Computational modelling shows evidence in support of both sensory and frontal theories of consciousness

Kavindu H. Bandara[1], Elise G. Rowe[1], Marta I. Garrido[1, 2]

[1] Melbourne School of Psychological Sciences, The University of Melbourne, Australia

[2] Graeme Clark Institute for Biomedical Engineering, The University of Melbourne, Australia

Abstract

The role of the prefrontal cortex (PFC) in consciousness is hotly debated. Frontal theories argue that the PFC is necessary for consciousness, while sensory theories propose that consciousness arises from recurrent activity in the posterior cortex alone, with activity in the PFC resulting from the mere act of reporting. To resolve this dispute, we re-analysed an EEG dataset of 30 participants from a no-report inattentional blindness paradigm where faces are (un)consciously perceived. Dynamic causal modelling was used to estimate the effective connectivity between the key contended brain regions, the prefrontal and the posterior cortices. Then, a second-level parametric empirical Bayesian model was conducted to determine how connectivity was modulated by awareness and task-relevance. While an initial data-driven search could not corroborate neither sensory nor frontal theories of consciousness, a more directed hypothesis-driven analysis revealed strong evidence that both theories could explain the data, with a very slight preference for frontal theories. Specifically, a model with backward connections switched off within the posterior cortex explained awareness better (53%) than a model without backward connections from the PFC to sensory regions. Our findings provide some support for a subtle, yet crucial, contribution of the frontal cortex in consciousness, and highlight the need to revise current theories of consciousness.

Keywords: Consciousness, PFC, Computational Modelling, Inattentional Blindness, Dynamic Causal Modelling, Parametric Empirical Bayes, Bayesian

1.1 Introduction

The necessity of the prefrontal cortex (PFC) in generating consciousness remains hotly debated (Michel & Morales, 2020; Cogitate Consortium, 2023). To date, empirical support has been largely mixed, and the role of the PFC has emerged as a key point of divergence between the current major theories of consciousness. These theories can be split into two theoretical camps suggesting distinct brain regions are involved in conscious perception (Figure 1a and 1b). The sensory (or posterior) family of theories argues that consciousness is accounted for fully by activity within posterior regions (as in Information Integration Theory, IIT, [Oizumi et al., 2014; Tononi et al., 2016] and Recurrent Processing Theories, RPT, [Lamme, 2006; Lamme & Roelfsema, 2000]). Frontal theories, on the other hand, propose that the PFC is necessary in the generation of consciousness (as in Global Neuronal Workspace Theory, GNWT, [Dehaene & Naccache, 2001; Mashour et al., 2020] and Higher Order Thought Theories, HOTs, [Brown et al., 2019; Lau & Rosenthal, 2011]).

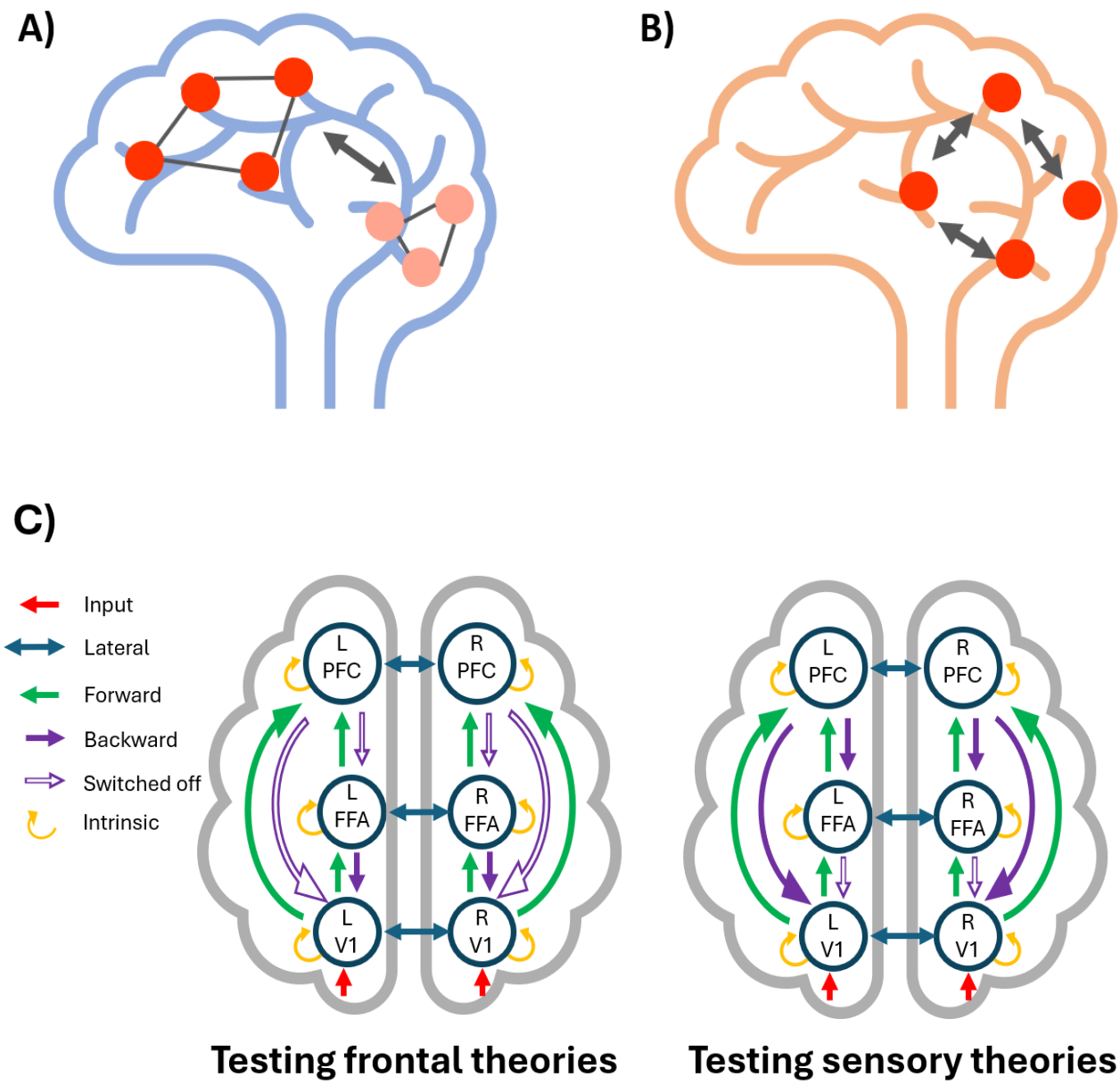

Figure 1. *Comparison of Global Neuronal Workspace Theory and Recurrency Theory and the 2 Models Used in our Hypothesis-driven Analysis to Test Predictions of Sensory and Frontal Theories.* A comparison of two major theories of consciousness with differing views on the necessity of the PFC in consciousness. A) Global neuronal workspace theory proposes that a frontoparietal network is involved in generating consciousness. When activated, this network broadcasts information across the cortex, enabling access of information from local cortical units. The theory itself predicts more local processors are involved, but for the sake of illustration, a simplified model is depicted here. The key point is that the activation of prefrontal areas is causally involved in the generation of consciousness. B) In recurrent

processing theory, feedback activity in recurrent loops within sensory cortices results in consciousness. C) The two models used in a direct model comparison in this paper using two reduced (or nested) PEB models. These models differed such that key backward connections from either model were fixed at a prior mean of 0 (effectively, switching off the connection). Connections switched off are denoted by hollow arrows in the diagram. On the left model, backward connections from the PFC to lower sensory regions (i.e. PFC-FFA and PFC-V1) are switched off. If losing these connections greatly impacts the model fit, then this suggests that PFC connections are important for awareness, corroborating frontal theories. On the right model, backward connections from the FFA to V1 were turned off. Here, if the model suffers, it suggests that these FFA to V1 connections are important, favouring sensory theories. Here, PFC refers to the prefrontal cortex (Brodmann area 45), FFA is the fusiform face area, and V1 is the primary visual cortex.

Given the well-established role of the PFC in executive control, working memory and other higher-cognitive functions, posterior theories argue that PFC activity is elicited only in accessing and reporting on consciousness, rather than in generating consciousness itself. As the confound of report is difficult to control for, according to sensory theories PFC activity contaminates the true neural correlates of consciousness (NCC; Brascamp et al., 2015; Frässle et al., 2014; Koch et al., 2016; Tsuchiya et al., 2015).

No-report paradigms have been particularly powerful in dissociating the various functions involved in reporting a stimulus from the conscious experience of that stimulus itself. Typically, this is accomplished by delaying the point of report from the time of experience or using other measures which do not rely on explicit report, such as eye-movements (see Frässle et al., 2014). No-report paradigms have already challenged previous interpretations of NCCs; such as the event-related potential component, the P3b, not being a reliable measure of consciousness (Dembski et al., 2021) and the reduction or absence of PFC involvement when stimuli are not perceived (Boly et al., 2017).

In contrast, frontal theories contend that the PFC (or at least some of its subregions) is necessary for generating consciousness. While no-report studies have highlighted that report may be a more robust correlate of PFC activation, frontal theorists argue that these findings do not yet rule out the PFC as a key brain region involved in the generation of consciousness (Boly et al., 2017; Odegaard et al., 2017). Indeed, it is true that "the absence of evidence is not evidence of absence" (Lau, 2022, p. 48), some frontal theories advocate for a nuanced role of the PFC cortex in consciousness. Higher order thought theories of consciousness hypothesise that the neural architecture involved in targeting sensory, first-order representations may not be that extensive and are instead quite subtle (Lau, 2019). They posit that the failure to detect an influence of PFC activity under conditions of awareness may simply be due to traditional imaging methods (e.g., fMRI, MEG) lacking the sensitivity to detect a subtle PFC activity (Michel & Morales, 2020; Morales & Lau, 2020). Additionally, it is well-established that the neurons in the PFC are highly functionally dispersed and it may well be that relatively few neurons support conscious perception; a position that might be difficult to corroborate through univariate analysis tools (Kapoor et al., 2018). However, it is also useful to note that GNWT predicts a more widespread, global ignition of activity from sensory areas to PFC, and may not be compatible with a subtle role of the PFC in consciousness. What is abundantly clear, however, is that the role of the PFC in consciousness is not well understood and demands further study.

In this study, we used computational modelling to reanalyse EEG data from an inattentional blindness (IB) no-report task first described by Shafto and Pitts (2015). IB is a particularly useful paradigm for consciousness researchers as the information entering the visual system is clearly visible and kept constant across participants, yet the IB stimulus only reaches awareness for some people while others fail to detect it (Mack, 2003; for a recent meta-analysis see Hutchinson et al., 2022). Thus, any differences in neural activity across

aware and unaware groups in an IB experiment should, in theory, be accounting for consciousness.

To adjudicate between the aforementioned theories, we used Dynamic Causal Modelling (DCM), a technique which estimates the effective connectivity between brain regions, or the directed influence that a brain region has on another (David et al., 2005). This allowed us to investigate the roles of key prefrontal and posterior brain regions under fluctuating conditions of awareness and task-relevance.

2.1 Methods

All data for this study have been kindly provided to us from Shafto and Pitts (2015). For detailed methods please refer to their original paper. However, to provide a brief overview of their experiment (as shown in Figure 2), their task consisted of a three-phase design, where task instructions and expectation were carefully manipulated to isolate task-relevance from awareness. Participants were shown two sets of stimuli – a jittering pattern of lines which, on occasion, would form a face (this was the critical inattentional blindness stimuli), and a set of green circles which formed a ring. Crucially, all trials across phases were kept identical with the only difference across phases being the task-instructions. In phase 1, participants were instructed to ignore the jittering background lines and only respond when one of the dots increased slightly in brightness. After the first phase, participants were given an awareness assessment where they were asked whether they noticed any patterns in the background stimuli. This awareness assessment determined that half of the participants were inattentionally blind (or unaware) of the face and the other half noticed (or were aware) of the face. As it was unclear when participants became aware of the face during phase 1, data from the aware participants were excluded from further analysis. After surprise questioning, in phase 2 expectancy to notice any patterns in the background stimuli increased while participants were instructed to continue to respond to the circle. Another awareness

assessment after phase 2 confirmed that all participants became aware of the face in the second phase. This meant that the face remained task-irrelevant, but participants were aware of it. Finally, in phase 3, participants were instructed to respond directly to the faces, and the face became task-relevant.

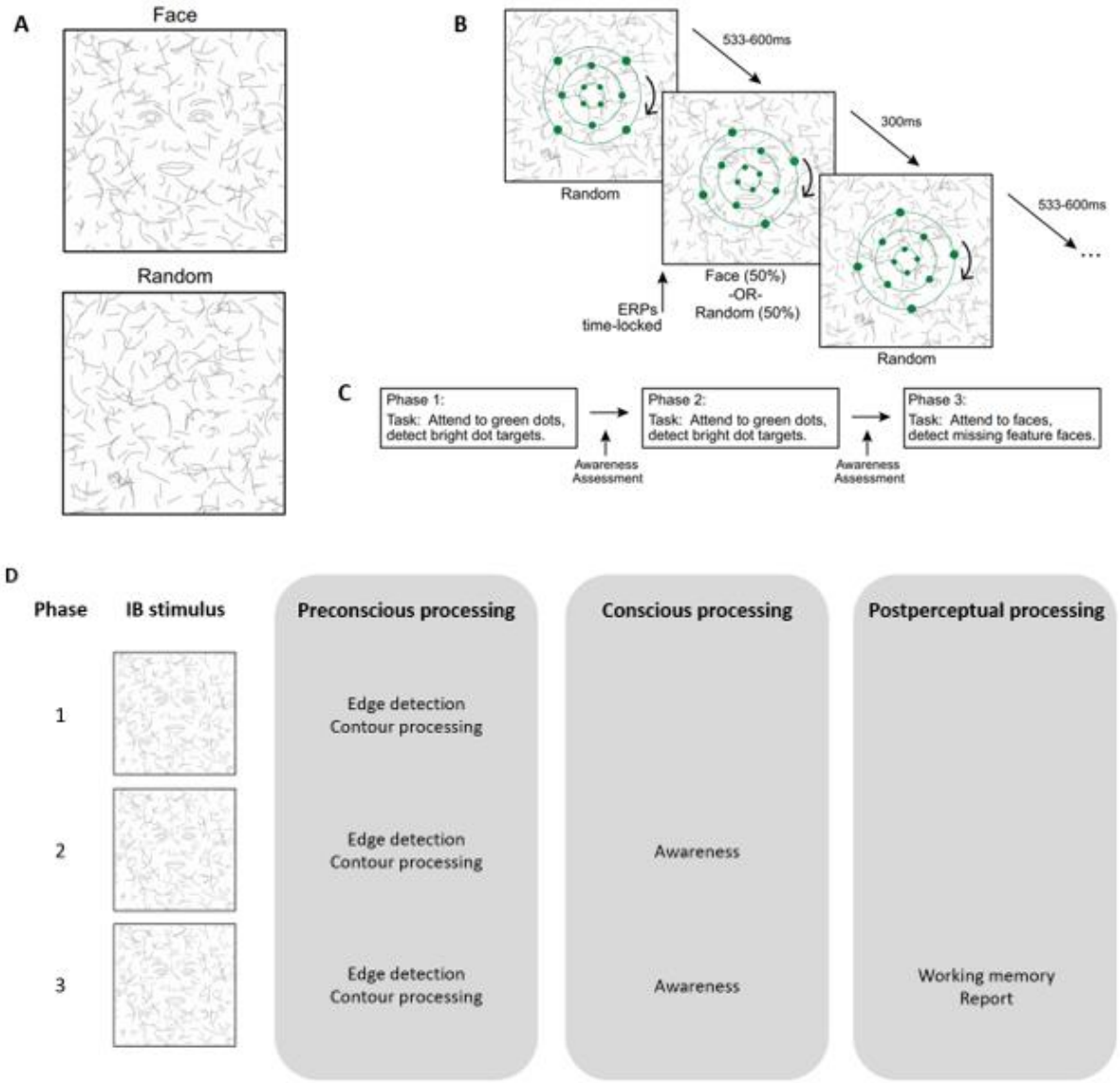

Figure 2. *Shafto and Pitts (2015) Inattentional Blindness Task.* An overview of the experimental setup in Shafto and Pitts (2015) as well as the key manipulations. A): The two types of background stimuli presented to participants in this task. Here, the face was the critical inattentional blindness stimuli. B): In this task, participants were instructed to respond to an increase in brightness in one of the dots along the green rings. The background lines jittered throughout each phase and participants were told to ignore these jittering lines. On occasion, the jittering lines aligned to form a face – the critical IB stimulus. C): The three-

phase design of the task where all trials across phases were kept identical. In phase 1, participants were asked to attend to the green dots, where half the participants were unaware and the other half were aware of the face. In phase 2, all participants became aware of the face but it remained task-irrelevant. Finally, in phase 3, participants were instructed to respond to the faces, and the face became task-relevant. D): Adapted from Pitts et al., 2014, this diagram walks through the different stages of conscious processing throughout the various phases of the task, assuming a participant is inattentionally blind. Preconscious in this sense is used to refer to processing that occurs immediately prior to the conscious representation. This information is supraliminal, in the sense that it *could* be consciously perceived, but it is not made available for report (Dehaene et al., 2006; Pitts et al., 2014).

2.2 EEG Preprocessing and Analysis

Data preprocessing was slightly different to the original Shafto and Pitts (2015) paper. SPM12 (https://www.fil.ion.ucl.ac.uk/spm/) was primarily used for preprocessing. Data was also preprocessed using a combination of EEGLAB (Delorme & Makeig, 2004; https://sccn.ucsd.edu/eeglab/index.php), and the Chronux software package (Mitra & Bokil, 2007; http://chronux.org/). The raw EEG data originally sampled at 500Hz was first converted to a readable format and then epoched from −100ms to 500ms relative to stimulus onset. Trials containing eye blinks were rejected. These were classified using VEOG channels. Trials containing artefacts, defined as activity greater than 100µV in any channel, and those containing button presses were also removed from analysis. Bad channels were defined as trials which contained these artefacts on more than 20% of trials, and these were also removed from analysis. This led to 16.20% of trials rejected in the unaware group and 11.83% rejected in the aware group. Robust averaging was then performed, and the data was subsequently low-pass filtered at 40 Hz, as robust averaging can introduce high frequency noise. Finally, the data was baseline corrected between -100ms and stimulus onset.

2.3 Source Reconstruction

To establish the nodes in the DCM models, we can take either a data-driven, or a priori, literature-based approach. As slightly different regions are theorized to be important across different theories of consciousness (even within the same family of theories) a data-driven approach was implemented to guide DCM model specification. Source reconstructed EEG images were used to identify regions of highest activation, which would provide the best chance of finding connectivity changes amongst the regions of interest, should they exist.

To test the hypotheses of frontal and sensory theories of consciousness, at least one PFC node and two sensory ROIs were needed in our network. The need for a PFC node is self-evident; the necessity of the PFC in consciousness is a key point of contention between these two families of theories and the key question of the current study. The two sensory ROIs were provided as a test for recurrent connectivity modulation during awareness, within sensory cortices (a prediction of sensory theories). Sensory nodes were also required as an input point for sensory information to enter the network. These considerations guided the model specification of the DCM analysis.

Source reconstruction on the preprocessed EEG files was conducted in SPM12 (https://www.fil.ion.ucl.ac.uk/spm/), which takes an empirical Bayesian approach to source reconstruction (Phillips et al. 2005). The default SPM template head model was aligned using the fiducial sensors at the nasion, and the left and right periauricular points. This model was first used to run a forward model which computed the effect of source activity on sensor activity. A boundary-element model cortical mesh was used for the forward model, as recommended by the SPM manual (Ashburner et al., 2014). To compute source estimates, this forward model was then inverted using a multiple sparse priors algorithm, optimised using a greedy search routine, which progressively searches for different combinations of source configurations until model evidence no longer improves (Friston et al., 2008). This

inversion was conducted using all trials across all conditions and with a time window from -100ms to 500ms relative to stimulus onset (corresponding to the trial epochs). No specific frequency band was of theoretic interest here, and thus a frequency band window covering the entire frequency range of the data, from 0 to 256 Hz, was chosen. The 3D source reconstructed images were smoothed at a resolution of $12mm^3$ and each image contained 8196 voxels.

The individual source reconstructed data was optimised via a group inversion scheme within SPM. This method takes the individual source estimates and adds an additional constraint to align activity across all participants, but allows the magnitude of activation for sources to vary. This yields better source estimates compared to individual inversion, as it is likely that the same sources are producing activity across all subjects as they are completing the same task (Litvak & Friston, 2008).

2.4 Dynamic Causal Modelling

DCM for evoked responses was used first to model how connection strengths modulate in a network when viewing face stimuli (using random stimuli as a baseline) in each subject per phase. The DCM was specified using the nodes obtained from the source-level analysis (see 3.2 Results for details). Here, random and face trials across all phases were inputted as a between-trial effect across a time-window from 1ms to 500ms. An equivalent current dipole model was used to model single dipoles for each source location, as shown in Figure 5. No specific DCM network architecture was specified. Instead, a fully connected network was created, with connections at all levels for forward, backward, intrinsic, and lateral connections. Input connections, however, were restricted to the left and right occipital nodes (i.e., it is well-established that visual information is processed initially in primary visual areas). The modulatory effect of face stimuli was modelled such that all connections were allowed to modulate. These DCMs were inverted per participant per each of the 3

phases (i.e. 3 DCMs per participant to compare phases where faces were either unaware or aware, and task-relevant or irrelevant).

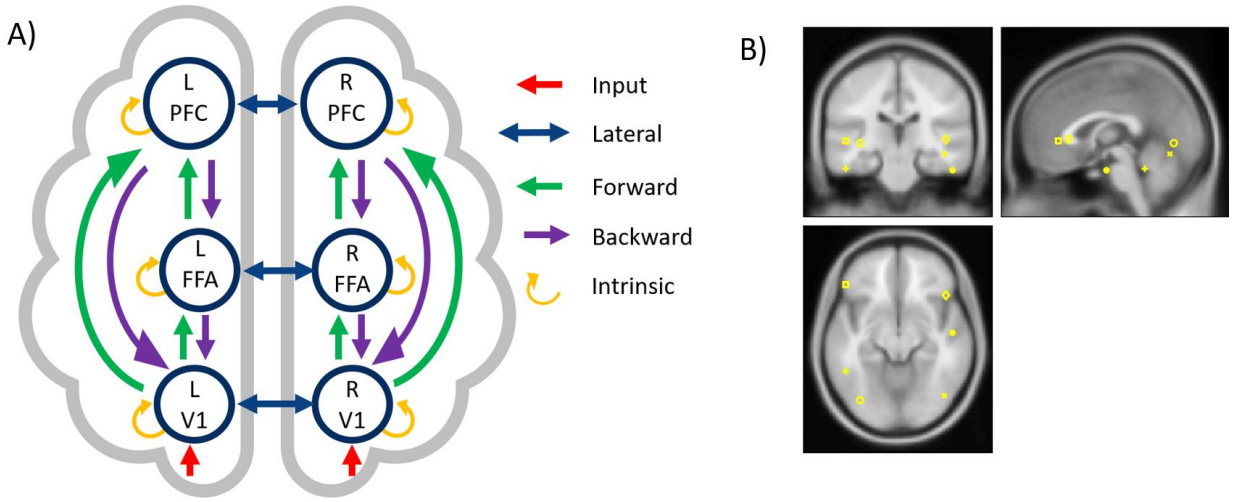

Figure 5. *The Full DCM Model Architecture and Dipoles.* A) The fully connected DCM architecture used in the DCM analysis. All nodes specified from the source reconstruction image analysis were used and connected at all levels of the architecture. The forward, backward, intrinsic, and lateral connections were free to be modulated by the effect of face. L and R refer to the left and right hemispheres; PFC: the prefrontal cortex node, here, Brodmann area 45 within the inferior frontal gyrus; FFA: fusiform face area; and, Occ: occipital area. B) The dipoles used as nodes in the DCM architecture displayed on a template head model.

2.5 Parametric Empirical Bayes

Parametric empirical Bayesian (PEB) modelling was used to examine how connection parameters in the network were modulated at the group-level across the main effects of awareness and task-relevance. This involved a second-order hierarchical Bayesian model which was conducted over the first-order DCM parameters (Friston et al., 2016). The two main effects of awareness and task-relevance were modelled to observe their corresponding changes in modulatory activity (i.e. the B-matrix in the DCMs) across the network. From

here, two types of PEB analyses were performed - an exploratory approach and a hypothesis-driven approach.

2.5.1 Exploratory automatic search PEB analysis

First, we took an exploratory approach (also called automatic search), estimating a fully connected model and then reducing redundant connections which did not contribute to model evidence via Bayesian model reduction (Friston et al., 2016; Zeidman et al., 2019). Here, three regressors were entered into the PEB model. The first regressor modelled the group mean (i.e. face vs. random stimuli), the second modelled how connectivity differed as a function of awareness of the face (0 for unaware and 1 for aware), and finally, the third regressor modelled the difference between task-relevance of the face (0 for task-irrelevant and 1 for task-relevant). For the awareness regressor, phase one was coded as 0 (recall that aware participants phase one data was excluded), while phase two and three was coded as 1. For the task-relevance regressor, phases one and two were coded as 0, and phase three was coded as 1. The full PEB was then estimated and reduced to prune redundant connections.

2.5.2 Hypothesis-driven PEB analysis

A second PEB analysis was also performed across a reduced hypothesis space to test specific predictions from either family of theories. Instead of an automatic search, specific parameters within the B-matrix were fixed at a prior mean of 0 (effectively, switching these parameters off). Differences in model evidence when specific connection parameters are switched-off reveal the importance of those connections (Zeidman et al., 2019).

Two candidate models were specified to test frontal and sensory theories, respectively. The interpretation of these models may be counterintuitive at first glance. That is, to determine support for sensory theories, connections from higher to lower sensory regions (VI and FFA) were ‘switched off’ (i.e. fixed at a prior mean of zero), and if this impacted the performance of the model greatly (as measured by model evidence), this would

suggest that those connections are important for awareness. On the other hand, to gather evidence for frontal theories, connections from the PFC to both V1 and the FFA were switched off. If such a model suffered greatly by this change in the network architecture, this would suggest that those connections are important for awareness.

As both families of theories agree with respect to how task-relevance would affect activity across the network (i.e. PFC should influence sensory regions), only the awareness main effect was analysed here. The same coding scheme used in the exploratory approach above was used here (i.e. 0 for unaware and 1 for aware). As shown in Figure 6c, two nested PEB models were specified which switched off backward connections from the PFC to sensory regions (testing key connections predicted by frontal theories) or backward connections from higher to lower sensory regions (testing key connections predicted by sensory theories). Model evidence and posterior probabilities of these template PEBs were then compared.

3.1 Results

Grand mean ERPs are shown, which through visual inspection, appeared consistent with the ERPs published in the original Shafto & Pitts (2015) paper, despite our slightly different preprocessing. Example ERPs are plotted in Figure 3.

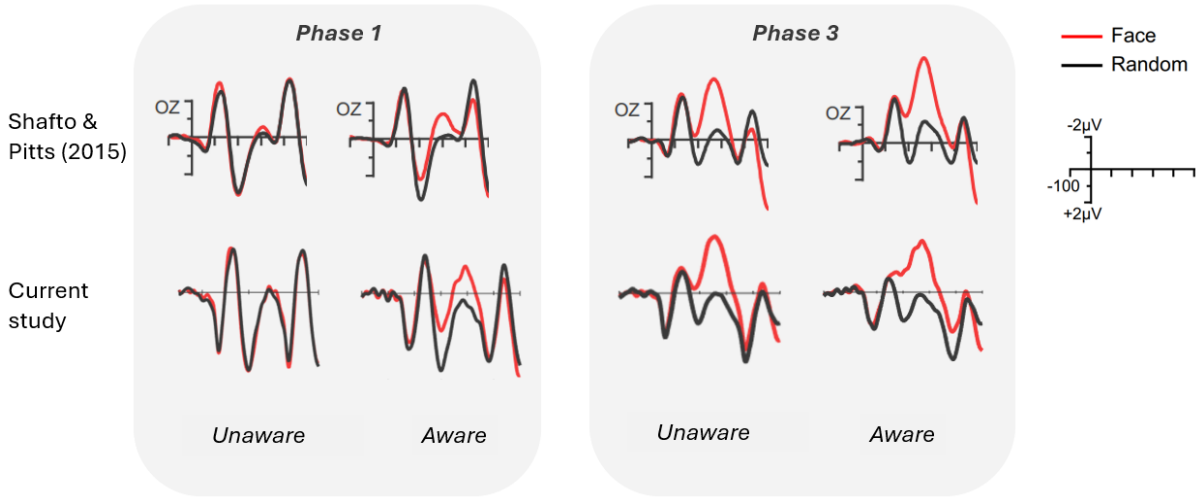

Figure 3. *Comparison of Selected ERPs in Current Study and Shafto and Pitts (2015).* Comparison of ERPs at electrode Oz in original study (Shafto & Pitts, 2015) with the ERPs computed in the current study after preprocessing. This shows the grand averaged ERPs for unaware and aware participants across phases 1 (when faces are task-irrelevant) and phase 3 (when faces are task-relevant) for face and random trials.

3.2 Source-level contrasts

A general linear model was used to determine regions of greatest activation under the experimental conditions. A full factorial model was conducted with a 2 (aware or unaware) x 2 (face or random) x 3 (phase) design. The contrast of interest was the interaction between face and random stimuli in phase three. Phase three was chosen because it is expected that the nodes underlying awareness and task-relevance will exhibit the greatest activity in this condition (where faces are both task-relevant and conscious). Face and random trials were chosen for this comparison to eliminate any common (or baseline) voxel activity. Voxel clusters with the greatest T-statistic were used as nodes in the construction of the DCM. Figure 4 shows the results of these source level contrasts.

Only four clusters with significant activation were found: the right occipital region ([44 -72 -12], peak-level T_{max} = 6.35, peak-level $p_{FWE\text{-}corr}$ <0.001), right fusiform region ([52 -12 -28], peak-level T_{max} = 6.2, peak -level $p_{FWE\text{-}corr}$ <0.001), left fusiform region ([50 -48 -26], peak-level T_{max} = 5.68, peak-level $p_{FWE\text{-}corr}$ <0.001) and left occipital region ([-36 -76 -2],

peak-level T_{max} = 5.66, peak-level $p_{FWE\text{-}corr}$ <0.001). No other significant clusters were found at this adjusted p-value threshold.

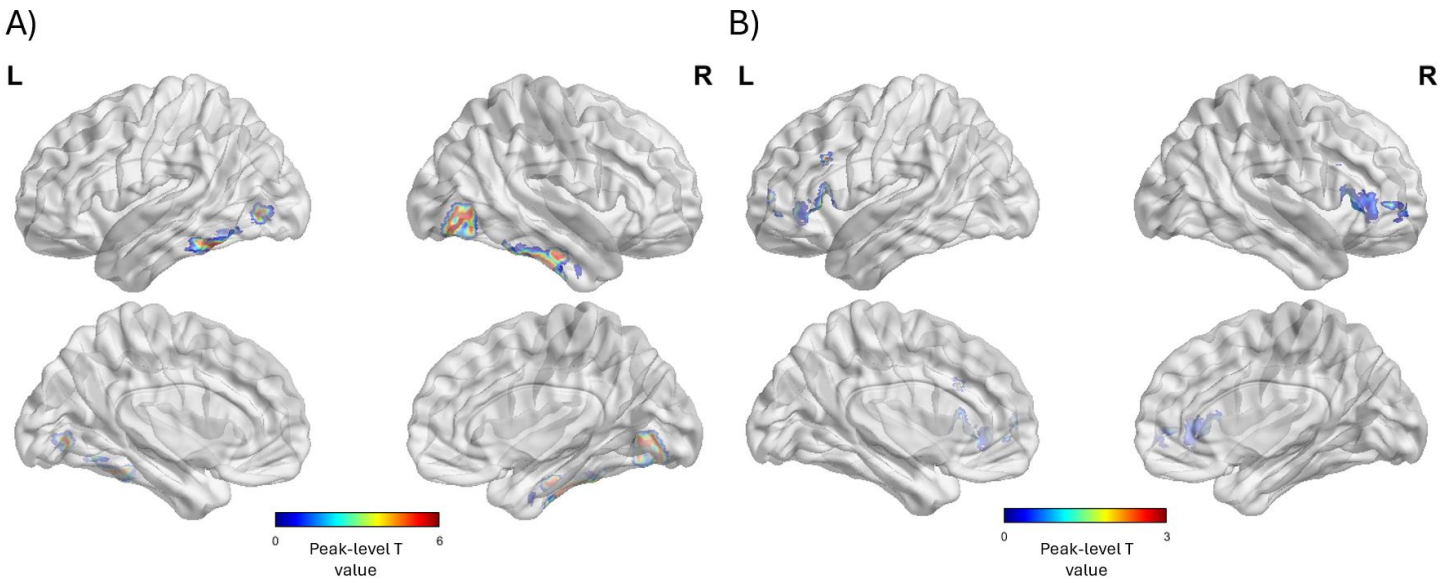

Figure 4. *Analysis of Source Reconstructed Images.* Significant regions of activation comparing face and random trials in phase three across all participants. A) Source-level contrasts for the interaction between face and random in phase three revealed two clusters across both hemispheres corresponding to occipital and fusiform areas respectively. A family-wise error (FWE) corrected threshold of $p < 0.05$ was used here. B) As no significant prefrontal clusters were found in the first analysis, a less conservative significance threshold of $p < 0.05$ uncorrected was used over an anatomical PFC mask. Peak activity revealed significant clusters at the left and right inferior frontal gyri. L and R denote left and right hemispheres, respectively.

In line with our hypotheses set, we identified significant clusters corresponding to one lower-order sensory region and one higher-order sensory-region, from either hemisphere. However, to establish a prefrontal ROI, the same full-factorial model was run but this time with an anatomical mask over the PFC and an uncorrected p-value threshold of p<0.05 was used. This PFC mask was created over the Brodmann areas 8, 9, 10, 11, 12, 13, 14, 24, 25, 32, 44, 45, 46, and 47 for left and right hemispheres, respectively (as defined by Fuster, 2015). The most significant prefrontal clusters from these analyses were the left inferior

frontal gyrus ([-50 34 0] peak-level T_{max} = 2.66, peak-level p_{unc} = 0.004) and the right inferior frontal gyrus ([46 24 2] peak-level T_{max} = 2.38, peak-level p_{unc} = 0.009).

3.3 DCM and PEB analysis

The PEB analysis revealed connectivity changes across the inverted DCMs as a function of awareness and task-relevance. To reiterate, two approaches were taken here - an exploratory automatic search to determine how connection strengths are modulated by awareness and task-relevance, and a hypothesis-driven search to examine whether a frontal or sensory model best explained the data.

The automatic search revealed how awareness and task-relevance modulated the strength of face-specific connections across the network and throughout the three phases. The estimated group-level connection strengths as a function of awareness and task-relevance is shown in Figure 6 for the winning model. For the main effect of awareness, connectivity increased from right fusiform to the right PFC. Though this activity was lateralised, this may provide partial support in line with the idea of global ignition from PFC to sensory areas postulated by a prominent frontal theory (GNWT; Mashour et al., 2020). We also saw decreased connectivity from the left PFC and left fusiform nodes to the left occipital node as a function of awareness (posterior probability > 0.75). For the main effect of task-relevancy, we found that connectivity from the right fusiform to the right PFC decreased, while intrinsic connections within the left and right occipital, left and right PFC, and right fusiform all increased (posterior probability > 0.75).

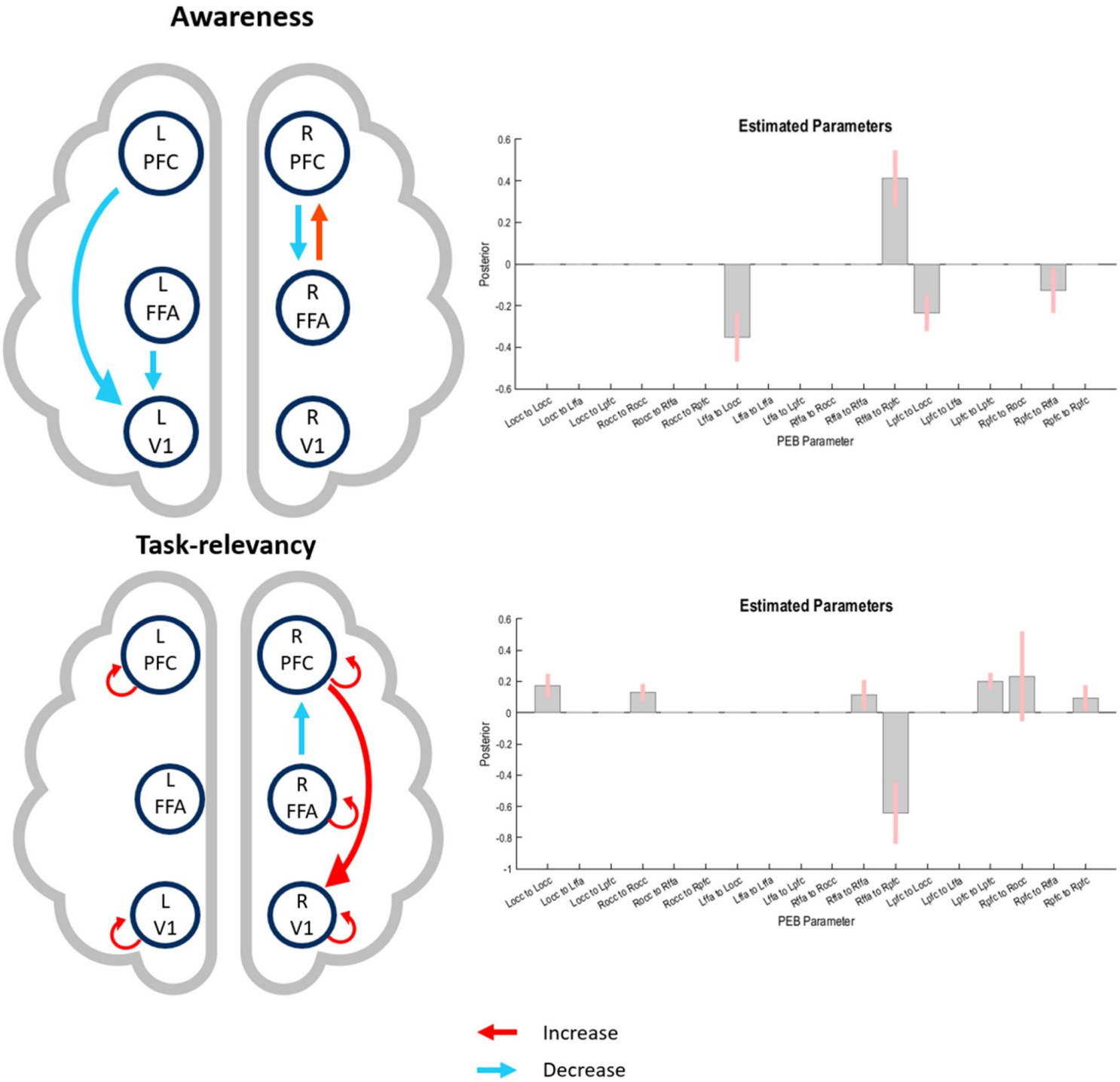

Figure 6. *Connection Strengths Across the Network Modulated by Awareness and Task-relevancy using automatic search PEB.* This figure shows the results of the winning model via the automated PEB search. The top half shows how the main effect of awareness modulated connectivity across the network. Here, awareness increased connectivity from the right-fusiform to the right-PFC node and decreased from the left-PFC and left-FFA ROIs to

the left-occipital node. The bottom half of the figure shows how task-relevance modulated connectivity parameters across the network. Connectivity from the right-fusiform to the right-PFC decreased, while intrinsic connections within the left- and right-occipital, left- and right-PFC all increased (all posterior probability>0.95). PFC refers to the prefrontal cortex (Brodmann area 45), FFA is the fusiform face area, and V1 is the primary visual cortex.

As these results did not provide clear support for predictions from either theoretical camp, a more directed hypothesis-driven approach was taken to constrain the model space to test specific hypotheses. When comparing two nested PEB models to explain modulatory changes in awareness, with either FFA to V1 connections switched off (to assess sensory theories) or PFC to both FFA and V1 connections switched off (to assess frontal theories), to, there was a slight preference for the model with FFA to V1 connections switched off with a posterior probability of 53% compared to 47% for the model with PFC backward connections switched off. Hence, slightly favouring the frontal theories, Crucially, to ensure that these backward connections were indeed important in contributing to model evidence, a null model (where all backward connections were switched off) was added into the model space. This null model performed poorly and had a negligible effect on the model probabilities, confirming the importance of these backward connections in the model.

4.1 Discussion

This study aimed to investigate the function of the PFC in visual consciousness. We conducted second-order modelling using parametric empirical Bayes on first-level DCMs to investigate how particular regions contribute to consciousness under conditions of awareness and unawareness. Through an exploratory search based on Bayesian model averaging, we found inconclusive results which sided with neither frontal nor sensory theories of consciousness, albeit a feedback connection from the right PFC to FFA provide partial

support for a frontal theory (as in GNWT; Mashour et al., 2020). Through a subsequent hypothesis-driven approach, we found evidence supporting both theories of consciousness, with a very slight preference for frontal theories (53% for frontal vs 47% for sensory).

The exploratory automatic search PEB results revealed that in the winning model, awareness of the face increased connectivity from the right fusiform to the right PFC nodes and decreased from the left PFC to the left occipital ROIs. These results do not align with either family of theories. Various sensory theories predict feedback connectivity would increase under awareness within higher and lower-sensory regions – however this was not observed. Secondly, though frontal theories predict the PFC would increase feedback connectivity to sensory areas, instead a decrease was observed. Having said that, an increase in connectivity was observed from the right fusiform to the right PFC, providing partial support for the idea of global ignition via the PFC as predicted by GNWT (Mashour et al., 2020), one of the frontal theories.

The results for task-relevancy were similarly unexpected. For the main effect of task-relevancy, we observed a decrease in connection strengths from the right fusiform to the right PFC, while intrinsic connection strengths for the right occipital, left PFC, and right PFC all increased. This was unexpected as both theories predict PFC connectivity to sensory regions should increase with task-relevancy.

As these results did not clearly align with predictions from either theoretical camp, a more directed hypothesis-driven approach was taken to limit the model space to test specific hypotheses. Two candidate models were specified to test frontal and sensory theories, and key connections were switched-off and model evidences was computed and compared. The results from this analysis revealed a slight preference for frontal theories, as the model which

did not allow backward connections from the PFC to modulate, had a relatively lower posterior probability.

At face value, these findings provide some support for frontal theories, albeit only slightly. Our results reveal that a model with key connections from the PFC to higher- and lower-sensory regions switched off, fare slightly worse compared to a model where only higher- to lower-sensory connections are switched off. This suggests the backward connections from the PFC may be important for awareness and the model suffers without them. This could suggest a subtle influence of the PFC in consciousness as predicted by frontal theories.

Thus far, we have discussed frontal theories as a family of theories; however, these theories do differ in how frontal areas precisely instantiate consciousness. Most notably, various proponents of higher-order thought theories advocate for a relatively subtle, yet essential, role of the PFC in consciousness (Michel & Morales, 2020; Morales & Lau, 2020; Odegaard et al., 2017). Higher-order thought theories propose that consciousness arises when higher-order regions (meta-) represent lower-order perceptual. These theories hypothesise that the neural architecture involved in targeting first-order representations may not be that extensive, and quite sparse (Lau, 2019). Accordingly, various findings showing no or little PFC activity under conditions of (non-reported) awareness are argued to be due to traditional imaging methods (e.g., fMRI, MEG) lacking the sensitivity to detect this subtle activity (Michel & Morales, 2020; Morales & Lau, 2020).

There are several studies which support the view that activity in the PFC is reduced, but not absent, during awareness in no-report tasks (Dellert et al. 2019; Frässle et al., 2014; Hatamimajoumerd et al., 2022; and using more sensitive measures in Noy et al., 2015; and Fang et al., 2023). Using a visual masking paradigm, a recent study by Hatamimajoumerd et

al. (2022) provided partial support for this view. They demonstrated that fMRI activity from frontal regions was modulated by the requirement to report on a visible stimulus; such that, frontal activation was elicited in report conditions but was absent in no-report conditions. However, when using more sensitive analysis techniques (i.e. multivoxel pattern analysis), a classifier could reliably decode conscious awareness from frontal regions, irrespective of the requirement to report. This could be suggestive of a subtle, albeit crucial, role of the PFC in consciousness, as predicted by some frontal theories. Nonetheless, this conclusion should be treated with caution as frontal areas were the only region (out of the regions studied) where decoding accuracy was significantly modulated by report. Accordingly, under no-report conditions, awareness may indeed be decodable, but to a lesser extent, due to a baseline level of executive processing involved in completing the task (Block, 2019).

The ever so slight preference for frontal theories found here, suggests the role of the PFC in consciousness may be more complex than previously thought. A recent study used Granger causality to investigate the role of the PFC in awareness using the same dataset analysed here (Rowe et al., 2024). Granger causality is similar to DCM as it examines connectivity in networks across the brain, but does so by determining whether region X is *G-causal* of Y, that is, the past of Y better explains X than the past of X itself (Seth et al., 2015). This study revealed that the PFC contained information about the current sensory environment or stimulus, irrespective of awareness and task-relevance of that very sensory information. Our exploratory PEB results partially align with these findings, where an increase in feedforward connectivity strength from the right FFA to PFC was observed for the effect of awareness. However, the lack of recurrent modulation in these findings suggests that the PFC may be required in the feedforward processing of sensory information only, and not in feedback processing. Corroborating these findings using practically identical methods to Shafto & Pitts (2015) and alongside concurrent EEG-fMRI, Dellert et al. (2021) found that

conscious face perception was associated with activation in the fusiform gyrus, left inferior lateral occipital cortex, and the right inferior frontal junction, within the PFC. The fact that these results which may not be consistent with different theories of consciousness but are consistent across conventional univariate imaging and computational approaches, and across different samples may highlight the need for a revision of long-held ideas within consciousness science.

Our approach is timely and coincides with a broader push in the field toward pitting theories of consciousness against each other in a systematic and unbiased way. By directly comparing key points of divergence between theories, this approach hopes to accelerate progress toward an understanding of consciousness (Melloni et al., 2021). This has been best exemplified by a recent large-scale adversarial collaboration (Cogitate Consortium, 2023). Importantly, given the (relatively) inconclusive results of that study, ours, and that of Rowe et al. (2024), we argue for a need to revise the current theories of consciousness, including re-examining precisely what the role of the PFC is in consciousness.

Acknowledgements. The authors would like to thank Dr. Peter Zeidman and Dr. Kit Melissa Larsen for their extremely helpful guidance during the data analysis process, specifically in relation to the DCM-PEB analysis. We would also like to thank Dr. Juliet Shafto and Dr. Michael Pitts for collecting the data and sharing this data with us for reanalysis. KB was supported by the Melbourne School of Psychological Sciences PhD Studentship and the Australian Government Research Training Program Scholarship.

References

Ashburner, J., Barnes, G., Chen, C. C., Daunizeau, J., Flandin, G., Friston, K., ... & Penny, W. (2014). SPM12 manual. *Wellcome Trust Centre for Neuroimaging, London, UK*, *2464*(4).

Block, N. (2019). What is wrong with the no-report paradigm and how to fix it. *Trends in Cognitive Sciences*, *23*(12), 1003–1013. https://doi.org/10.1016/j.tics.2019.10.001

Boly, M., Massimini, M., Tsuchiya, N., Postle, B. R., Koch, C., & Tononi, G. (2017). Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? Clinical and neuroimaging evidence. *Journal of Neuroscience*, *37*(40), 9603–9613. https://doi.org/10.1523/JNEUROSCI.3218-16.2017

Brascamp, J., Blake, R., & Knapen, T. (2015). Negligible fronto-parietal BOLD activity accompanying unreportable switches in bistable perception. *Nature Neuroscience*, *18*(11). https://doi.org/10.1038/nn.4130

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, *23*(9), 754–768. https://doi.org/10.1016/j.tics.2019.06.009

Cogitate Consortium, Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., ... & Melloni, L. (2023). An adversarial collaboration to critically evaluate theories of consciousness. *BioRxiv*, 2023-06.

Cohen, M. A., Ortego, K., Kyroudis, A., & Pitts, M. (2020). Distinguishing the neural correlates of perceptual awareness and postperceptual processing. *Journal of Neuroscience*, *40*(25), 4925-4935.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, *2*, 263–275.

David, O., Harrison, L., & Friston, K. J. (2005). Modelling event-related responses in the brain. *NeuroImage*, *25*(3), 756–770. https://doi.org/10.1016/j.neuroimage.2004.12.030

Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, *10*(5), 204–211. https://doi.org/10.1016/j.tics.2006.03.007

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79*(1), 1–37. https://doi.org/10.1016/S0010-0277(00)00123-2

Dellert, T., Müller-Bardorff, M., Schlossmacher, I., Pitts, M., Hofmann, D., Bruchmann, M., & Straube, T. (2021). Dissociating the neural correlates of consciousness and task relevance in face perception using simultaneous EEG-fMRI. *The Journal of Neuroscience*, *41*(37), 7864–7875. https://doi.org/10.1523/JNEUROSCI.2799-20.2021

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dembski, C., Koch, C., & Pitts, M. (2021). Perceptual awareness negativity: A physiological correlate of sensory consciousness. *Trends in Cognitive Sciences*, *25*(8), 660–670. https://doi.org/10.1016/j.tics.2021.05.009

Fang, Z., Dang, Y., Ling, Z., Han, Y., Zhao, H., Xu, X., & Zhang, M. (2023). The involvement of the human prefrontal cortex in the emergence of visual awareness. *eLife*, *12*. https://doi.org/10.7554/eLife.89076.2

Förster, J., Koivisto, M., & Revonsuo, A. (2020). ERP and MEG correlates of visual consciousness: The second decade. *Consciousness and Cognition*, *80*, 102917. https://doi.org/10.1016/j.concog.2020.102917

Frässle, S., Sommer, J., Jansen, A., Naber, M., & Einhäuser, W. (2014). Binocular rivalry: Frontal activity relates to introspection and action but not to perception. *Journal of Neuroscience*, *34*(5), 1738–1747. https://doi.org/10.1523/JNEUROSCI.4403-13.2014

Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, *2*(1–2), 56–78. https://doi.org/10.1002/hbm.460020107

Friston, K., Harrison, L., Daunizeau, J., Kiebel, S., Phillips, C., Trujillo-Barreto, N., ... & Mattout, J. (2008). Multiple sparse priors for the M/EEG inverse problem. *NeuroImage*, *39*(3), 1104-1120.

Friston, K. J., Litvak, V., Oswal, A., Razi, A., Stephan, K. E., Van Wijk, B. C., ... & Zeidman, P. (2016). Bayesian model reduction and empirical Bayes for group (DCM) studies. *NeuroImage*, *128*, 413-431.

Friston, K., Moran, R., & Seth, A. K. (2013). Analysing connectivity with Granger causality and dynamic causal modelling. *Current Opinion in Neurobiology*, *23*(2), 172–178. https://doi.org/10.1016/j.conb.2012.11.010

Fuster, J. (2015). *The prefrontal cortex*. Academic press.

Hatamimajoumerd, E., Ratan Murty, N. A., Pitts, M., & Cohen, M. A. (2022). Decoding perceptual awareness across the brain with a no-report fMRI masking paradigm. *Current Biology*. https://doi.org/10.1016/j.cub.2022.07.068

Hutchinson, B. T., Pammer, K., Bandara, K., & Jack, B. N. (2022). A tale of two theories: A meta-analysis of the attention set and load theories of inattentional blindness. *Psychological Bulletin*, *148*(5-6), 370.

Kapoor, V., Besserve, M., Logothetis, N. K., & Panagiotaropoulos, T. I. (2018). Parallel and functionally segregated processing of task phase and conscious content in the prefrontal cortex. *Communications Biology*, *1*(1), 215.

Kiebel, S. J., Garrido, M. I., Moran, R. J., & Friston, K. J. (2008). Dynamic causal modelling for EEG and MEG. *Cognitive Neurodynamics*, *2*(2), 121. https://doi.org/10.1007/s11571-008-9038-0

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, *17*(5), 307-321.

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494–501. https://doi.org/10.1016/j.tics.2006.09.001

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571-579.

Lau, H. (2019). *Consciousness, Metacognition, & Perceptual Reality Monitoring*. PsyArXiv. https://doi.org/10.31234/osf.io/ckbyf

Lau, H. (2022). *In Consciousness We Trust: The Cognitive Neuroscience of Subjective Experience*. Oxford University Press.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15*(8), 365–373. https://doi.org/10.1016/j.tics.2011.05.009

Litvak, V., & Friston, K. (2008). Electromagnetic source reconstruction for group studies. *NeuroImage*, *42*(4), 1490–1498. https://doi.org/10.1016/j.neuroimage.2008.06.022

Mack, A. (2003). Inattentional blindness: Looking without seeing. *Current Directions in Psychological Science*, *12*(5), 180–184. https://doi.org/10.1111/1467-8721.01256

Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron*, *105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

Melloni, L., Mudrik, L., Pitts, M., & Koch, C. (2021). Making the hard problem of consciousness easier. *Science*, *372*(6545), 911–912. https://doi.org/10.1126/science.abj3259

Michel, M., & Morales, J. (2020). Minority reports: Consciousness and the prefrontal cortex. *Mind & Language*, *35*(4), 493–513. https://doi.org/10.1111/mila.12264

Mitra, P., & Bokil, H. (2007). *Observed Brain Dynamics*. Oxford University Press, USA.

Morales, J., & Lau, H. (2020). The Neural Correlates of Consciousness. In U. Kriegel (Ed.), *The Oxford Handbook of the Philosophy of Consciousness* (pp. 233–260). Oxford University Press.

Naccache, L. (2018). Why and how access consciousness can account for phenomenal consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170357. https://doi.org/10.1098/rstb.2017.0357

Noy, N., Bickel, S., Zion-Golumbic, E., Harel, M., Golan, T., Davidesco, I., ... & Malach, R. (2015). Ignition's glow: Ultra-fast spread of global cortical activity accompanying local "ignitions" in visual cortex during conscious visual perception. *Consciousness and Cognition*, *35*, 206-224.

Odegaard, B., Knight, R. T., & Lau, H. (2017). Should a few null findings falsify prefrontal theories of conscious perception? *Journal of Neuroscience*, *37*(40), 9593–9602. https://doi.org/10.1523/JNEUROSCI.3217-16.2017

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Computational Biology*, *10*(5), e1003588.

Phillips, C., Mattout, J., Rugg, M. D., Maquet, P., & Friston, K. J. (2005). An empirical Bayesian solution to the source reconstruction problem in EEG. *NeuroImage*, *24*(4), 997-1011.

Pitts, M. A., Lutsyshyna, L. A., & Hillyard, S. A. (2018). The relationship between attention and consciousness: An expanded taxonomy and implications for 'no-report' paradigms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170348. https://doi.org/10.1098/rstb.2017.0348

Pitts, M., Metzler, S., & Hillyard, S. (2014). Isolating neural correlates of conscious perception from neural correlates of reporting one's perception. *Frontiers in Psychology*, *5*, 1078. https://doi.org/10.3389/fpsyg.2014.01078

Rowe, E. G., Garrido, M. I., & Tsuchiya, N. (2024). Feedforward connectivity patterns from visual areas to the front of the brain contain information about sensory stimuli regardless of awareness or report. *Cortex*, *172*, 284-300.

Seth, A. K., Barrett, A. B., & Barnett, L. (2015). Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, *35*(8), 3293–3297. https://doi.org/10.1523/JNEUROSCI.4399-14.2015

Shafto, J. P., & Pitts, M. A. (2015). Neural signatures of conscious face perception in an inattentional blindness paradigm. *Journal of Neuroscience*, *35*(31), 10940-10948.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450–461.

Tsuchiya, N., Wilke, M., Frässle, S., & Lamme, V. A. F. (2015). No-report paradigms: Extracting the true neural correlates of consciousness. *Trends in Cognitive Sciences*, *19*(12), 757–770. https://doi.org/10.1016/j.tics.2015.10.002

Zeidman, P., Jafarian, A., Seghier, M. L., Litvak, V., Cagnan, H., Price, C. J., & Friston, K. J. (2019). A guide to group effective connectivity analysis, part 2: Second level analysis with PEB. *Neuroimage*, *200*, 12-25.