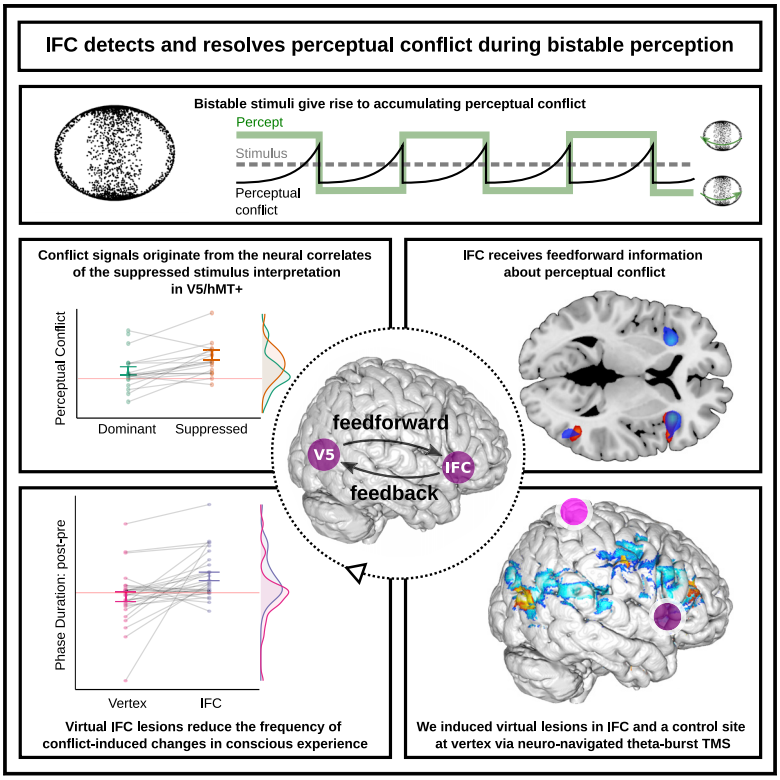
Article

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

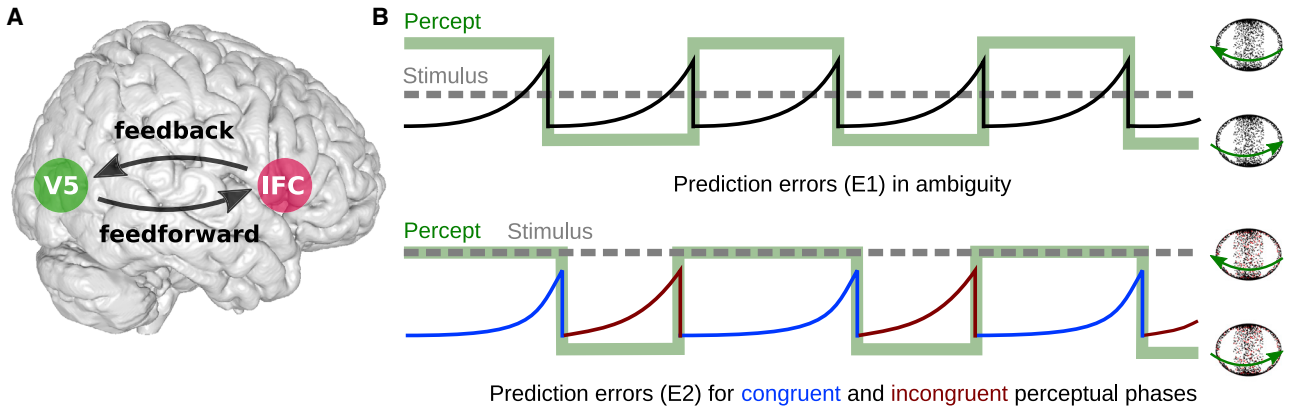


Figure 1. Concept

(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

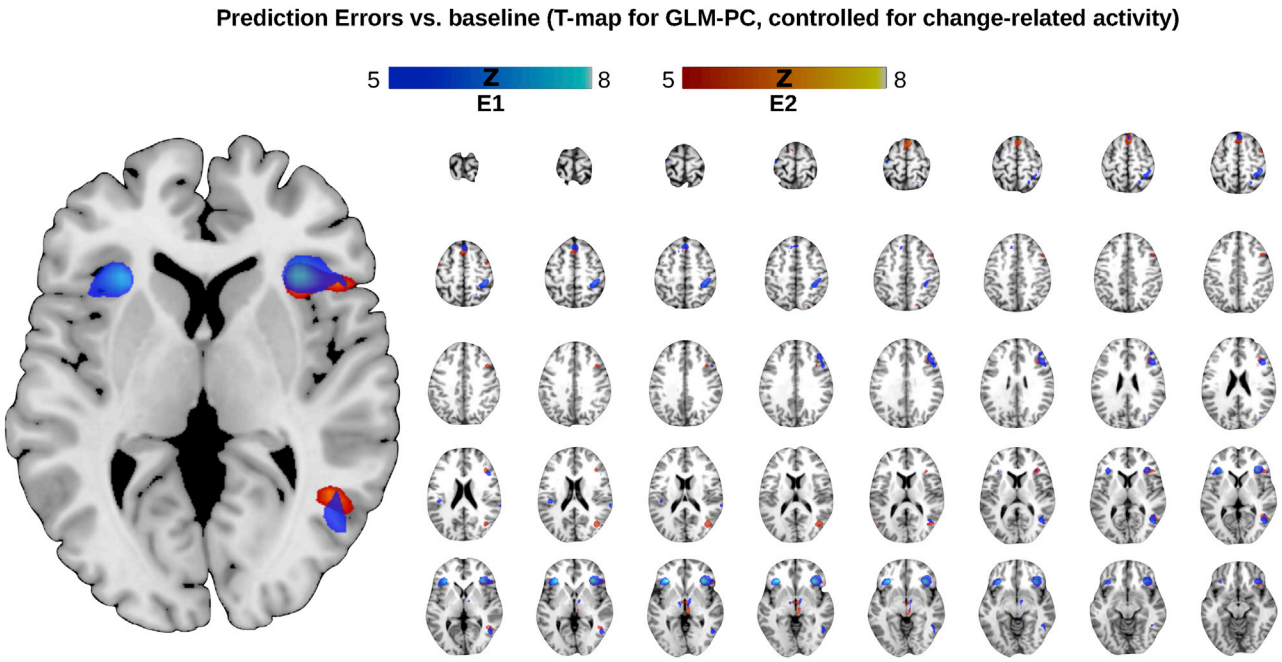


Figure 2. The neural correlates of accumulating perceptual conflict

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]

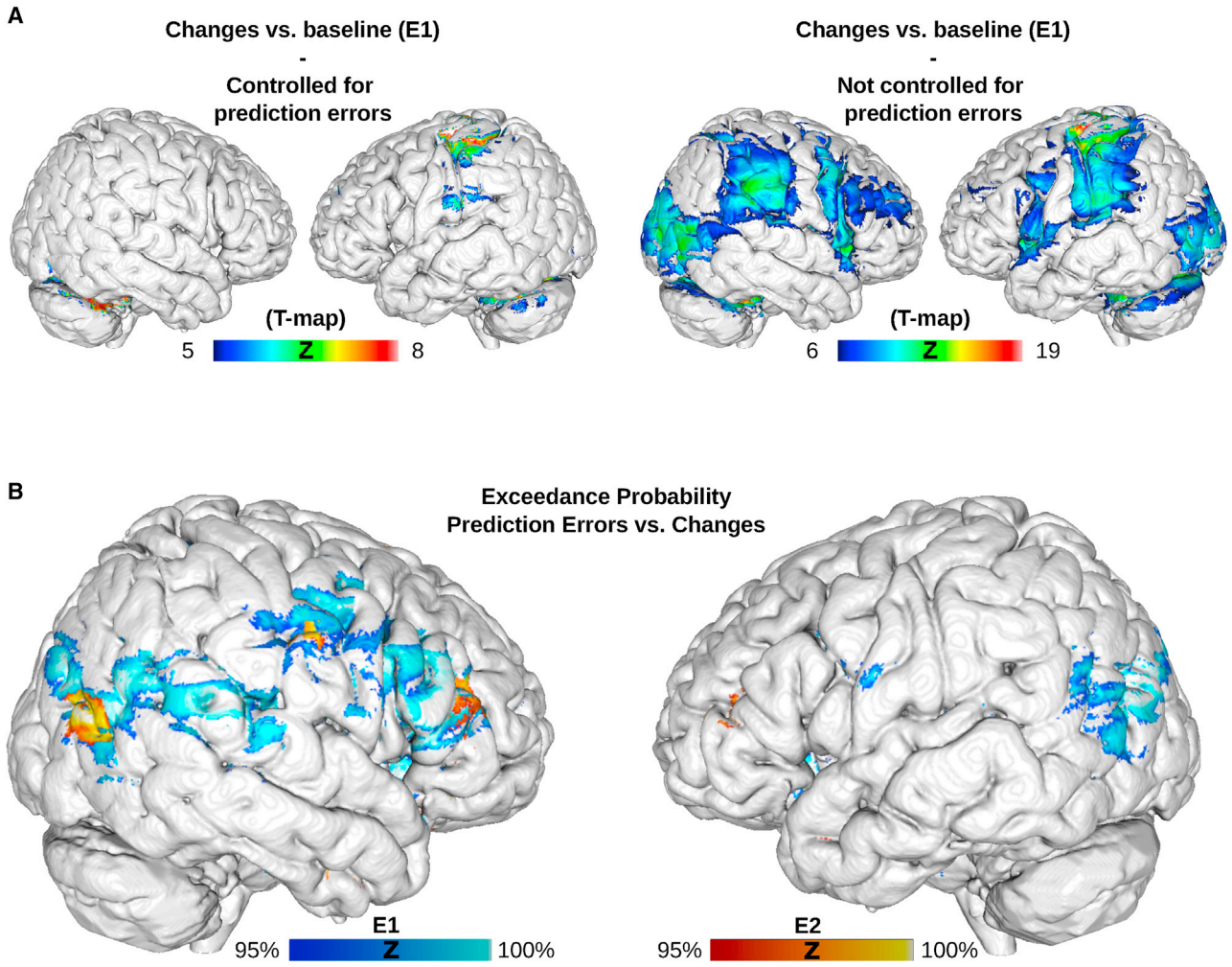


Figure 3. Conflict- versus change-related BOLD activity

(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).

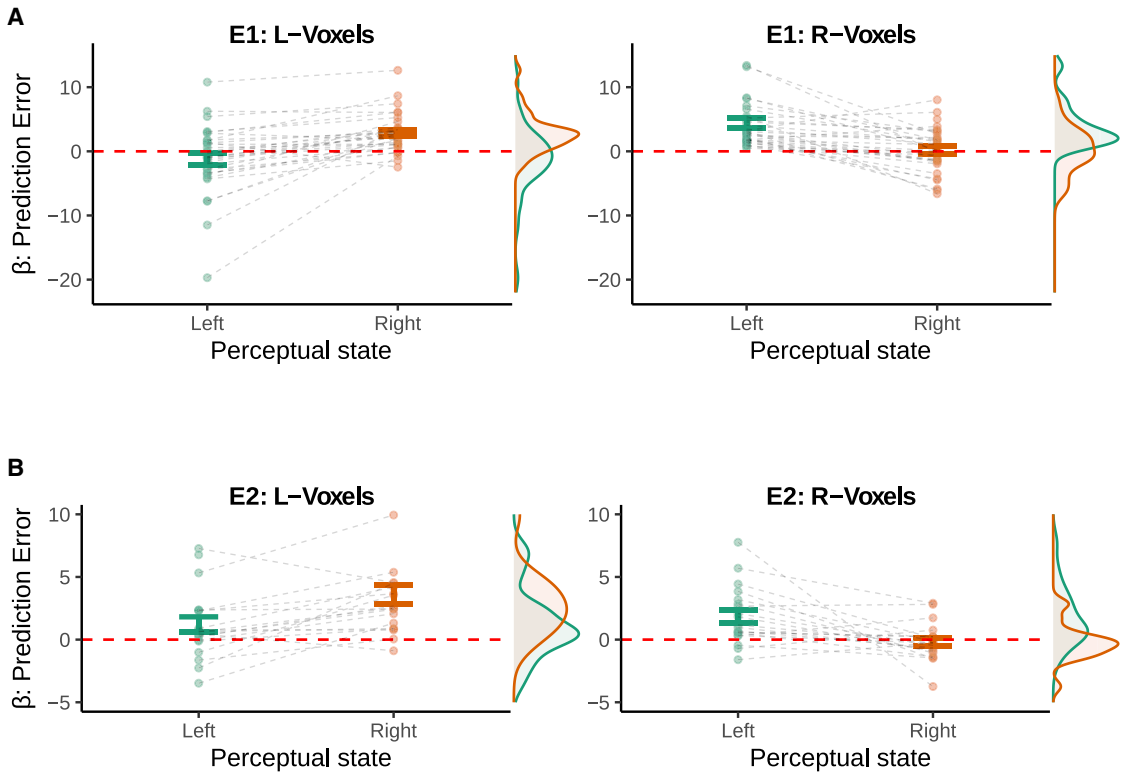


Figure 4. Neural correlates of perceptual conflict in V5/hMT+

(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

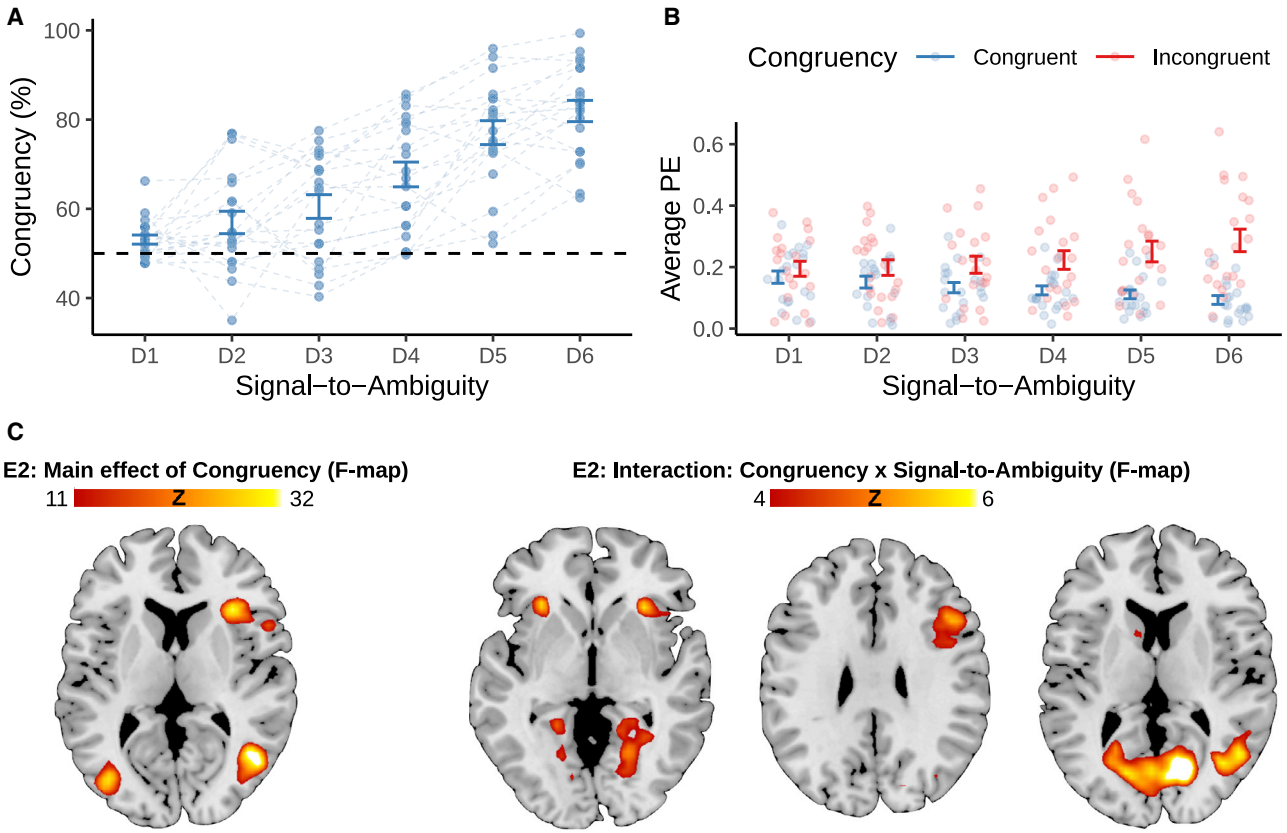


Figure 5. Perceptual conflict during graded ambiguity

(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.

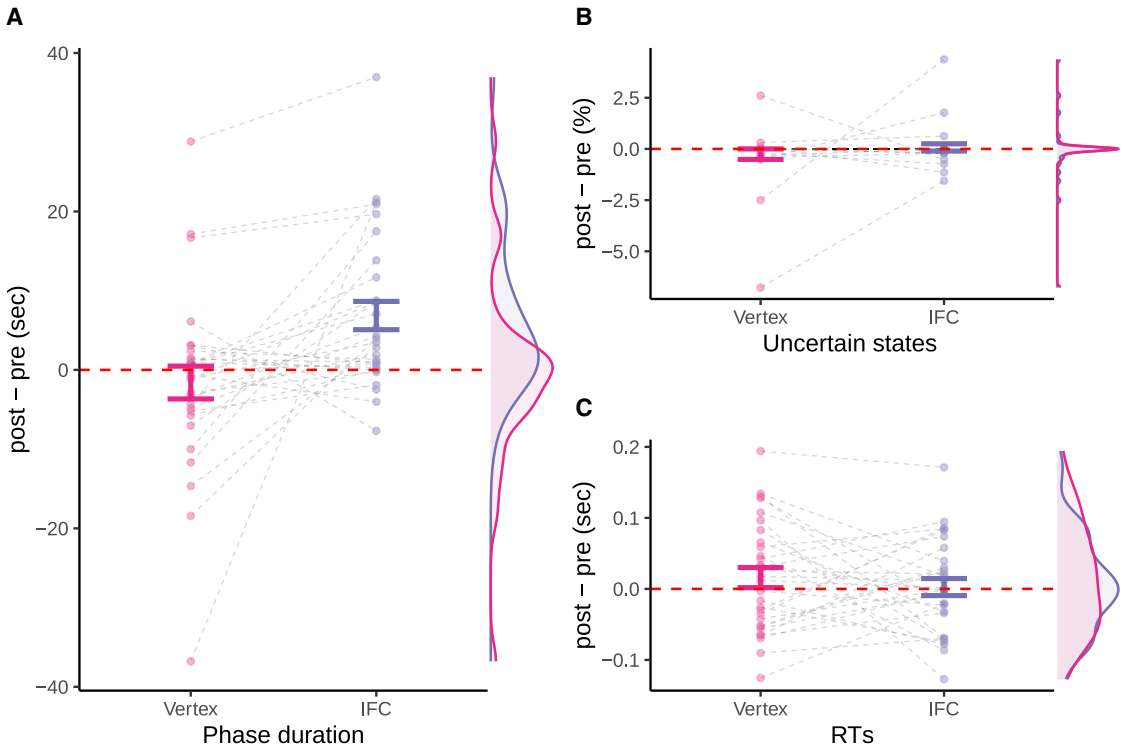


Figure 6. TMS effects on perception

(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).

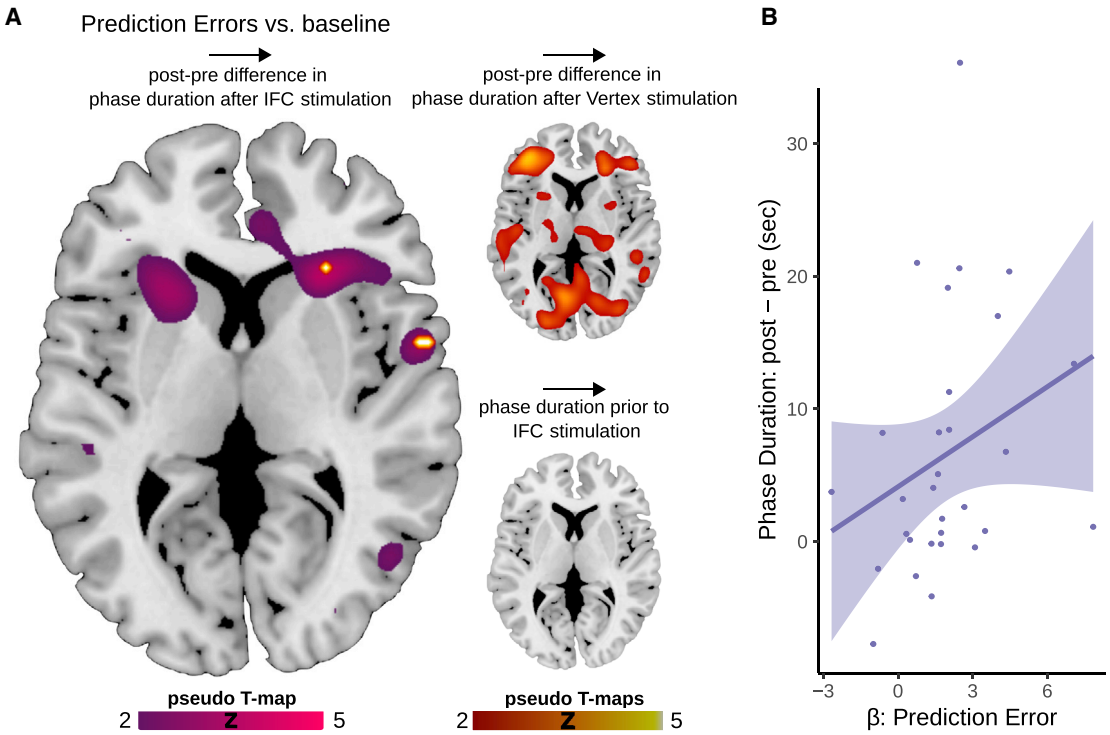


Figure 7. Brain-behavior associations

(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 $P(\theta)$:

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

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

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

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 $(\mathcal{N}(\mu_{stabiliy}, \pi^{-1}_{stability}))$ 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 $\theta_1 = 1$:

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

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

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.