description of the computational model of bistable perception. In addition, we describe how the model was inverted based on behavioral data. In simulation analyses, we illustrate the relation between model parameters (π_{IPS}: the initial belief in the stability of the visual environment; π_{ERROR}: the impact of perceptual conflict on the belief in the stability of the visual environment; π_{DIS}: the participants' sensitivity to disambiguating stimulus information) and the temporal characteristics of conscious experience *y*. With this, we derive quantitative predictions for the behavioral and imaging analyses outlined in the Results section.

Model description

Throughout the experiments E1 to E3, we presented a rotating discontinuous structure-from-motion stimulus. Participants reported whether they perceived the front surface of the object as rotating to the left or right. During full ambiguity, the direction of rotation spontaneously changed at a specific frequency (phase duration) in each participant. During graded ambiguity,[33] we experimentally manipulated the stimulus by introducing *disambiguating stimulus information* in form of 3D cues. Depending on the signal-to-ambiguity ratio, this disambiguating stimulus information biased conscious experience toward stimulus-congruent perceptual states.

Here, we explain how sensory data and implicit beliefs about the stability of the sensory environment give rise to perceptual states y during full and graded ambiguity. We adopt a Bayesian approach assuming that perceptual states are determined by *posterior probability distributions*. Posterior probability distributions result from the combination of currently available sensory data (the *likelihood distribution*) with information acquired from previous visual experience (the *prior distribution*).

During full ambiguity, our model assumes a bi-modal likelihood distribution representing balanced evidence for both perceptual interpretations. Graded ambiguity shifts the balance of the likelihood in the direction of one perceptual interpretation at the expense of the other. In this context, sensory information is described by the direction of disambiguation (μ_{DIS}) and the amount of disambiguation (i.e., the *signal-to-ambiguity* ratio; defined for the condition D1-D6 of experiment E2). As a free parameter, π_{DIS} reflects the individual impact of disambiguating stimulus information on conscious experience. This is equivalent to the participants' sensitivity to disambiguating stimulus information during graded ambiguity.

The prior, in turn, is modeled as a uni-modal distribution centered on the previously dominant perceptual interpretation. It acts as an implicit belief in the stability of the environment. The prior is defined by the current perceptual state $(\mu_{stability})$ and its impact on future conscious experience $(\pi_{stability})$. Two free parameters define the temporal evolution of $\pi_{stability}$: π_{IPS} reflects the maximum value of $\pi_{stability}$, which we allocate to the beginning of a perceptual phase. In addition, we assume that $\pi_{stability}$ decays linearly during a perceptual phase. This linear decay (with a lower bound at 0) occurs relative to the impact (or *precision*) of perceptual prediction errors (π_{ERROR}, see below).

The model combines the bimodal likelihood and the unimodal stability prior. This computes the available evidence for both interpretations of the sensory data. Crucially, once a percept is established, the residual evidence for the suppressed perceptual state constitutes a perceptual prediction error. Relative to the precision of the prediction error (π_{ERROR}), this quantitative representation of perceptual conflict leads to a linear reduction in the precision of the stability prior. Over time, this results in escalating prediction errors and a dynamic shift of the posterior distribution toward the currently suppressed perceptual interpretation. This entails an increasing probability of a change in conscious experience. Once the change has occurred, the stability prior shifts to the now-dominant stimulus interpretation and its precision is re-set to an initial value (π_{IPS}). As predicted by predictive-coding theories of perceptual inference,[10,23] prediction errors are thus minimized after the observer adopts a new perceptual interpretation.

In addition, our model assumes a modulation of prediction error accumulation by disambiguating stimulus information: When the current perceptual state is congruent with the disambiguating sensory evidence, our model predicts that prediction errors are reduced relative to full sensory ambiguity. Conversely, when perception is incongruent with the disambiguating sensory evidence, our model assumes enhanced perceptual prediction errors. Importantly, the strength of this enhancement/reduction in prediction errors scales with the amount of sensory evidence during graded ambiguity (i.e., the *signal-to-ambiguity* ratio) and the participants' sensitivity to disambiguating stimulus evidence (π_{DIS}).

Hence, three free parameters control the perceptual dynamics produced by our model: The intial precision of the stability prior π_{IPS}, the precision of perceptual prediction errors π_{ERROR}, which governs the rate of linear decay in the precision of the stability prior over time, and, in the case of graded ambiguity, the participants' sensitivity to disambiguating stimulus evidence π_{DIS}. We infer these parameters by inverting our model based on the sequence of percepts y indicated by the participants and, in the case of graded ambiguity, the available sensory information (μ_{DIS}: direction of disambiguation; *SAR*: signal-to-ambiguity ratio)

Since changes in conscious experience for non-depth-symmetrical structure-from-motion stimuli occur exclusively at overlapping stimulus configurations,[24,53] we represent percepts and all further model quantities in discrete time points t defined by stimulus overlaps. For computational expediency, our model assumes Gaussian probability distributions defined by mean and precision (inverse of variance).

At each time point t, we compute the probability of the two percepts based on the posterior distribution $P(\theta)$:

$$\theta = \begin{cases} >0.5: & \rightarrow & (rotation) \\ <0.5: & \leftarrow & (rotation) \end{cases} \quad \text{(Equation 1)}$$

The currently perceived direction at time point t is defined by:

$$y(t) = \begin{cases} 1: & \rightarrow & (rotation) \\ 0: & \leftarrow & (rotation) \end{cases} \quad \text{(Equation 2)}$$

We manipulate the level of sensory information by changing the fraction of dots associated with a stereo-disparity signal. This is captured by a Gaussian distribution *Disambiguation* $(\mathcal{N}(\mu_{DIS}, \pi_{DIS}^{-1}))$. The direction of disambiguation at time point t is represented by μ_{Dis}:

$$\mu_{Dis}(t) = \begin{cases} 1: & \rightarrow & (disambiguation) \\ 0.5: & \leftrightarrow & (ambiguous) \\ 0: & \leftarrow & (disambiguation) \end{cases} \quad \text{(Equation 3)}$$

π_{DIS} represents the participants' sensitivity to disambiguating stimulus information. The amount of disambiguating stimulus information was varied systematically in 120 s blocks of visual stimulation. The signal-to-ambiguity ratio (*SAR*) was defined by the fraction of stimulus dots that carried a 3D cue (level D1: 0.15, D2: 0.30, D3: 0.45, D4: 0.60, D5: 0.75 and D6; 1.00). If set to zero, π_{DIS} is removed from the model.

$$\pi_{Graded} = \pi_{DIS} * SAR \quad \text{(Equation 4)}$$

Furthermore, our model assumes that an implicit prior belief in the stability of the visual environment controls the frequency of changes in conscious experience during bistability. The mean of the Gaussian distribution "stability" $(\mathcal{N}(\mu_{stability}, \pi^{-1}_{stability}))$ is determined by the perceptual state indicated by the participants at the overlap preceding time point t:

$$\mu_{stability}(t) = y(t-1) \quad \text{(Equation 5)}$$

$\pi_{stability}$ describes the impact of the "stability" prior on perceptual state. If a change in conscious experience occurred at the preceding overlap $(t = t_0)$, $\pi_{stability}(t)$ is set to the initial stability precision π_{IPS}:

$$\pi_{stability}(t = t_0) = \pi_{IPS} \quad \text{(Equation 6)}$$

Inversion of our model during graded ambiguity allows for the estimation of π_{IPS}. If fixed to zero, the parameter is removed from the model.

If no perceptual event occurred at the preceding overlap $(t \neq t_0)$, we calculate $\pi_{stability}(t)$ by updating the previous precision of the stability prior $\pi_{stability}(t-1)$ with a precision-weighted prediction error (PE). The precision of the prediction error (π_{ERROR}) reflects how quickly $\pi_{stability}$ decays over time and is estimated as a free parameter:

$$\pi_{stability}(t \neq t_0) = \pi_{stability}(t-1) - \pi_{ERROR} * |PE(t-1)| \quad \text{(Equation 7)}$$

By combining the stability prior $(\mathcal{N}(\mu_{stability}, \pi^{-1}_{stability}))$ with the signal-to-ambiguity-adjusted likelihood $(\mathcal{N}(\mu_{DIS}, \pi^{-1}_{Graded}))$, we adjust the density ratio r of the posterior $P(\theta)$ for the two peak locations $\theta_0 = 0$ and $\theta_1 = 1$:

$$m(t) = \frac{\pi_{stability} * \mu_{stability}(t) + \pi_{graded} * \mu_{DIS}(t)}{\pi_{stability} + \pi_{graded}} \quad \text{(Equation 8)}$$

$$r(t) = exp\left(\frac{(\theta_1 - m(t))^2 - (\theta_0 - m(t))^2}{2 * (\pi_{stability} + \pi_{graded})^{-2}}\right) \quad \text{(Equation 9)}$$

The posterior probability of right-ward rotation predicts the perceptual response $y(t)$:

$$y_{predicted}(t) = \frac{1}{r(t)+1} \quad \text{(Equation 10)}$$

We infer on the free parameters (π_{DIS}, π_{ERROR}, π_{IPS}) by optimizing the model with regard to the difference between the prediction and the actual perceptual response ($y_{predicted}$ and y). Once a new percept $y(t)$ has been established, we compute the residual evidence for the alternative perceptual interpretation. This model quantity reflects a quantitative representation of dynamic changes in perceptual conflict. Given the inspiration of our model by predictive coding, we refer to this quantity as the *perceptual prediction error* $PE(t)$:

$$PE(t) = y(t) - y_{predicted}(t) \quad \text{(Equation 11)}$$

Model inversion

For model inversion, we used a free energy minimization approach,[79] which maximized a lower bound on the log-model evidence for the individual participants' data. We modeled π_{IPS}, π_{ERROR} and π_{DIS} either as free parameters defined by log-normal distributions or fixed these entities to zero, thereby effectively removing them from the model. We optimized parameters using quasi-Newton Broyden-Fletcher-Goldfarb-Shanno minimization as implemented in the HGF3.0 toolbox (TAPAS toolbox, https://www.translationalneuromodeling.org/hgf-toolbox-v3-0/).

For ambiguous visual stimulation, parameters were inverted using the following priors: π_{IPS} = prior mean of log(2) and prior variance of 1; π_{ERROR} = prior mean of log(1) and prior variance of 0.1. For graded ambiguity, prior means for π_{IPS} and π_{ERROR} were defined by the posterior estimates obtained from the preceding ambiguous runs. Prior variance was reduced to 0.01 for π_{IPS} and to 0.001 for π_{ERROR}. π_{DIS} was estimated with a prior mean of log(2) and a prior variance of 1.

We used the inverted models for model-based fMRI in experiment E1 (posterior parameter estimates: π_{IPS} = 2.83 ± 0.22; π_{ERROR} = 0.7 ± 0.08) and E2 (π_{IPS} = 2.25 ± 0.13; π_{ERROR} = 0.57 ± 0.09; π_{DIS} = 1.05 ± 0.15). Relative to model variants in which free parameters

were systematically removed, models incorporating the full set of parameters (Ambiguity: π_{IPS} and π_{ERROR}, Graded ambiguity: π_{IPS}, π_{ERROR} and π_{DIS}) were superior in explaining the participants' behavior (protected exceedance probability E1: 100%; E2: 99.98%). Furthermore, posterior model parameters were uncorrelated, indicating successful model inversion in E1 (π_{IPS} to π_{ERROR}: $\rho = -0.18$, $p = 0.32$) and E2 (π_{IPS} to π_{ERROR}: $\rho = -0.35$, $p = 0.13$; π_{IPS} to π_{DIS}: $\rho = 0.17$, $p = 0.47$; π_{ERROR} to π_{DIS}: $\rho = 0.3$, $p = 0.2$).

Based on previous work,[13] our model-based fMRI approach focused on perceptual prediction errors, since this model quantity provides a dynamic and quantitative representation of perceptual conflict.

Simulation

To visualize the predictions of our model, we simulated experiment E2 (one run of ambiguous stimulation; three runs of graded ambiguity across six levels of sensory evidence D1 to D6) for a total of 100 hypothetical participants. Parameters for simulation were drawn randomly between the 30% and 70% quantile of posterior parameters estimated in the behavioral pretest (π_{IPS} = 2.23 ± 0.14, π_{ERROR} = 0.36 ± 0.07; π_{DIS} = 1.73 ± 0.30).

As expected, the distribution of simulated phase durations (Figure S7A) obtained during ambiguous stimulation showed a sharp rise and long tail. It was best fit by a gamma distribution (Bayesian Information Criterion = 3.83×10^4; shape = 1.66, rate = 0.14) as compared to a lognormal (BIC = 3.85×10^4) and a normal distribution (BIC = 4.11×10^4).

When simulating graded ambiguity, we observed that disambiguating stimulus information biased the model predictions toward congruent perceptual states (Figure S7B). This congruency effect was stronger at higher levels of signal-to-ambiguity ($F(495) = 195.1$, $p = 1.49\times10^{-114}$, $BF_{10} = 4.45\times10^{120}$). Simulated prediction errors signaled elevated perceptual conflict during incongruent as opposed to congruent perceptual states (main effect of *Congruency*: $F(1.09\times10^3) = 4.15\times10^3$, $p = 0$, $BF_{10} = 6.8\times10^{275}$, Figure S7C). Differences in prediction errors between congruent and incongruent perceptual states scaled with the signal-to-ambiguity ratio (interaction between *Congruency* and *Signal-to-Ambiguity*: $F(1.09\times10^3) = 148.71$, $p = 2.1\times10^{-120}$, $BF_{10} = 2.13\times10^{116}$). We also observed a main effect of *Signal-to-Ambiguity* ($F(1.09\times10^3) = 81.43$, $p = 1.07\times10^{-72}$, $BF_{10} = 1.18\times10^{42}$).

Thus, when simulating from this computational model, we observed that the model's predictions closely followed the behavioral characteristics of both full and graded ambiguity.

2.5 Bistable perception alternates between internal and external modes of sensory processing

Weilnhammer VA, Chikermane M, Sterzer P. iScience 24, 102234 (2021). DOI: https://doi.org/10.1016/j.isci.2021.102234

The above publication provides causal evidence for an active role of prefrontal cortex in balancing internal predictions against externally-driven prediction errors[52]. In this study, we asked how the balance between predictions and prediction errors evolves over time. Healthy observers reported their conscious experience using button presses on consecutive trials at which we presented partially disambiguated Lissajous figures at varying levels of signal-to-ambiguity (Figure 2C). We used an adaptive staircase to titrate the proportion of stimulus-congruent conscious experiences to 75%. In recurring intervals lasting more than 20 trials, conscious experience was determined by internal predictions that over-rode otherwise effective sensory signals (*internal mode* processing). Such internal mode processing alternated with intervals during which conscious experience closely followed external sensory information (*external mode* processing). Computational modeling suggested that slow fluctuations in mode may play an important adaptive functions for perceptual inference, such as generating stable internal predictions despite ongoing sensory inputs, or calibrating beliefs about the reliability of external and internal sources of information.

The following text corresponds to the abstract of the article[37]:

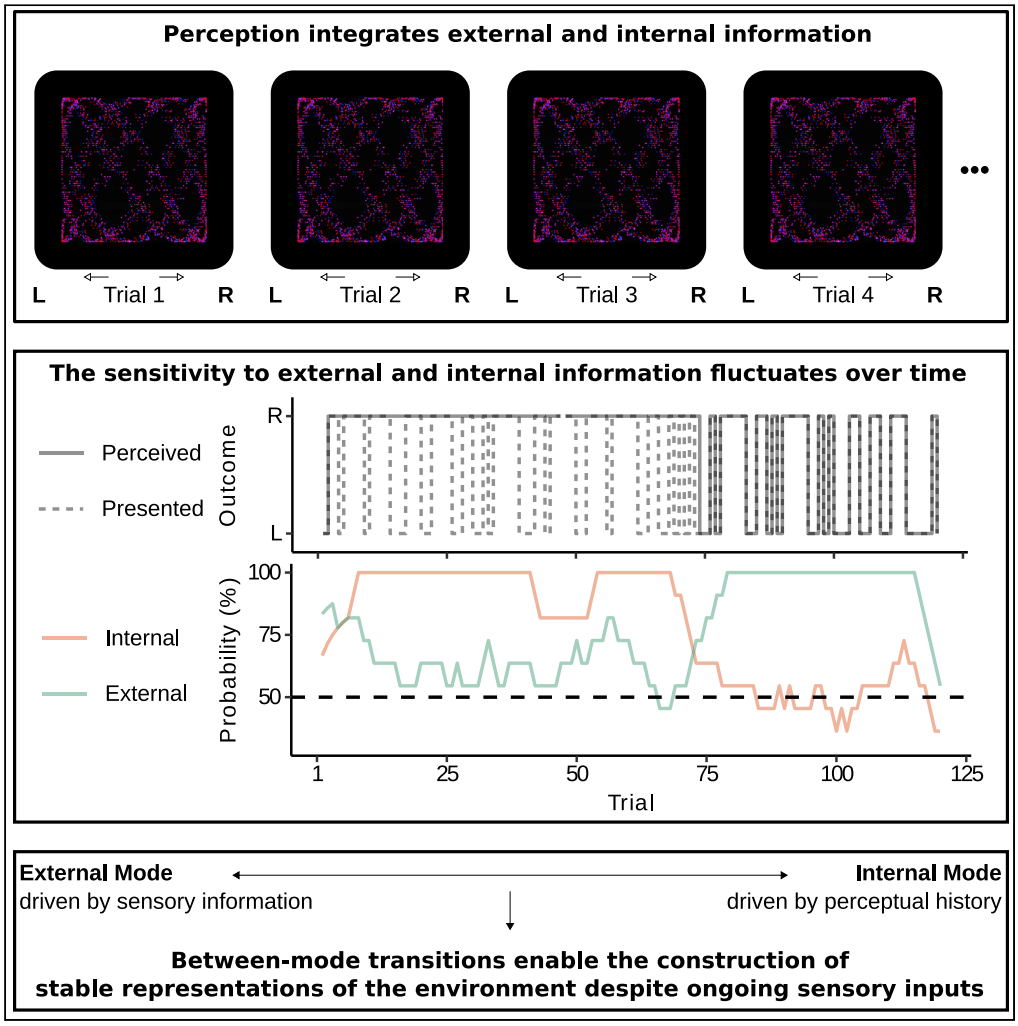
"Perceptual history can exert pronounced effects on the contents of conscious experience: when confronted with completely ambiguous stimuli, perception does not waver at random between diverging stimulus interpretations but sticks with recent percepts for prolonged intervals. Here, we investigated the relevance of perceptual history in situations more similar to everyday experience, where sensory stimuli are usually not completely ambiguous. Using partially ambiguous visual stimuli, we found that the balance between past and present is not stable over time but slowly fluctuates between two opposing modes. For time periods of up to several minutes, perception was either largely determined by perceptual history or driven predominantly by disambiguating sensory evidence. Computational modeling suggested that the construction of unambiguous conscious experiences is modulated by slow fluctuations between internally and externally oriented modes of sensory processing."

iScience

CellPress
OPEN ACCESS

Article

Bistable perception alternates between internal and external modes of sensory processing



Veith Weilnhammer, Meera Chikermane, Philipp Sterzer

veith-andreas.weilnhammer@charite.de,
https://osf.io/y2cfm/

HIGHLIGHTS

Perception fluctuates between external and internal modes of sensory processing

Weilnhammer et al., iScience *24*, 102234
March 19, 2021
https://doi.org/10.1016/j.isci.2021.102234

iScience

CellPress
OPEN ACCESS

Article

Bistable perception alternates between internal and external modes of sensory processing

Veith Weilnhammer,[1,2,5,*] Meera Chikermane,[1] and Philipp Sterzer[1,2,3,4]

SUMMARY

Perceptual history can exert pronounced effects on the contents of conscious experience: when confronted with completely ambiguous stimuli, perception does not waver at random between diverging stimulus interpretations but sticks with recent percepts for prolonged intervals. Here, we investigated the relevance of perceptual history in situations more similar to everyday experience, where sensory stimuli are usually not completely ambiguous. Using partially ambiguous visual stimuli, we found that the balance between past and present is not stable over time but slowly fluctuates between two opposing modes. For time periods of up to several minutes, perception was either largely determined by perceptual history or driven predominantly by disambiguating sensory evidence. Computational modeling suggested that the construction of unambiguous conscious experiences is modulated by slow fluctuations between internally and externally oriented modes of sensory processing.

INTRODUCTION

Imagine walking down a dark and unfamiliar street. As you struggle to identify potential obstacles, you are confronted with an ongoing stream of sensory signals, each compatible with multiple interpretations. In such situations, your previous perceptual experiences may provide valuable clues about how to interpret the ambiguous sensory data. Yet, relying too heavily on the past is risky, as you may end up overlooking unexpected changes in the environment.

Experimentally, the influence of preceding experiences on perception is usually investigated in tasks that require participants to perform perceptual decisions in a sequence of consecutive trials (Bergen and Jehee, 2019; Fründ et al., 2014). Such experiments reveal that, even in the absence of any correlation between the stimuli that are presented on successive trials, perception is significantly biased toward preceding choices (Abrahamyan et al., 2016; Fischer and Whitney, 2014; Fritsche et al., 2017; Hsu and Wu, 2020; Liberman et al., 2014; Urai et al., 2017, Urai et al., 2019). Importantly, *perceptual history* effects increase when sensory information becomes unreliable (Bergen and Jehee, 2019; Fründ et al., 2014). This reflects the idea that, when making perceptual decisions in situations of uncertainty, the brain may rely more strongly on internal predictions (Friston, 2005, 2010) that reflect the continuity of the sensory environment.

Integrating the *internal* information provided by perceptual history with the available *external* stimulus information may thus benefit perception by preventing erratic responses to unreliable sensory signals (Friston, 2005, 2010). However, the effects of perceptual history may also become mal-adaptive: when relying too strongly on preceding experiences, observers may become prone to ignore conflicting stimuli, which may lead to *hallucinatory* perceptual states that diverge from the true cause of the sensory data (Horga and Abi-Dargham, 2019; Powers et al., 2017).

In this work, we studied how visual perception balances external with internal sources of information in situations where perceptual history has a particularly strong effect. To this end, we investigated how preceding experiences impact the perception of ambiguous stimuli, i.e., stimuli that are compatible with two mutually exclusive perceptual states and typically give rise to bistable perception (Leopold et al., 2002). During bistable perception, observers experience spontaneous transitions between the two perceptual states, whereas the sensory data remain constant (Logothetis et al., 1996). Importantly, when the ambiguous stimuli are presented in successive trials separated by blank intervals, perception tends to stabilize

[1]Department of Psychiatry, Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charitéplatz 1, 10117 Berlin, Germany

[2]Berlin Institute of Health, Charité-Universitätsmedizin Berlin and Max Delbrück Center, 10178 Berlin, Germany

[3]Bernstein Center for Computational Neuroscience, Charité-Universitätsmedizin Berlin, 10117 Berlin, Germany

[4]Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, 10099 Berlin, Germany

[5]Lead contact

*Correspondence: veith-andreas.weilnhammer@charite.de, https://osf.io/y2cfm/

https://doi.org/10.1016/j.isci.2021.102234

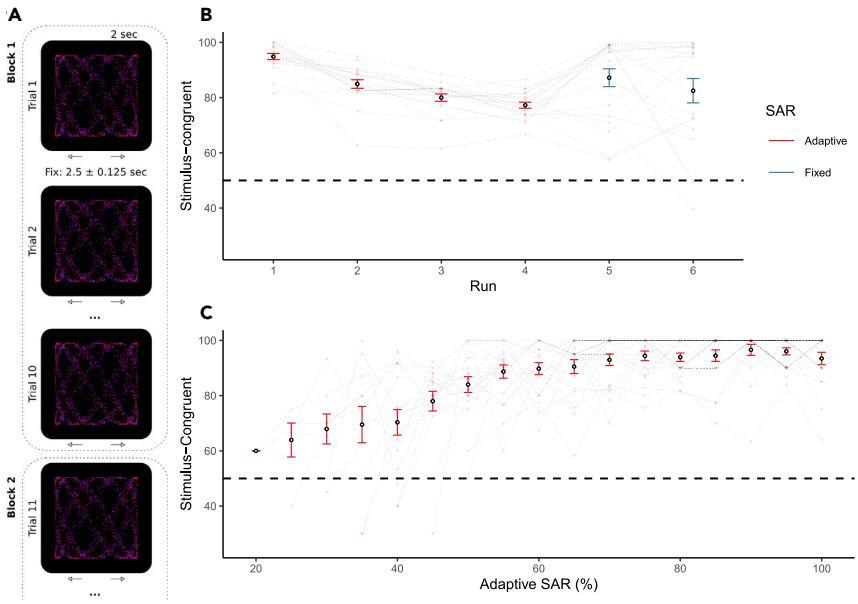


Figure 1. Psychophysical staircase

(A) **Graded ambiguity**. Participants viewed partially ambiguous structure-from-motion stimuli and indicated whether they perceived 3D rotation to the left or to the right. In runs R1-4, we dynamically adjusted the signal-to-ambiguity ratio (SAR) according to a staircase procedure that was based on the number of stimulus-congruent trials computed within blocks of 10 successive trials. During the final runs R5 and R6, we fixed the SAR to the average SAR obtained during runs R1-4.

(B) **Stimulus-congruent percepts across runs.** In runs R1-4 (depicted in red), the staircase procedure introduced dynamic adjustments in the SAR, reducing the frequency of stimulus-congruent percepts to approximately 75% (R1: 94.88 ± 1.1%; R2: 84.92 ± 1.55%; R3: 80 ± 1.33%; R4: 77.25 ± 1.09%). In runs R5-6 (depicted in blue), the SAR was fixed to the average SAR from the preceding runs R1-4 (60.25 ± 2.36%). Stimulus-congruent percepts amounted to 87.21 ± 3.23% in R5 and 82.5 ± 4.41% in R6.

(C) **Stimulus-congruent percepts across levels of SAR.** Stimulus-congruent percepts were more frequent at higher levels of disambiguating sensory information, ceiling at 100%. Pooled data are represented as mean ± SEM.

in one of the two interpretations (Maloney et al., 2005), indicating a pronounced effect of perceptual history (Pearson and Brascamp, 2008).

Here, we estimated the strength of perceptual history during bistable perception using a staircase procedure that dynamically adjusted the degree of perceptual ambiguity of structure-from-motion stimuli. By quantifying the effect of perceptual history relative to graded levels of sensory ambiguity, we investigated the computational mechanisms of integrating internal with external information during bistable perception.

RESULTS

To study how perceptual history is balanced against external sensory information during bistable perception, we asked 20 participants to indicate whether they perceived partially ambiguous random-dot-kinematograms as rotating to the left or the right (Figure 1A and Video S1). At each trial, we attached a 3D signal to a subset of the stimulus dots. This enabled us to parametrically manipulate the stimulus' *signal-to-ambiguity ratio* (Weilnhammer et al., 2020) (SAR). Ranging between 0% and 100%, these varying *levels* of disambiguating sensory information enforced one of the two stimulus interpretations (i.e., the *direction* of disambiguation). Within each experimental run, both directions of disambiguation occurred in equal number and in random sequence.

Perception integrates perceptual history with disambiguating sensory information

In the first four runs (R1-4, Figure 1B), we estimated individual threshold SARs necessary to induce balanced frequencies of stimulus-congruent and stimulus-incongruent percepts (i.e., trials perceived as congruent or incongruent with the disambiguating sensory information, respectively; Figure 2A). To this end, we dynamically adjusted the SAR based on the proportion of stimulus-congruent responses in consecutive 10-trial blocks. This psychophysical staircase decreased the SAR if less than 80% of trials were perceived as stimulus-congruent.

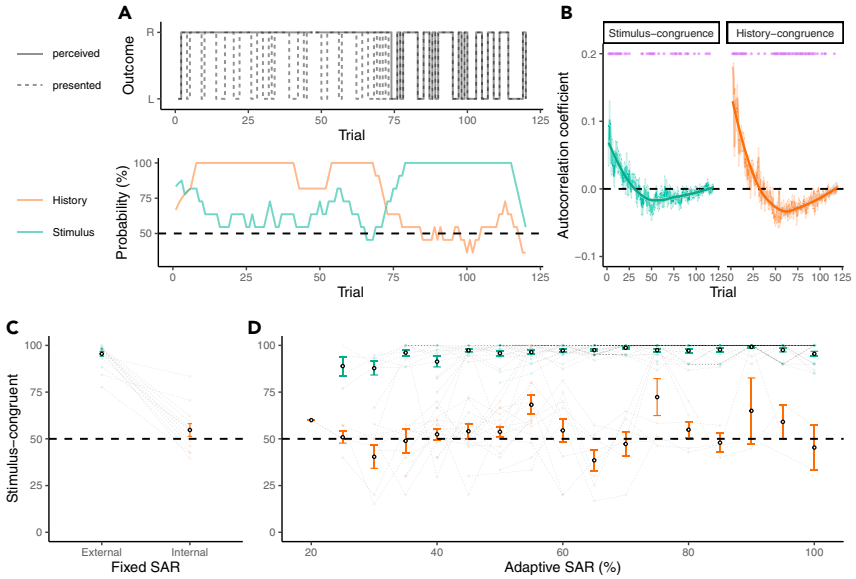


Figure 2. External and internal modes

(A) **Stimulus- and history-congruent perceptual states**. To visualize the influence of disambiguating sensory information and perceptual history, the upper panel depicts the time course of presented stimuli (L/R: disambiguating stimulus information for leftward/rightward rotation; dashed line) and the associated time course of perception (solid line). Perception is stimulus-congruent when the presented stimulus matches the associated perceptual state (i.e., overlap between the dashed and the solid line). History-congruent perception occurs when the perceptual state at a given trial matches the perceptual state at the preceding trial. The lower panel depicts the dynamic probabilities of stimulus-congruent percepts (green) and history-congruent percepts (orange) computed in sliding windows of ±5 trials for a representative participant. Perceptual processing switched between prolonged intervals of *internal mode* (green line below orange line), *external mode* (green line above orange line), and *intermediate mode* (overlap between green an orange line).

(B) **Average autocorrelation coefficients of stimulus- and history-congruence.** Despite constant SAR at threshold, both stimulus and history congruent were highly autocorrelated. If the index trial was perceived as congruent with visual stimulation (left panel) or perceptual history (right panel), the observer was more likely to experience stimulus- or history-congruent perceptual states, respectively, for approximately 25 trials. After that, the observer was more likely to experience incongruent states. The opposite relation holds for incongruent perceptual states at the index trial. Group-level averages were fitted using local polynomial regression fitting. Purple dots indicate trials at which the autocorrelation coefficients differed significantly from chance level ($p < 0.05$, two-sided one-sample t tests).

(C) **Stimulus-congruent percepts during internal and external mode for SAR at threshold.** During external mode, stimulus-congruent percepts made up for almost 100% of trials (95.49 ± 1.28%) but, interestingly, did not differ significantly from chance level during internal mode (54.71 ± 3.39%).

(D) **Stimulus-congruent percepts during internal and external mode across the full range of SAR.** Linear mixed effects modeling indicated that the frequency of stimulus-congruent percepts increased with levels of *SAR*. Internal mode was associated with a strong reduction of stimulus-congruent percepts (main effect of *mode*), which was more pronounced at low levels of SAR (*mode* × *SAR* interaction). Please note that any main effect of *mode* was expected, because external and internal mode were defined based on the dynamic probability of stimulus congruence. Pooled data are represented as mean ± SEM.

Conversely, we increased the SAR if the proportion of stimulus-congruent trials fell below 80%. As expected, stimulus-congruent percepts were less frequent at lower SARs ($F(1, 265.07) = 181.5$, $p = 7.25 \times 10^{-32}$, $BF_{10} = 5.22 \times 10^{28}$, main effect of *SAR*, Figure 1C). In runs R5-6, stimuli were presented at the individual threshold SAR (i.e., the average SAR from runs R1-4), which yielded stimulus-congruent percepts in 84.85 ± 3.12% of trials (Figure 1B).

Conversely, higher SARs reduced the impact of perceptual history (Figure S1). This resulted in a strong inverse relationship between stimulus- and history-congruent percepts (i.e., trials perceived in congruence with the immediately preceding percept), which were anti-correlated both within (average Pearson correlation coefficient $\rho = -0.9 \pm 0.02$, $T(19) = -49.25$, $p = 1.66 \times 10^{-21}$, $BF_{10} = 1.34 \times 10^{18}$, one-sample t test; Figure S2A) and across participants ($\rho = -0.77$, $p = 7.2 \times 10^{-5}$, $BF_{10} = 203.27$, Pearson correlation; Figure S2B).

We did not find any systematic bias toward one of the two perceptual interpretations (average probability of rightward rotation: 51.86 ± 3.04%; T(19) = 0.61, p = 0.55, BF_{10} = 0.27, one-sample t test). Absolute biases were small, amounting to 13.99 ± 1.57% across participants. Error responses were negligible, occurring in only 1.6 ± 1.57% of trials. Unclear percepts were not reported by the participants.

In logistic regression applied to each individual participant's behavioral data, trial-wise perceptual responses were best predicted based on both the current sensory information and the previous percept, as compared with reduced logistic regression models (Figure S2C) that used only stimulus information (T(19) = −9.39, p = 1.45×10^{-8}, BF_{10} = 8.89×10^{5}, paired t test) or only perceptual history (T(19) = −16.46, p = 1.06×10^{-12}, BF_{10} = 6.54×10^{9}) for prediction.

Two additional control analyses confirmed that both disambiguating sensory information and perceptual history significantly modulated the perception of partially disambiguated stimuli. Firstly, general linear mixed effects modeling with a binomial link function indicated a highly significant effect of both disambiguating sensory evidence (z = 45.55; p = 0) and perceptual history (z = 28.51; p = 8.62×10^{-179}), while controlling for the within-participant correlations using random intercepts.

A second possibility for this group-level inference is provided by general estimating equations (Hanley, 2003), which offer a non-parametric way of accounting for within-participant correlation by estimating population average effects. Likewise, this approach revealed a highly significant effect of disambiguating sensory evidence (Wald = 38.6; p = 5.2×10^{-10}) and perceptual history (Wald = 74.33; p = 0, correlation structure = "independence").

These results indicate that the effect of perceptual history is not limited to fully ambiguous stimuli (Pearson and Brascamp, 2008) but modulates perception through a weighted integration with varying levels of disambiguating sensory information (Bergen and Jehee, 2019). This finding aligns with the well-known observation that perception is co-determined by both sensory data and past experiences (Chopin and Mamassian, 2012; Fischer and Whitney, 2014; Fritsche et al., 2017; Hsu and Wu, 2020; Liberman et al., 2014). Perceptual history may benefit perception as an internal representation (Friston, 2005, 2010; Körding and Wolpert, 2004; Teufel and Fletcher, 2020) that stabilizes conscious experience when external sensory information is incomplete or unreliable. On your night-time walks, previous experiences may thus help you to avoid responding to irrelevant fluctuations in the ongoing stream of ambiguous sensory signals.

Perception fluctuates between temporally extended modes that are biased toward either external or internal information

In a next step, we examined how the probabilities of stimulus- and history-congruent percepts evolved within individual runs of the experiment (Figure 2A). Intriguingly, we found that both stimulus- and history-congruence were significantly autocorrelated (Figure 2B), indicating that the integration of perceptual history with sensory information was highly variable over time. For partially ambiguous stimuli presented at constant SARs (R5-6), we observed marked switches between intervals in which perception was either strongly driven by disambiguating sensory information (*external mode*, 73.25 ± 6.17% of trials) or determined by perceptual history (*internal mode*; 23.94 ± 5.84%), in addition to shorter intermediate intervals (2.81 ± 0.77%; Figure 2A, lower panel). Switches between these modes occurred on average every 39.9 ± 7.31 trials (179.53 ± 32.91 s).

Our analyses therefore revealed prolonged intervals of alternating biases toward either internal or external information. This finding is incompatible with the view that perception is best explained by integrating uncertain sensory data with only the immediately preceding perceptual state. As indicated by simulation analyses (Figure S3), such a Markovian assumption did not reproduce the autocorrelation of stimulus and history congruence (Figure S3B) and predicted longer external (T(19) = 2.75, p = 0.01, BF_{10} = 4.17, paired t test; Figure S3C) as well as shorter internal modes (T(19) = −3.49, p = 2.44×10^{-3}, BF_{10} = 16.92).

In sum, these results imply that a stable moment-by-moment integration of current sensory information with the immediately preceding percept is not sufficient to explain the perceptual dynamics during graded ambiguity. Rather, our findings suggest that participants transition between temporally extended perceptual *modes* (Honey et al., 2017) that are biased toward either external information (i.e., disambiguating sensory information) or internal information (i.e., perceptual history).

Importantly, switches between internal and external modes could not be attributed to small fluctuations in the participants' sensitivity to disambiguating sensory information. At threshold (R5-6), stimulus-congruent percepts were close to 100% during external mode but ranged at chance level during internal mode ($T(12) = 1.39$, $p = 0.19$, $BF_{10} = 0.61$, one-sample t test, Figure 2C). Please note that the overall difference in stimulus-congruency between modes is expected, because external and internal mode were defined based on the dynamic probability of stimulus-congruent perceptual states.

Moreover, internal mode suppressed the sensitivity to disambiguating sensory information not only at the threshold but across the full range of SAR ($F(2, 484.41) = 35.26$, $p = 5.04 \times 10^{-15}$, $BF_{10} = 4.78 \times 10^{66}$; main effect of *mode*; Figure 2D). During runs in which the SAR was adjusted dynamically (R1-R4), transitions from internal to mode were more likely to occur when the available sensory information was reduced ($F(2, 472.71) = 5.25$, $p = 5.58 \times 10^{-3}$, $BF_{10} = 3.57$, *mode* × *SAR* interaction). In sum, these control analyses argue against the view that between-mode transition may result exclusively from a threshold phenomenon.

As a second caveat, we asked whether the observed transitions between internal and external mode constitute a perceptual phenomenon or, alternatively, occur only due to cognitive processes that are situated downstream of perception (Brascamp et al., 2018). In this context, it may be argued that the participants' attention to the experimental task may have fluctuated over time (Rosenberg et al., 2013; Zalta et al., 2020), leading to intervals of stereotypic reporting behavior. We addressed this potential confound by analyzing response times (*RTs*, see Figure S4), which have been shown to link closely with on-task attention (Prado et al., 2011; Rosenberg et al., 2013).

In contrast to stimulus and history congruence, response times remained stable across the experimental runs (Figure S4A) and did not vary across levels of SAR (R1-R4; $F(1, 261.5) = 0.05$, $p = 0.82$, $BF_{10} = 0.15$; Figure S4B). At threshold, RTs did not differ between external and internal mode ($T(12) = 0.74$, $p = 0.48$, $BF_{10} = 0.35$, paired t test; Figure S4C). Moreover, when analyzing RTs according to the factors *mode* and *SAR* in runs R1-R4 (Figure S4D), we found that, during internal mode, RTs increased for escalating levels of SAR. Speculatively, this *mode* × *SAR* interaction ($F(2, 476.5) = 10.73$, $p = 2.77 \times 10^{-5}$, $BF_{10} = 538.42$) could reflect the increase in conflict between the history-congruent state and the available sensory information (Weilnhammer et al., 2020).

At the same time, both the absence of any mode effect on RTs at threshold as well as the sensitivity of internal-mode RTs to levels of *SAR* argue against the notion that internal mode is caused by the participants paying less attention to the experimental task. In line with this observation, we found no changes in the distribution of normalized RTs, as participants transitioned between internal and external mode (see Figure 4E for group RTs collapsed across participants and Figure 4F for individual distributions).

As a final control analysis, we checked whether internal mode was associated with an enhanced impact of the perceptual state *experienced* at the preceding trial (i.e., perceptual history), as opposed to the disambiguating sensory information *presented* at the preceding trial (i.e., stimulus history). As expected, history-congruent perceptual states dominated periods of internal mode processing ($94.15 \pm 1.03\%$), whereas stimulus history had no detectable influence on perception in these intervals ($49.79 \pm 1.03\%$; $T(19) = -0.2$, $p = 0.84$, $BF_{10} = 0.24$, one-sample t test).

During internal mode, perceptual history thus strongly determines conscious experience, overriding otherwise effective sensory information. As you interpret ambiguous sensory information on your walk through the dark, relying on an internal representation of your surroundings may dramatically increase the energy efficiency of perception. However, this is only adaptive in stable environments, i.e., when sensory events are highly auto-correlated. In volatile environments, internally biased sensory processing may cause perception to get stuck in the past, resulting in hallucinatory experiences that ignore relevant conflicts (Weilnhammer et al., 2020) with sensory information (Horga and Abi-Dargham, 2019).

Computational modeling indicates that between-mode transitions are best explained by a fluctuating impact of accumulating perceptual history

How can perception achieve an adaptive balance between external and internal mode? To address this question, we investigated the potential computational mechanisms that could lead to the observed oscillations between internally and externally biased modes of perceptual processing. To this end, we constructed a set of four generative behavioral models (Wilson and Collins, 2019) (Figure 3) that differed across two dimensions.

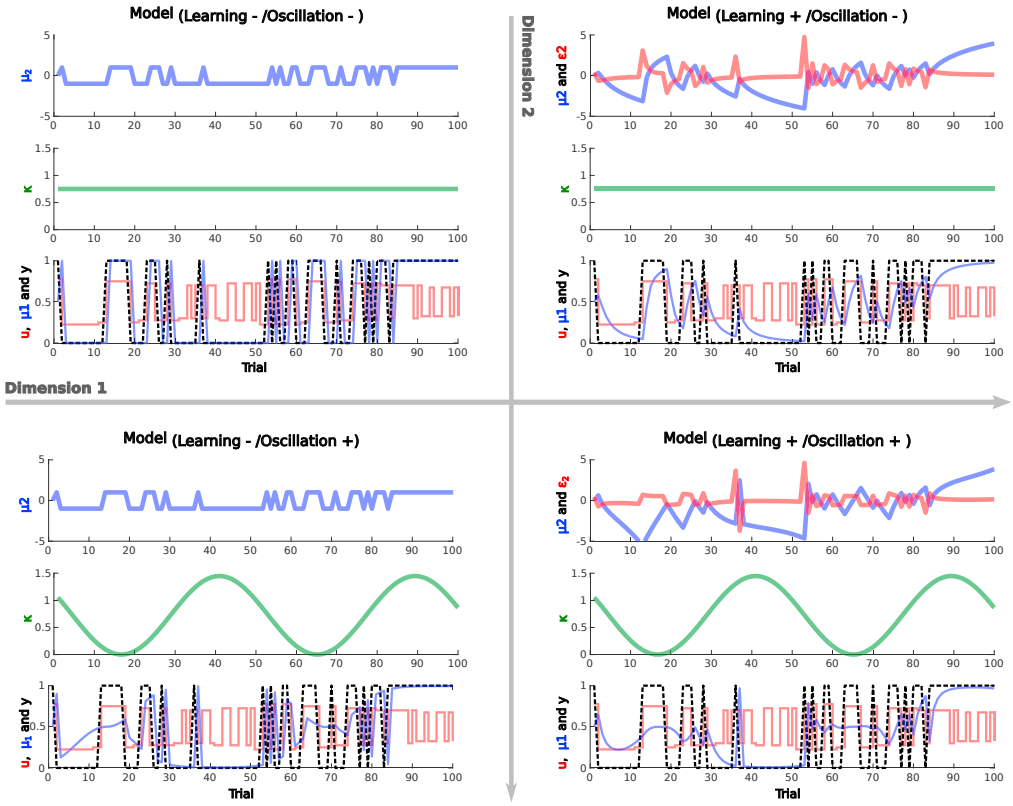


Figure 3. Computational modeling: Modelspace

To investigate the computational mechanisms of between-mode transitions, we constructed a space of four behavioral models that differed along two dimensions. Each model's quantities are shown in three separate panels. Along a **first dimension** (horizontal arrow), we manipulated whether perceptual history effects were represented exclusively by the perceptual state at the preceding trial (left side) or, alternatively, dynamically accumulated according to a learning rate ω (right). Perceptual history and its updating are displayed in the upper panel of each model. The blue line represents μ_2, i.e., the tendency to expect rightward (above zero) or leftward (below zero) rotation at the upcoming trial. The red line depicts dynamic precision-weighted prediction errors ε_2 that update μ_2 in response to the sequence of perceptual experiences. Along a **second dimension** (vertical arrow), we contrasted models that assumed a stable influence of perceptual history on perception (top) against models that assumed a systematic fluctuation in the impact of perceptual history. κ (green line; middle panel) represents the weight at which perceptual history impacts on perception. The lower panel shows the perceptual prediction $\widehat{\mu_2}$ (blue line, provided by a sigmoid transform of $\mu * \kappa$), the disambiguating sensory information *u* (red) and the participants' response *y* (black).

On the first dimension, we asked whether biases toward internal mode arise from the sequence of previous experiences. We reasoned that, if perceptual history effects dynamically accumulate over time (Brascamp et al., 2008; Pearson and Brascamp, 2008), perception would be more strongly biased toward a perceptual state if the current trial was preceded by a long sequence of history-congruent trials. Accumulating perceptual history effects could eventually become strong enough to override otherwise effective sensory information, thereby creating intervals during which perception is strongly determined by internal information.

To this end, we adopted a Bayesian modeling approach that frames perception as an inferential process in which perceptual decisions are determined by posterior distributions (Friston, 2010). Following Bayes' rule, such posterior distributions are computed by integrating a likelihood distribution representing the sensory evidence (i.e., disambiguating sensory information for left- or rightward rotation at a given SAR) with the prior probability of perceptual states (i.e., perceptual history).

The null model $M_{Learning-/Oscillation-}$ (see transparent method section and Figure 3 for details) assumes that the effect of perceptual history (i.e., the estimated prior probability of perceptual states) depends only on

the perceptual state at the immediately preceding trial. Its weight on perception is determined by the parameter κ. The impact of sensory information, in turn, depends on the sensitivity parameter α.

By contrast, in the alternative model $M_{Learning+/Oscillation-}$, the estimated prior probability of perceptual states depends not only on the response at the preceding trial but dynamically accumulates over time according to a two-level Hierarchical Gaussian Filter (Mathys et al., 2014). Thus, the implicit *belief* in the probability of perceiving leftward rotation increases as a function of the number of preceding trials that have been experienced as rotating toward the left (and vice versa). The second-level accumulation of perceptual history is governed by the learning-rate parameter ω.

In this model, switches between modes can only be driven by experience. Once perceptual history effects have accumulated and caused the estimated probability of leftward rotation to increase significantly above chance level, switches to external mode are enabled by prediction errors that are caused by the experience of rightward rotation (and vice versa).

As an alternative explanation, we reasoned that switches between modes could additionally be facilitated by systematic fluctuations in κ, the parameter governing the impact of perceptual history on perception. When κ is low, perceptual states are more likely to be history incongruent, increasing the likelihood of prediction errors that enable the transition from internal to external mode. To test whether such fluctuations provide a plausible explanation of our behavioral data, we introduced a second dimension to our model space by constructing $M_{Learning+/Oscillation+}$ and $M_{Learning-/Oscillation+}$. Instead of estimating κ as a stable parameter, these models enable oscillations in κ that are governed by parameters for amplitude *amp*, frequency *f* (in *nb trials*$^{-1}$), and phase *p*.

We inverted all models based the trial-wise perceptual responses given by our participants and used random-effects Bayesian model family selection (Stephan et al., 2009) to determine whether the dynamic accumulation of perceptual history (dimension 1) and systematic fluctuations in its impact (dimension 2) were likely to represent a computational mechanism of mode switches.

On the first dimension, we found that models assuming a dynamic accumulation of perceptual history (*Learning+*) outperformed *Learning−* models at a protected exceedance probability of 100%. On the second dimension, Bayesian model selection indicated that our data were better explained by models that assumed a fluctuating impact of perceptual history (*Oscillation+*) as compared with *Oscillation−* models at a protected exceedance probability of 99.98%. $M_{Learning+/Oscillation+}$ was therefore identified as the clear winning model (protected exceedance probability = 99.82%; see Figure 4A for model-level inference at the participant level and Figure 4B for posterior parameter estimates).

With this, our computational approach suggests that switches between internal and external mode are governed by two interlinked processes: In line with previous findings (Brascamp et al., 2008; Pearson and Brascamp, 2008), we found that perceptual history accumulates over time. Eventually, accumulating perceptual history may override disambiguating sensory information, causing a transition to from external to internal mode. In isolation, however, such a process falls short of explaining transitions in the opposite direction. Because perceptual history effects continue to accumulate during internal mode, they should eventually become impossible to overcome (Wexler et al., 2015). Crucially, our modeling results propose that fluctuations in the impact of perceptual history enable transition from internal to external mode by temporarily de-coupling the perceptual decision from implicit internal representations of the environment.

DISCUSSION

In this work, we show that perceptual history modulates perception through a weighted integration (Bergen and Jehee, 2019) with varying levels of sensory information. Perceptual history therefore acts as an internal representation (Friston, 2005, 2010; Teufel and Fletcher, 2020) that stabilizes perception when sensory signals are ambiguous. Intriguingly, we found that the balance between perceptual history and disambiguating sensory information slowly alternates between internally and externally oriented modes of sensory processing. Computational modeling indicated that between-mode transitions were likely to be caused by fluctuations in how strongly perception was driven by the accumulating effects of perceptual history.

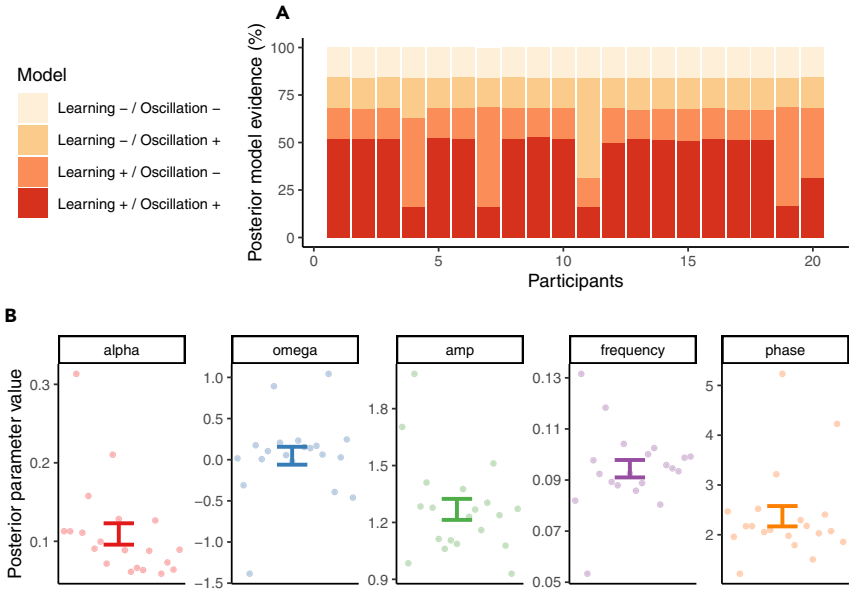


Figure 4. Computational modeling: Results

(A) **Model-level inference.** Random-effects Bayesian model selection identified $M_{Learning+/Oscillation+}$ as the clear winning model (group-level protected exceedance probability = 99.82%).

(B) **Parameter-level inference.** This model assumes that the external sensory signals is detected with a sensitivity parameter of $\alpha = 0.11 \pm 0.01$. The internal representation derived from perceptual history is updated as a function of the sequence of percepts according to learning rate $\omega = 0.05 \pm 0.11$. κ, the impact of accumulating perceptual history on perception, fluctuated according to a sine function with an amplitude of 1.27 ± 0.06, a frequency of $0.09 \pm 3.41 \times 10^{-3}$ (in *nb trials*$^{-1}$), and a phase of 2.37 ± 0.2. Pooled data are represented as mean $\pm$ SEM.

It may be argued that temporally extended biases toward internal information are not generic but specific to the class of structure-from-motion stimuli (Longuet-Higgins, 1986) investigated here. Indeed, structure-from-motion induces relatively long perceptual dominance durations (Weilnhammer et al, 2014, 2016, 2020). In addition, individual observers have been shown to exhibit stable idiosyncratic biases toward one of the two stimulus interpretations (Mamassian and Wallace, 2010; Weilnhammer et al., 2020), which can become strong enough to override disambiguating 3D cues (Wexler et al., 2015).

In this work, however, two factors speak against the view that transitions to internal mode were caused exclusively by strong perceptual biases. Firstly, we found relatively weak imbalances between the two possible states induced by our partially ambiguous structure-from-motion stimulus (see Results section). Secondly, we observed frequent transitions from internal to external mode while sensory information was held constant at threshold (see Figure 2), arguing against stable biases as the primary determinant of internally biased processing during graded ambiguity. Yet, to empirically assess this caveat, future work should investigate whether between-mode transitions occur also for ambiguous stimuli that induce shorter dominance durations, such as the Necker cube (Kornmeier and Bach, 2005). This would help understand whether fluctuations between internal and external mode depend on the type, strength, and temporal characteristics of bistable perception or, alternatively, occur independently of these factors and thus constitute a more general feature of perceptual processing.

As a second alternative explanation of our results, it may be proposed that fluctuating biases toward internal or external mode do not represent a perceptual phenomenon but, conversely, occur only due to processes that are situated downstream of perception, such as changes in reporting behavior (Brascamp et al., 2018) that are caused by periodic changes in how well participants attended to the experimental task (Zalta et al., 2020). Our analysis of response times (Figure S4), which are classically linked to fluctuating attention in paradigms such as the *Continuous Performance Task* (Rosenberg et al., 2013), did not yield any evidence for systematic differences in response behavior between internal and external mode. Yet, future experiments should apply no-report paradigms (Frässle et al., 2014), pupillometry (Lawson et al., 2020), or experimental manipulations of on-task attention (Alais et al., 2010) to dissociate post-perceptual processes from the perceptual phenomenon of mode-switching proposed in this work.

In a similar vein, it may be argued that slow fluctuations between externally and internally biased perception reflect epiphenomena that may arise from arbitrary constraints of neural processing (Honey et al., 2017). On the other hand, there may also be a specific computational benefit to slow transitions between external and internal model (Honey et al., 2017; Palva and Palva, 2011; VanRullen, 2016): In stable environments, internal mode may come with the benefit of a dramatic reduction in the energy demands of perception (Friston, 2010). Periodic switches to external mode may ensure that internal representations are updated in response to potential changes in the environment (Honey et al., 2017). In contrast to simultaneous processing, periodic mode switches may allow the brain to differentiate between internal and external sources of information (Honey et al., 2017). This may help perception to solve the credit-assignment problem, i.e., deciding whether to update internal representations of the environment or, alternatively, to modify beliefs about the reliability of sensory information (Weilnhammer et al., 2018). Thus, mode switching may represent a process that helps constructing stable representations of the environment despite ongoing sensory inputs (Bengio et al., 2015).

Indeed, fluctuations between externally and internally biased processing have been described in a variety of cognitive domains, including perception (Monto et al., 2008), episodic memory (Duncan et al., 2012), and waking state (McGinley et al., 2015). Switching between external and internal processing modes may thus represent a general computational mechanism that helps to adaptively integrate prior predictions with new information (Honey et al., 2017). Alterations in the temporal dynamics of mode switching may therefore represent the neurocomputational basis of psychotic experiences that often co-occur across cognitive domains, such as hallucinations, delusions, and altered sense of agency (Horga and Abi-Dargham, 2019; Sterzer et al., 2018).

To test the hypothesis that mode switches represent an adaptive mechanism that occurs across cognitive domains, future research should investigate whether transitions between external and internal mode can be induced experimentally. Based on the results of our computational modeling analysis, it may be hypothesized that participants should be more prone to transition from internal to external mode when repeatedly confronted with information that contradicts past experiences. Conversely, transitions from external to internal mode should occur more swiftly when participants receive information that is in line with prior predictions. Further down the line, it may be speculated that the overall frequency of mode switches could be altered by experimentally manipulating the volatility of the input data (Iglesias et al., 2013; Mathys et al., 2014).

Likewise, the existence of mode switches should be further substantiated by investigating whether external and internal modes can be determined based on markers that are independent of the perceptual response such as pupillary response or heart rate (Lawson et al., 2020). Together with an experimental manipulation of between-mode transitions, such markers could help to understand whether transitions between internal and external modes indeed represent an adaptive cognitive strategy that aids learning, or, alternatively, result from independent phenomena such as adaption (Chopin and Mamassian, 2012), attention (Alais et al., 2010), or response behavior (Frässle et al., 2014).

Limitations of the study

In this study, we have shown that bistable perception cycles through prolonged periods of enhanced and reduced sensitivity to disambiguating stimulus information. This finding suggests that conscious experience is characterized by slow fluctuations between internally and externally oriented modes of sensory processing. As a first limitation, our work investigated between-mode transitions only for a specific class of bistable stimuli (ambiguous structure-from-motion). Future work should test whether alternations between internally and externally oriented modes of processing also occur in other bistable stimuli, in particular in relation to paradigms that induce shorter dominance durations. As a second limitation, our work defines internal and external modes solely on the basis of behavior. Future studies should apply independent markers for internal and external modes (such as pupillometry) to probe between-mode transitions irrespective of behavioral reports. As a third limitation, our study does not provide an experimental control of pre- and post-perceptual processes such as attention or response behavior. Future experiments should use no-report paradigms or experimental manipulations of on-task attention to confirm that mode-switching represents a perceptual phenomenon rather than a process that occurs up- or downstream of perception.

Resource availability

Lead contact

The lead contact for this study is Veith Weilnhammer (veith-andreas.weilnhammer@charite.de).

Material availability

Materials have been deposited at OSF: https://doi.org/10.17605/OSF.IO/Y2CFM.

Data and code availability

Original data and code have been deposited at OSF: https://doi.org/10.17605/OSF.IO/Y2CFM.

METHODS

All methods can be found in the accompanying transparent methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102234.

ACKNOWLEDGMENTS

Author VW is a fellow of the Clinician Scientist Program funded by the Charité – Universitätsmedizin Berlin and the Berlin Institute of Health. This program was initiated and led by Prof. Dr. Duska Dragun to enable physicians to pursue a parallel career in academic research. With great sadness we have received the news that Prof. Dragun passed away on December 28th of 2020. We dedicate this publication to her as a mentor, friend, role model, and stellar scientist.

PS is funded by the German Research Foundation (STE 1430/8-1) and the German Ministry for Research and Education (ERA-NET NEURON program 01EW2007A). We acknowledge support from the German Research Foundation (DFG) and the Open Access Publication Fund of Charité – Universitätsmedizin Berlin.

AUTHOR CONTRIBUTIONS

- VW and PS conceptualized the study.
- VW designed the experiment.
- VW and MC collected the data.
- VW and PS wrote the initial draft and edited the manuscript.
- VW, MC, and PS reviewed the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 28, 2020
Revised: January 20, 2021
Accepted: February 21, 2021
Published: March 19, 2021

SUPPORTING CITATIONS

The following references appear in the supplemental information: Gekas et al., 2019.

REFERENCES

Abrahamyan, A., Silva, L.L., Dakin, S.C., Carandini, M., and Gardner, J.L. (2016). Adaptable history biases in human perceptual decisions. Proc. Natl. Acad. Sci. U S A *113*, E3548–E3557.

Alais, D., Boxtel, J.J.van, Parker, A., and Ee, R.van (2010). Attending to auditory signals slows visual alternations in binocular rivalry. Vis. Res. *50*, 929–935.

Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., and Lin, Z. (2015). Towards biologically plausible deep learning. arRxiv, 1502.04156.

Bergen, R.S.van, and Jehee, J.F. (2019). Probabilistic representation in human visual cortex reflects uncertainty in serial decisions. J. Neurosci. *39*, 8164–8176.

Brascamp, J., Sterzer, P., Blake, R., and Knapen, T. (2018). Multistable perception and the role of the frontoparietal cortex in perceptual inference. Annu. Rev. Psychol. *69*, 77–103.

Brascamp, J.W., Knapen, T.H.J., Kanai, R., Noest, A.J., van Ee, R., and van den Berg, A.V. (2008). Multi-timescale perceptual history resolves visual ambiguity. PLoS One *3*, e1497.

Chopin, A., and Mamassian, P. (2012). Predictive properties of visual adaptation. Curr. Biol. *22*, 622–626.

Duncan, K., Sadanand, A., and Davachi, L. (2012). Memory's Penumbra: episodic memory decisions induce lingering mnemonic biases. Science *337*, 485–487.

Fischer, J., and Whitney, D. (2014). Serial dependence in visual perception. Nat. Neurosci. *17*, 738–743.

Frässle, S., Sommer, J., Jansen, A., Naber, M., and Einhäuser, W. (2014). Binocular rivalry:

frontal activity relates to introspection and action but not to perception. J. Neurosci. *34*, 1738–1747.

Friston, K. (2010). The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. *11*, 127–138.

Friston, K. (2005). A theory of cortical responses. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. *360*, 815–836.

Fritsche, M., Mostert, P., and Lange, F.P.de (2017). Opposite effects of recent history on perception and decision. Curr. Biol. *27*, 590–595.

Fründ, I., Wichmann, F.A., and Macke, J.H. (2014). Quantifying the effect of intertrial dependence on perceptual decisions. J. Vis. *14*, 1–16.

Gekas, N., McDermott, K.C., and Mamassian, P. (2019). Disambiguating serial effects of multiple timescales. J. Vis. *19*, 1–14.

Hanley, J.A. (2003). Statistical analysis of correlated data using generalized estimating equations: an orientation. Am. J. Epidemiol. *157*, 364–375.

Honey, C.J., Newman, E.L., and Schapiro, A.C. (2017). Switching between internal and external modes: a multiscale learning principle. Netw. Neurosci. *1*, 339–356.

Horga, G., and Abi-Dargham, A. (2019). An integrative framework for perceptual disturbances in psychosis. Nat. Rev. Neurosci. *20*, 763–778.

Hsu, S.M., and Wu, Z.R. (2020). The roles of preceding stimuli and preceding responses on assimilative and contrastive sequential effects during facial expression perception. Cogn. Emot. *34*, 890–905.

Iglesias, S., Mathys, C., Brodersen, K.H., Kasper, L., Piccirelli, M., Ouden, H.E.den, and Stephan, K.E. (2013). Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. Neuron *80*, 519–530.

Kornmeier, J., and Bach, M. (2005). The Necker cube–an ambiguous figure disambiguated in early visual processing. Vis. Res. *45*, 955–960.

Körding, K.P., and Wolpert, D.M. (2004). Bayesian integration in sensorimotor learning. Nature *427*, 244–247.

Lawson, R.P., Bisby, J., Nord, C.L., Burgess, N., and Rees, G. (2020). The computational, pharmacological, and physiological determinants of sensory learning under uncertainty. Curr. Biol. *31*, 163–172.e4.

Leopold, D.A., Wilke, M., Maier, A., and Logothetis, N.K. (2002). Stable perception of visually ambiguous patterns. Nat. Neurosci. *5*, 605–609.

Liberman, A., Fischer, J., and Whitney, D. (2014). Serial dependence in the perception of faces. Curr. Biol. *24*, 2569–2574.

Logothetis, N.K., Leopold, D.A., and Sheinberg, D.L. (1996). What is rivalling during binocular rivalry? Nature *380*, 621–624.

Longuet-Higgins, H.C. (1986). Visual motion ambiguity. Vis. Res. *26*, 181–183.

Maloney, L.T., Dal Martello, M.F., Sahm, C., and Spillmann, L. (2005). Past trials influence perception of ambiguous motion quartets through pattern completion. Proc. Natl. Acad. Sci. U S A *102*, 3164–3169.

Mamassian, P., and Wallace, J.M. (2010). Sustained directional biases in motion transparency. J. Vis. *10*, 23.

Mathys, C.D., Lomakina, E.I., Daunizeau, J., Iglesias, S., Brodersen, K.H., Friston, K.J., and Stephan, K.E. (2014). Uncertainty in perception and the hierarchical Gaussian filter. Front. Hum. Neurosci. *8*, 825.

McGinley, M.J., Vinck, M., Reimer, J., Batista-Brito, R., Zagha, E., Cadwell, C.R., Tolias, A.S., Cardin, J.A., and McCormick, D.A. (2015). Waking state: rapid variations modulate neural and behavioral responses. Neuron *87*, 1143–1161.

Monto, S., Palva, S., Voipio, J., and Palva, J.M. (2008). Very slow EEG fluctuations predict the dynamics of stimulus detection and oscillation amplitudes in humans. J. Neurosci. *28*, 8268–8272.

Palva, J.M., and Palva, S. (2011). Roles of multiscale brain activity fluctuations in shaping the variability and dynamics of psychophysical performance. In Progress in Brain Research (Elsevier B.V.), pp. 335–350.

Pearson, J., and Brascamp, J. (2008). Sensory memory for ambiguous vision. Trends Cogn. Sci. (Regul. Ed.) *12*, 334–341.

Powers, A.R., Mathys, C., and Corlett, P.R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. Science *357*, 596–600.

Prado, J., Carp, J., and Weissman, D.H. (2011). Variations of response time in a selective attention task are linked to variations of functional connectivity in the attentional network. NeuroImage *54*, 541–549.

Rosenberg, M., Noonan, S., DeGutis, J., and Esterman, M. (2013). Sustaining visual attention in the face of distraction: a novel gradual-onset continuous performance task. Atten. Percept. Psycho. *75*, 426–439.

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., and Friston, K.J. (2009). Bayesian model selection for group studies. NeuroImage *46*, 1004–1017.

Sterzer, P., Adams, R.A., Fletcher, P., Frith, C., Lawrie, S.M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., and Corlett, P.R. (2018). The predictive coding account of psychosis. Biol. Psychiatr. *84*, 634–643.

Teufel, C., and Fletcher, P.C. (2020). Forms of prediction in the nervous system. Nat. Rev. Neurosci. *21*, 231–242.

Urai, A.E., Braun, A., and Donner, T.H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. Nat. Commun. *8*, 14637.

Urai, A.E., De Gee, J.W., Tsetsos, K., and Donner, T.H. (2019). Choice history biases subsequent evidence accumulation. eLife *8*, e46331.

VanRullen, R. (2016). Perceptual cycles. Trends Cogn. Sci. *20*, 723–735.

Weilnhammer, V., Röd, L., Eckert, A.-L., Stuke, H., Heinz, A., and Sterzer, P. (2020). Psychotic experiences in schizophrenia and sensitivity to sensory evidence. Schizophrenia Bull. *46*, 927–936.

Weilnhammer, V.A., Ludwig, K., Sterzer, P., and Hesselmann, G. (2014). Revisiting the Lissajous figure as a tool to study bistable perception. Vis. Res. *98*, 107–112.

Weilnhammer, V.A., Sterzer, P., and Hesselmann, G. (2016). Perceptual stability of the lissajous figure is modulated by the speed of illusory rotation. PLoS one *11*, e0160772.

Weilnhammer, V.A., Stuke, H., Sterzer, P., and Schmack, K. (2018). The neural correlates of hierarchical predictions for perceptual decisions. J. Neurosci. *38*, 5008–5021.

Wexler, M., Duyck, M., and Mamassian, P. (2015). Persistent states in vision break universality and time invariance. Proc. Natl. Acad. Sci. U S A *112*, 14990–14995.

Wilson, R.C., and Collins, A.G. (2019). Ten simple rules for the computational modeling of behavioral data. eLife *8*, e49547.

Zalta, A., Petkoski, S., and Morillon, B. (2020). Natural rhythms of periodic temporal attention. Nat. Commun. *11*, 1–12.

3. Discussion

The research presented above investigates how the brain generates unambiguous conscious experiences from ambiguous sensory information, and how alterations in this process may lead to psychotic symptoms[37,43,50,52,73,80]. At the level of behavior, we found that conscious experience in strongly biased by internal predictions[37,50,73,80]. At the neural level, we observed that these predictive processes are mediated by neural activity that reflects internal predictions and prediction errors at multiple levels of the cortical hierarchy, including supra-modal brain regions in the frontoparietal network[50,52,80]. With respect to psychotic symptoms, we found that inter-individual differences in the sensitivity to prediction errors correlate with inter-individual differences in the severity of hallucinations in patients with paranoid schizophrenia[43].

3.1 Predictive processes shape perceptual inference

A growing body of work suggests that hierarchical predictive coding provides a powerful algorithm for understanding the neural mechanisms of perceptual inference[30,31,83–85]. Our results overlap with these findings by showing that predictions about the stability of the sensory environment[80] and multi-level beliefs induced by cross-modal associative learning[50] shape the experience of ambiguous stimuli. To this end, the studies presented in this habilitation thesis rely on predictive coding algorithms[27–31] that are fitted to the participants' behavior.

One important strength of our predictive coding model[80] is its ability to link across different levels of description regarding perceptual inference. On a computational level, the model understands the construction of unambiguous conscious experience from ambiguous sensory information in the light of Bayesian brain theory[15–17,86] and the free-energy principle[25]. According to these views, the central nervous system has evolved to reduce surprise about itself and the world, providing a unifying view on action, perception and sense-of-self[25,26]. Our work introduces a predictive coding algorithm that transfers these computational ideas to the paradigm of bistable perception. Importantly, it can be applied directly to empirical data[50,80]. The proposed model allowed us not only to study the factors that contribute to the resolution of sensory ambiguity in healthy participants[50,80] and patients diagnosed with paranoid schizophrenia[43], but also to map the neural implementation of predictive processing via model-based fMRI[50,52].

With respect bistable perception, our predictive coding approach enables us to investigate the processing of ambiguous and disambiguated stimuli within the same modeling framework[80]. This bridges the gap between fully ambiguous versus fully disambiguated stimuli (i.e., as used in *replay* conditions[71,73]), extending our investigation to a multitude of factors that contribute to the resolution of sensory ambiguity, such as partial disambiguation[43,52] (Figure 2C), cross-modal learning[50] or perceptual history[37,38]. Our work thereby showcases bistable perception not only as an extreme phenomenon that occurs under artificial lab-conditions, but as a tool to study the construction of unambiguous conscious experiences from ambiguous sensory information in every-day perception[26].

Despite the progress toward partially ambiguous stimuli[50,52], ecological validity remains an important limitation to consider. Our work has used a *shallow* predictive coding model[31,50,80] that explains binary perceptual responses to simplistic stimuli designed for the artificial context of psychophysical experiments[52,80,87]. Importantly, our model lacks the depth and granularity necessary to explain how perceptual inference is realized at the level of biological neural networks that support the recognition of complex naturalistic stimuli[88,89]. Deep artificial neural networks[90–93] may provide one way to progress toward a better understanding of predictive processing in general object recognition. Indeed, activity at the nodes of deep artificial neural networks trained in the task of object recognition correlate with neural activity in biological neural networks dedicated to object recognition in humans[94] and

macaques[89,95], suggesting that artificial and biological neural networks may rely functionally equivalent computations in object recognition[89,91,94]. Biologically plausible deep artificial neural networks[92] may therefore provide important insights on how perception, cognition and behavior are linked to predictive processes[96] at the level of individual neurons and across the hierarchy of cortical processing[89,95].

3.2 Inferior frontal cortex regulates the access of conflicting information into conscious experience

Bistable perception is a key paradigm for understanding the computational principles of perceptual inference[21]. In this context, one of the most heavily debated questions concerns the role of frontoparietal cortex in bistable perception[69]: While some authors have proposed that frontoparietal cortex selects what is consciously perceived[62,66,72,80], others have related frontoparietal activity to cognitive functions that unfold as a consequence of transitions in conscious experience[97,98], most notably the processing of perceptual uncertainty[71] and motor behavior[74].

The research presented in this habilitation thesis contributes to this debate in four ways: First, while controlling for the potential confound of perceptual uncertainty[71], we show that frontoparietal BOLD activity is elevated at the time of transitions in conscious experience during bistable perception compared to an unambiguous control condition[73]. Second, we show that effective connectivity from IFC to visual cortex is elevated at the time of transitions in conscious experience during bistable perception, supporting the causal role of frontoparietal cortex in triggering these events[73]. Third, our findings indicate that virtual lesions in IFC reduce the frequency of transitions in conscious experience during bistable perception[52] without effecting reports on perceptual uncertainty[71] and motor responses[74,97]. Forth, we link IFC activity to the accumulation of prediction errors[52,80], suggesting that frontoparietal cortex regulates the access of conflicting sensory information into conscious experience perception[69,99]

These results provide compelling evidence for an active role of frontoparietal cortex in bistable perception in particular and conscious experience in general. However, more studies are needed to better dissociate the function of regulating access to conscious experience from alternative cognitive functions that occur up- or downstream of phenomenal consciousness[26]. Most notably, this concerns potential links between frontoparietal cortex and attention[100,101] and report[74,97].

With respect to the functional neuroanatomy of attention, the frontoparietal network has been divided into a ventral subset that interrupts ongoing activity, and a dorsal subset that links incoming sensory inputs to behavioral outputs[100]. As part of the ventral frontoparietal network, the IFC may therefore exert its influence on conscious experience during bistable perception through its role in attention. This link attention and the role of IFC in bistable perception is further substantiated by studies showing that withdrawing attention from bistable stimuli reduces the frequency of transitions in conscious experience[101]. While our work has found no change in response times (a proxy of on-task attention[102]) following virtual IFC-lesions, more work is needed to better understand the role of attention for IFC activity and the processing of sensory ambiguity. For example, future studies could test whether parametric modulations of on-task attention correlate with changes in results obtained from computational modeling or TMS-induced IFC-lesion-effects.

An additional caveat concerns the role of prefrontal cortex in reporting behavior. Previous work has shown that activity in the frontoparietal network is reduced in paradigms that detect transitions in conscious experience during bistable perception in the absence of active report (e.g., by decoding the content of conscious experience via pupil-size or optokinetic nystagmus[74]). Importantly, however, the observed reduction in BOLD activity[74] typically spares the IFC[69,74]. Moreover, our own results do not show an effect of virtual IFC lesions on reporting behavior[52]. Overall, the evidence available to date is

thus mixed with regard to the question whether frontoparietal cortex has a causal role in bistable perception[74,97,103–106].

In a broader sense, frontoparietal activity may be related to a range of post-perceptual cognitive processes beyond motor report, such as self-monitoring, introspection, meta-cognition or cognitive control[107–109]. For example, an important fMRI study has shown that frontoparietal activity is absent when transitions in conscious experience become inconspicuous to the extent that they are not noticed by the observer[97]. To further corroborate that the functions supported by frontoparietal brain activity are not exclusively situated down-stream of the resolution of sensory ambiguity, future experiments could optimize the assessment of self-monitoring, introspection or meta-cognition to test whether virtual IFC lesions impact additional cognitive functions beyond the frequency of transitions in conscious experience during bistable perception[74,97,103,109].

3.3 Imbalances in psychotic experiences drive psychotic experiences

The above considerations link frontoparietal cortex to an important homeostatic function: By regulating the processing of conflicting information, frontoparietal cortex may stabilize conscious experience against uninformative fluctuations in ambiguous sensory data. A breakdown of this function may thus trigger mal-adaptive changes in conscious experience that contribute to psychotic symptoms[40]. In line with this, the research presented here shows that inter-individual differences in hallucinatory experiences correlate to inter-individual differences in the balance between internal predictions and prediction errors that are driven by ambiguous external sensory information[43]. Our work thereby links the study of bistable perception to one of the most heavily debated problems in computational psychiatry[56–58], namely whether hallucinations are driven by relying to strongly on internal predictions, or, alternatively, by excessive prediction errors[40–42,59].

Our results show that, in a paradigm based on partially disambiguated bistable stimuli (Figure 2C), the sensitivity to conflicting yet ambiguous sensory information correlates with the severity of hallucinatory experiences in patients diagnosed with paranoid schizophrenia[43]. These findings therefore support the weak prior account[40,59] and align with previous work on contextual illusions that suggests a link between excessive prediction errors and psychotic symptoms[51,60].

However, a growing body of work has argued for the opposite by associating psychotic symptoms with an overly strong reliance on internal predictions[39,44,110,111] (i.e., the strong prior account). Lastly, there may be a bi-directional relationship between internal predictions and prediction errors[61] that may lead to corresponding changes across multiple levels of the cortical hierarchy[40,42]. One way to resolve this discrepancy may be to use interventional designs (i.e., pharmacological models based on dopamine[44,53,54] and ketamine[44,55,112] or TMS-induced virtual-lesions in prediction- and prediction-error-related brain areas[50,52]) that modulate psychotic experiences via the interplay of internal predictions with prediction errors. Such a research program could help to identify the neuorocomputational anatomy of psychosis[59] and pave the way toward new possibilities for the therapeutic modulation of hallucinatory experiences.

4. Summary

The research presented in this habilition thesis seeks to understand the construction of unambiguous conscious experiences from ambiguous sensory sensory information[43,50,52,73,80].

At the computational level, our work builds on the idea that the brain applies predictive processes to resolve the ambiguity inherent in sensory information[15–20]. Using the phenomenon of bistable perception, we developed an algorithm based on predictive coding[27–31] to measure the interaction of internal prediction with prediction errors that are driven my ambiguous sensory information[50,80].

At the level of neural implementation, we combined model-based fMRI and TMS-induced virtual lesions to show that the processing of sensory ambiguity is not limited to feature-selective regions in sensory brain areas, but involves supra-modal brain regions of the frontoparietal network[50,52,73,80]. In particular, our work proposes a key role for the IFC in regulating the access of conflicting information into conscious experience[52].

Finally, we show that the interaction of internal prediction with external sensory information may modulate hallucinatory experiences in patients diagnosed with paranoid schizophrenia[43].

In sum, the research presented in this habilitation thesis contributes to a neurocomputational understanding of how ambiguous sensory information is transformed into unambiguous conscious experiences, and suggests how alterations in this process may lead to hallucinations. Our computational[80] an lesion-based[52] approach will enable future research to advance the scientific understanding at two of the most important frontiers in contemporary neuroscience: the biology of consciousness[1–11] and neurocomputational theories of psychosis[39–42,44].

5. Literature

1. Crick, F. Consciousness and neuroscience. *Cerebral Cortex* **8**, 97–107 (1998).

2. Rees, G. *et al.* Neural correlates of consciousness in humans. *Nature reviews. Neuroscience* **3**, 261–70 (2002).

3. Miller, G. What is the biological basis of consciousness? *Science* **309**, 79 (2005).

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

5. Aru, J. *et al.* Distilling the neural correlates of consciousness. *Neuroscience and Biobehavioral Reviews* **36**, 737–746 (2012).

6. Boly, M. *et al.* Consciousness in humans and non-human animals: Recent advances and future directions. *Frontiers in Psychology* **4**, 1–20 (2013).

7. Adolphs, R. The unsolved problems of neuroscience. *Trends in Cognitive Sciences* **19**, 173–175 (2015).

8. Koch, C. *et al.* Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience* **17**, 307–321 (2016).

9. Dehaene, S. *et al.* What is consciousness, and could machines have it? *Science* **358**, 486–492 (2017).

10. Michel, M. *et al.* Opportunities and challenges for a maturing science of consciousness. *Nature Human Behaviour* **3**, 104–107 (2019).

11. Sohn, E. Decoding the neuroscience of consciousness. *Nature* **571**, S2—–S5 (2019).

12. Stewart, E. E. M. *et al.* A review of interactions between peripheral and foveal vision. *Journal of Vision* **20**, 1–35 (2020).

13. Kersten, D. *et al.* Object Perception as Bayesian Inference. *Annual Review of Psychology* **55**, 271–304 (2004).

14. Barthelmé, S. *et al.* Evaluation of objective uncertainty in the visual system. *PLoS computational biology* **5**, e1000504 (2009).

15. Hohwy, J. Attention and conscious perception in the hypothesis testing brain. *Frontiers in psychology* **3**, 96 (2012).

16. Lee, T. S. *et al.* Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America. A, Optics, image science, and vision* **20**, 1434–48 (2003).

17. Knill, D. C. *et al.* The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).

18. Friston, K. A theory of cortical responses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **360**, 815–836 (2005).

19. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and brain sciences* **36**, 181–204 (2013).

20. Clark, A. *Surfing Uncertainty: Prediction, Action and The Embodied Mind.* (Oxford University Press, 2016).

21. Hohwy, J. *et al.* Predictive coding explains binocular rivalry: an epistemological review. *Cognition* **108**, 687–701 (2008).

22. Herculano-Houzel, S. The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences* **109**, 10661–10668 (2012).

23. The Power Consumption Database. Game Consoles power consumption.

24. Helmholtz, H. von. *Handbuch der Physiologischen Optik*. (Voss, 1867).

25. Friston, K. The free-energy principle: a unified brain theory? *Nature reviews. Neuroscience* **11**, 127–38 (2010).

26. Seth, A. K. *Being you : a new science of consciousness*. (Dutton Books, 2022).

27. Rao, R. P. *et al.* Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience* **2**, 79–87 (1999).

28. Bastos, A. M. *et al.* Canonical microcircuits for predictive coding. *Neuron* **76**, 695–711 (2012).

29. Friston, K. *et al.* Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**, 1211–1221 (2009).

30. Iglesias, S. *et al.* Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. *Neuron* **80**, 519–530 (2013).

31. Mathys, C. D. *et al.* Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in human neuroscience* **8**, 825 (2014).

32. Eckmann, S. *et al.* Active efficient coding explains the development of binocular vision and its failure in amblyopia. *Proceedings of the National Academy of Sciences* **117**, 6156–6162 (2020).

33. Ouden, H. E. M. den *et al.* Striatal Prediction Error Modulates Cortical Coupling. The *Journal of Neuroscience* **30**, 3210–3219 (2010).

34. Odegaard, B. *et al.* Should a Few Null Findings Falsify Prefrontal Theories of Conscious Perception? *The Journal of Neuroscience* **37**, 9593–9602 (2017).

35. Boly, M. *et al.* Are the Neural Correlates of Consciousness in the Front or in the Back of the Cerebral Cortex? Clinical and Neuroimaging Evidence. *The Journal of Neuroscience* **37**, 9603–9613 (2017).

36. Schmack, K. *et al.* Learning What to See in a Changing World. *Frontiers in human neuroscience* **10**, 263 (2016).

37. Weilnhammer, V. *et al.* Bistable perception alternates between internal and external modes of sensory processing. *iScience* **24**, 102234 (2021).

38. Weilnhammer, V. *et al.* Humans and mice fluctuate between external and internal modes of sensory processing. *bioRxiv* 2021.08.20.457079 (2021) doi:10.1101/2021.08.20.457079.

39. Powers, A. R. *et al.* Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science* **357**, 596–600 (2017).

40. Sterzer, P. *et al.* The Predictive Coding Account of Psychosis. *Biological Psychiatry* **84**, 634–643 (2018).

41. Horga, G. *et al.* An integrative framework for perceptual disturbances in psychosis. *Nature Reviews Neuroscience* **20**, 763–778 (2019).

42. Corlett, P. R. *et al.* Hallucinations and Strong Priors. *Tics* **23**, 114–127 (2019).

43. Weilnhammer, V. *et al.* Psychotic Experiences in Schizophrenia and Sensitivity to Sensory Evidence. *Schizophrenia bulletin* **46**, 927–936 (2020).

44. Schmack, K. *et al.* Striatal dopamine mediates hallucination-like perception in mice. *Science* **372**, (2021).

45. Sommer, I. E. C. *et al.* Auditory verbal hallucinations predominantly activate the right inferior frontal area. *Brain* **131**, 3169–77 (2008).

46. Horga, G. *et al.* Deficits in predictive coding underlie hallucinations in schizophrenia. *The Journal of Neuroscience* **34**, 8072–82 (2014).

47. Penfield, W. *et al.* The brain's record of auditory and visual experience: A final summary and discussion. *Brain* **86**, 595–696 (1963).

48. Egner, T. *et al.* Expectation and surprise determine neural population responses in the ventral visual stream. The *Journal of Neuroscience* **30**, 16601–16608 (2010).

49. Keller, G. B. *et al.* Sensorimotor Mismatch Signals in Primary Visual Cortex of the Behaving Mouse. *Neuron* **74**, 809–815 (2012).

50. Weilnhammer, V. A. *et al.* The Neural Correlates of Hierarchical Predictions for Perceptual Decisions. *The Journal of Neuroscience* **38**, 5008–5021 (2018).

51. Dakin, S. *et al.* Weak suppression of visual context in chronic schizophrenia. *Current biology : CB* **15**, R822–4 (2005).

52. Weilnhammer, V. *et al.* An active role of inferior frontal cortex in conscious experience. *Current Biology* **31**, 2868–2880.e8 (2021).

53. Cassidy, C. M. *et al.* A Perceptual Inference Mechanism for Hallucinations Linked to Striatal Dopamine. *Current Biology* **28**, 503–514.e4 (2018).

54. Howes, O. D. *et al.* The dopamine hypothesis of schizophrenia: Version III - The final common pathway. *Schizophrenia Bulletin* **35**, 549–562 (2009).

55. Corlett, P. R. *et al.* Prediction error, ketamine and psychosis: An updated model. *Journal of Psychopharmacology* **30**, 1145–1155 (2016).

56. Stephan, K. E. *et al.* Computational approaches to psychiatry. *Current opinion in neurobiology* **25**, 85–92 (2014).

57. Wang, X.-J. *et al.* Computational Psychiatry. *Neuron* **84**, 638–654 (2014).

58. Huys, Q. J. M. *et al.* Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience* **19**, 404–413 (2016).

59. Adams, R. A. *et al.* The computational anatomy of psychosis. *Frontiers in psychiatry* **4**, 47 (2013).

60. Shergill, S. S. *et al.* Evidence for sensory prediction deficits in schizophrenia. *The American journal of psychiatry* **162**, 2384–6 (2005).

61. Jardri, R. *et al.* Experimental evidence for circular inference in schizophrenia. *Nature Communications* **8**, 14218 (2017).

62. Leopold, D. A. *et al.* Multistable phenomena: changing views in perception. *Trends in Cognitive Sciences* **3**, 254–264 (1999).

63. Logothetis, N. K. *et al.* What is rivalling during binocular rivalry? *Nature* **380**, 621–4 (1996).

64. Tong, F. *et al.* Neural bases of binocular rivalry. *Trends in cognitive sciences* **10**, 502–11 (2006).

65. Blake, R. *et al.* Visual competition. *Nature Reviews Neuroscience* **3**, 13–21 (2002).

66. Sterzer, P. *et al.* The neural bases of multistable perception. *Tics* **13**, 310–8 (2009).

67. Sterzer, P. *et al.* A neural basis for inference in perceptual ambiguity. *Proceedings of the National Academy of Sciences* **104**, 323–8 (2007).

68. Gershman, S. J. *et al.* Multistability and perceptual inference. *Neural computation* **24**, 1–24 (2012).

69. Brascamp, J. *et al.* Multistable Perception and the Role of the Frontoparietal Cortex in Perceptual Inference. *Annual Review of Psychology* **69**, 77–103 (2018).

70. Safavi, S. *et al.* Multistability, perceptual value, and internal foraging. *Neuron* **110**, 3076–3090 (2022).

71. Knapen, T. *et al.* The Role of Frontal and Parietal Brain Areas in Bistable Perception. *The Journal of Neuroscience* **31**, 10293–10301 (2011).

72. Lumer, E. D. *et al.* Neural correlates of perceptual rivalry in the human brain. *Science* **280**, 1930–4 (1998).

73. Weilnhammer, V. A. *et al.* Frontoparietal cortex mediates perceptual transitions in bistable perception. *The Journal of Neuroscience* **33**, 16009–15 (2013).

74. Frässle, S. *et al.* Binocular rivalry: frontal activity relates to introspection and action but not to perception. *The Journal of Neuroscience* **34**, 1738–47 (2014).

75. Sundareswara, R. *et al.* Perceptual multistability predicted by search model for Bayesian decisions. *Journal of Vision* **8**, 12.1–19 (2008).

76. Michel, M. *et al.* On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences* **1**, (2020).

77. Schmack, K. *et al.* Delusions and the Role of Beliefs in Perceptual Inference. The *Journal of Neuroscience* **33**, (2013).

78. Schmack, K. *et al.* Perceptual instability in schizophrenia: Probing predictive coding accounts of delusions with ambiguous stimuli. *Schizophrenia Research: Cognition* **2**, 72–77 (2015).

79. Schmack, K. *et al.* Enhanced predictive signalling in schizophrenia. *Human Brain Mapping* **38**, 1767–1779 (2017).

80. Weilnhammer, V. *et al.* A predictive coding account of bistable perception - a model-based fMRI study. *PLOS Computational Biology* **13**, e1005536 (2017).

81. Ramachandran, V. S. *et al.* The perception of apparent motion. *Scientific American* **254**, 102–109 (1986).

82. O'Shea, J. *et al.* Transcranial magnetic stimulation. *Current Biology* **17**, R196–R199 (2007).

83. Teufel, C. *et al.* The role of priors in Bayesian models of perception. *Frontiers in Computational Neuroscience* **7**, 25 (2013).

84. Teufel, C. *et al.* Forms of prediction in the nervous system. *Nature Reviews Neuroscience* **21**, 231–242 (2020).

85. Hohwy, J. *et al.* Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences* **1**, (2020).

86. Friston, K. J. *et al.* Free-energy and the brain. *Synthese* **159**, 417–458 (2007).

87. Weilnhammer, V. A. *et al.* Revisiting the Lissajous figure as a tool to study bistable perception. *Vision research* **98**, 107–12 (2014).

88. Hesse, J. K. *et al.* A new no-report paradigm reveals that face cells encode both consciously perceived and suppressed stimuli. *eLife* **9**, (2020).

89. Bao, P. *et al.* A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).

90. Krizhevsky, A. *et al.* ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems* **25**, (2012).

91. Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* **111**, 8619–8624 (2014).

92. Storrs, K. R. *et al.* Deep Learning for Cognitive Neuroscience. *Arxiv* **abs/1903.0**, (2019).

93. Zhou, Y. *et al.* Deep Learning in Next-Frame Prediction: A Benchmark Review. vol. 8 69273–69283 (2020).

94. Wen, H. *et al.* Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cerebral Cortex* **28**, 4136–4160 (2018).

95. Chang, L. *et al.* The Code for Facial Identity in the Primate Brain. *Cell* **169**, 1013–1028.e14 (2017).

96. Lotter, W. *et al.* Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings* (2016).

97. Brascamp, J. *et al.* Negligible fronto-parietal BOLD activity accompanying unreportable switches in bistable perception. *Nature neuroscience* **18**, 1672–1678 (2015).

98. Zou, J. *et al.* Binocular rivalry from invisible patterns. *Proceedings of the National Academy of Sciences* **113**, 8408–8413 (2016).

99. Kanai, R. *et al.* Structural and functional fractionation of right superior parietal cortex in bistable perception. *Current Biology* **21**, R106–R107 (2011).

100. Corbetta, M. *et al.* The Reorienting System of the Human Brain: From Environment to Theory of Mind. *Neuron* **58**, 306–324 (2008).

101. Alais, D. *et al.* Attending to auditory signals slows visual alternations in binocular rivalry. *Vision Research* **50**, 929–935 (2010).

102. Prado, J. *et al.* Variations of response time in a selective attention task are linked to variations of functional connectivity in the attentional network. *NeuroImage* **54**, 541–549 (2011).

103. Lumer, E. D. *et al.* Covariation of activity in visual and prefrontal cortex associated with subjective visual perception. *Proceedings of the National Academy of Sciences* **96**, 1669–1673 (1999).

104. Panagiotaropoulos, T. I. *et al.* Neuronal Discharges and Gamma Oscillations Explicitly Reflect Visual Consciousness in the Lateral Prefrontal Cortex. *Neuron* **74**, 924–935 (2012).

105. Kapoor, V. *et al.* Decoding the contents of consciousness from prefrontal ensembles. *bioRxiv* 2020.01.28.921841 (2020) doi:10.1101/2020.01.28.921841.

106. Dwarakanath, A. *et al.* Prefrontal state fluctuations control access to consciousness. *bioRxiv* 2020.01.29.924928 (2020) doi:10.1101/2020.01.29.924928.

107. Graaf, T. A. de *et al.* On the functional relevance of frontal cortex for passive and voluntarily controlled bistable vision. *Cerebral cortex (New York, N.Y. : 1991)* **21**, 2322–31 (2011).

108. Aron, A. R. *et al.* Inhibition and the right inferior frontal cortex: One decade on. *Tics* **18**, 177–185 (2014).

109. Tsuchiya, N. *et al.* No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. *Tics* **19**, 757–770 (2015).

110. Teufel, C. *et al.* Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. *Proceedings of the National Academy of Sciences* **112**, 13401–6 (2015).

111. Davies, D. J. *et al.* Anomalous Perceptions and Beliefs Are Associated With Shifts Toward Different Types of Prior Knowledge in Perceptual Inference. *Schizophrenia Bulletin* **44**, 1245–1253 (2018).

112. Corlett, P. R. *et al.* Glutamatergic model psychoses: prediction error, learning, and inference. *Neuropsychopharmacolog* **36**, 294–315 (2011).

Acknowledgements

The research that lead to this habilitation thesis has been shaped by countless conversations with friends, colleagues, teachers, and mentors. I am extremely grateful to all of you!

I would like to thank all the past and present colleagues at the Visual Perception Laboratory and at the Department of Psychiatry, Charité Universitätsmedizin Berlin, Campus Mitte. In particular, I would like to thank Meera Chikermane, Anna-Lena Eckert, Dr. med. Merve Fritsch, PD Dr. rer. nat. Dipl.-Phys. Matthias Guggenmos, Prof. Dr. med. Rainer Hellweg, Dr. med. Jakob Kaminski, Dr. med. Katharina Kanthak, Prof. Dr. med. Stephan Köhler, Dr. rer. nat. Karin Ludwig, Dr. med. Jochen Michely, PD Dr. med. Christian Müller, Lukas Röd, Dr. rer. medic. Dipl.-Psych. Marcus Rothkirch, Dr. rer. medic. M. Sc. Maria Sekutowicz, Dr. med. Heiner Stuke, and Dr. med. Constantin Volkmann.

I am beyond grateful to Dr. med. Katharina Schmack and to Prof. Dr. Guido Hesselmann for their invaluable mentoring and supervision.

I would like to thank Prof. Dr. med. Dr. phil. Andreas Heinz, director of the Department of Psychiatry, Charité Universitätsmedizin Berlin, Campus Mitte, for his generous and ongoing support that has enabled me to pursue a scientific career in parallel to my training as a psychiatrist.

I would like to express my deepest gratitude to Prof. Dr. med. Philipp Sterzer. Having worked with Philipp for more than ten years, I can say with confidence that I could not have hoped for a better supervisor. It has been and continues to be a privilege to work with such a stellar mentor and scientist.

I am very grateful to all the Berlin Institute of Health, who supported my research through the Clinician Scientist Program.

Finally, thank you to my family, my friends, and my partner Anouk. This journey would not have been possible without you!

Declaration

§ 4 Abs. 3 (k) der HabOMed der Charité

Hiermit erkläre ich, dass

- weder früher noch gleichzeitig ein Habilitationsverfahren durchgeführt oder angemeldet wurde,
- die vorgelegte Habilitationsschrift ohne fremde Hilfe verfasst, die beschriebenen Ergebnisse selbst gewonnen sowie die verwendeten Hilfsmittel, die Zusammenarbeit mit anderen Wissenschaftlern/Wissenschaftlerinnen und mit technischen Hilfskräften sowie die verwendete Literatur vollständig in der Habilitationsschrift angegeben wurden,
- mir die geltende Habilitationsordnung bekannt ist.

Ich erkläre ferner, dass mir die Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis bekannt ist und ich mich zur Einhaltung dieser Satzung verpflichte.

…… ……..

Datum Unterschrift