Current Research in Neurobiology 3 (2022) 100036

Contents lists available at ScienceDirect

Current Research in Neurobiology

journal homepage: www.sciencedirect.com/journal/current-research-in-neurobiology

An active inference model of conscious access: How cognitive action selection reconciles the results of report and no-report paradigms

Christopher J. Whyte [a,*], Jakob Hohwy [b], Ryan Smith [c]

[a] MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, UK
[b] Centre for Consciousness & Contemplative Studies, Monash University, Melbourne, Australia
[c] Laureate Institute for Brain Research, Tulsa, OK, USA

ARTICLE INFO

Keywords:
Active inference
Conscious access
Consciousness
Prefrontal cortex
No-report paradigms

ABSTRACT

Cognitive theories of consciousness, such as global workspace theory and higher-order theories, posit that frontoparietal circuits play a crucial role in conscious access. However, recent studies using no-report paradigms have posed a challenge to cognitive theories by demonstrating conscious accessibility in the apparent absence of prefrontal cortex (PFC) activation. To address this challenge, this paper presents a computational model of conscious access, based upon active inference, that treats working memory gating as a cognitive action. We simulate a visual masking task and show that late P3b-like event-related potentials (ERPs), and increased PFC activity, are induced by the working memory demands of self-report generation. When reporting demands are removed, these late ERPs vanish and PFC activity is reduced. These results therefore reproduce, and potentially explain, results from no-report paradigms. However, even without reporting demands, our model shows that simulated PFC activity on visible stimulus trials still crosses the threshold for reportability – maintaining the link between PFC and conscious access. Therefore, our simulations show that evidence provided by no-report paradigms does not necessarily contradict cognitive theories of consciousness.

1. Introduction

The neuroscience of consciousness has made considerable progress within the past two decades, with growing consensus in a number of areas. For example, almost all major theories of visual consciousness agree that recurrent activity is a necessary element (Lamme, 2006; Mashour et al., 2020; Tononi et al., 2016). Nevertheless, there is still substantial disagreement. Perceptual theories (e.g., local recurrency theory (Lamme, 2006) and integrated information theory (Tononi et al., 2016)) both argue that recurrent activity in posterior cortical regions is (in some sense) sufficient for consciousness. In contrast, cognitive theories, such as the global neuronal workspace (GNW; Dehaene and Changeux, 2011; Mashour et al., 2020) theory and higher order theories (HOT; Brown et al., 2019), argue that consciousness requires perceptual information to be accessed by, or at least be accessible to, structures within prefrontal cortex (PFC).

In support of cognitive theories, consciousness robustly correlates with both PFC activity and late event-related potentials (ERPs; i.e., indicative of post-perceptual cognitive processing) when measured via subjective report. For example, a meta-analysis of 19 bistable perception and phenomenal masking studies (Bisenius et al., 2015) found that a network of regions in extrastriate, temporal, prefrontal, and parietal cortices showed above-chance activation when contrasting conscious (seen) vs. unconscious (unseen) conditions. Likewise, in both masking and attentional blink paradigms, early ERP components have typically not displayed substantial differences between seen and unseen conditions, while the late P3b component has been found to show a large increase in amplitude during seen vs. unseen condition (Salti et al., 2012; Sergent et al., 2005; Ye et al., 2019).

More recently, however, these results have come under considerable scrutiny. Specifically, no-report paradigms, which do not require the collection of explicit subjective reports, have shown that prefrontal cortex activity is greatly reduced, and late ERPs vanish, when reporting demands are removed (Tsuchiya et al., 2015). In a no-report variant of a binocularly rivalry paradigm, Frassle and colleagues (Frassle et al., 2014) found that, when measured via reports, perceptual transitions were associated with the activation of a network of regions within the superior parietal cortex and bilateral middle frontal gyrus. However, when reporting demands were removed, the activity in bilateral middle frontal gyrus dropped below significance. Using large-scale intracranial

* Corresponding author. MRC Cognition and Brain Sciences Unit, 15 Chaucer Rd, Cambridge, CB2 7EF, UK.
E-mail address: christopherjackwhyte@gmail.com (C.J. Whyte).

https://doi.org/10.1016/j.crneur.2022.100036
Received 11 January 2022; Received in revised form 8 March 2022; Accepted 18 March 2022
Available online 1 April 2022

recordings (ECoG), Noy and colleagues (Noy et al., 2015) presented participants with a stream of images (all above threshold for conscious perception). In one condition participants also performed a one-back working memory task, while the other condition only required them to passively view the stimuli. They found that the magnitude of frontoparietal activity was substantially reduced in the passive perception condition. Similarly, Pitts and colleagues (Pitts et al., 2014) used a no-report variant of an inattentional blindness paradigm (i.e., where reports were gathered after blocks of trials as opposed to on each trial) and found that the P3b was driven by task relevance, and not perceptual awareness. This result has since been conceptually replicated and extended in a standard phenomenal masking paradigm by Cohen and colleagues (Cohen et al., 2020). Despite participants not reporting their experience in the no-report condition, Cohen and colleagues found that they still displayed a remarkable degree of accuracy on an incidental memory task for unmasked trials. This was not the case for trials with masked stimuli, reducing the plausibility of the objection that the unmasked stimuli may not have been consciously seen on no-report trials. This set of results has been interpreted by many as strong evidence in favour of perceptual theories of consciousness. It is suggested that PFC may modulate consciousness, but that it does not play a necessary role. In other words, posterior cortical processing is argued to be sufficient for experience on its own (Tsuchiya et al., 2015).

Despite these challenges to cognitive theories, new evidence from invasive neurophysiology in non-human primates has shown that, even in the absence of report, the contents of consciousness can be reliably decoded from PFC (Kapoor et al., 2020). Further, state fluctuations in PFC appear to precede perceptual transitions, suggesting a causal role for PFC in consciousness (Dwarakanath et al., 2020). Although not within a no-report paradigm, a very similar result was reported in a cohort of neurosurgical patients implanted with intracranial electrodes. Specifically, Gelbard-Sagiv et al. (2018) found that medial frontal activity preceded internally-driven perceptual switches. More recently still, in an auditory no-report paradigm conducted with human participants, Sergent and colleagues (Sergent et al., 2021) found that, even in the absence of report, conscious access correlated with a bifurcation in neural activity (recorded via EEG) reminiscent of the type of non-linear ignition predicted by GNW theory. Crucially, when they localised the EEG signal, conscious access was associated with activity in inferior PFC. Given that PFC activity seems to play a crucial role in consciousness in these studies, but is attenuated in no-report paradigms, it remains an open and important question why reporting demands induce late ERPs and enhance PFC activity. In other words, it remains unclear what the difference in *cognitive* processing is between report and no-report conditions that explains the dramatic change in the neural correlates of consciousness.

Here we present a computational model of visual conscious access, based upon active inference, that attempts to answer this question. In previous work (Whyte and Smith, 2021), we used a two-level partially observable Markov decision process (POMDP) to model the behavioural and electrophysiological correlates of visual consciousness in minimal contrast paradigms, based on the domain-general neural process theory accompanying active inference (Da Costa et al., 2020; Friston et al., 2017; Sajid et al., 2021; Smith et al., 2022). Specifically, we showed how this two-level POMDP could reproduce the dissociation between the P3b and conscious access. Although this work aimed to explain existing results, the model also generated further predictions about the interaction between attention and expectation that have subsequently been confirmed in an extension of the inattentional blindness paradigm, lending some credence to the validity of our explanation. Specifically, we predicted that valid expectations would reduce the amplitude of the P3b, but only when the stimulus was task-relevant, which is exactly what has subsequently been found (Schlossmacher et al., 2020). In this previous work, we modelled the dissociation of the P3b and conscious access as a result of the manipulation of attention. However, this explanation does not straightforwardly generalise to other no-report paradigms.

This paper presents an extension of our previous model to more general no-report paradigms by treating working memory as a type of cognitive action (Limanowski and Friston, 2018) – conceptually similar to other models positing action-like neural mechanisms for selectively gating contents into working memory (e.g., involving PFC-basal ganglia loops; O'Reilly and Frank, 2006) – but applied specifically to subjective reports. This is motivated by the fact that the frontoparietal networks appealed to by cognitive theories are known to overlap with selective attention and working memory processes (Rottschy et al., 2012), and by the fact that these processes have been explicitly appealed to by previous cognitive theorists as integral to conscious access (Prinz, 2012).

Building on this work, we use our model to simulate visual masking and show that the working memory demands associated with the generation of reports induce late P3b-like ERPs and enhance simulated PFC activity. In contrast, late ERPs vanish, and PFC activity is greatly reduced, when reporting demands are removed. Crucially, however, even when the model is not 'asked' to provide a subjective report (and therefore does not choose to maintain information in working memory), PFC activity on visible trials still rises above the threshold for reliable reportability immediately after stimulus presentation. We leverage this result to argue for an access-based account in which conscious perception is a matter of having sufficiently precise posterior beliefs to influence temporally deep policy selection (i.e., the selection of temporally extended action sequences, such as those involved in reporting). Specifically, our results suggest that if a participant *had* been asked to report their visual experience immediately after stimulus presentation, they *would have* reported seeing the stimulus, and that this is still a function of PFC activity in our model – associated with precise posterior beliefs at (temporally) deep levels of processing.

The paper is structured as follows. Section 2 provides a brief primer on active inference and the POMDP approach to modelling cognitive and perceptual processes. Section 3 outlines the structure of the specific generative model we employ (i.e., a model of how hidden causes outside of the brain generate sensory observations). Section 4 presents simulations of phenomenal masking *with* and *without* reporting demands within this generative model, showing its ability to reproduce the neural correlates of consciousness under both report and no-report conditions. We conclude in section 5 by highlighting the empirical predictions of the model. We also explore the implications of our results for the debate between cognitive and perceptual theories of consciousness and discuss the relationship between our model and prominent cognitive theories of consciousness.

2. Methods

2.1. A primer on active inference

Active inference is a framework for modelling (approximately) Bayes optimal behaviour under an internal (generative) model of the world (for a gentle but detailed introduction, see Smith et al., 2022). The POMDP (partially observable Markov decision process) formulation of active inference uses generative models that incorporate perception, learning, and decision-making under uncertainty. These models assume discrete state and outcome spaces and discrete time steps. Inference within these models is performed through the application of message passing algorithms – most commonly marginal message passing (Parr et al., 2019). This exemplifies the broader approach of minimizing variational free energy (VFE) that active inference models leverage more generally. VFE is a tractable approximation to the difference between a generative model and the true states of the world generating observed outcomes (i.e., the generative process). Minimizing VFE allows the perception-action cycle to be cast as an optimisation problem, where perception corresponds to the process of inferring the hidden states that maximise the probability of observations (while also minimizing model complexity; i.e., keeping the inferred explanations of observations as

simple as possible). In turn, action selection corresponds to the process of inferring the action sequences (i.e., policies) that will both minimise uncertainty and bring about preferred outcomes, formally cast as observations with high prior probabilities in the model (so-called prior preferences see Friston et al., 2016, 2017). Inference about optimal policies technically requires minimisation of the expected free energy (EFE) of future observations under each possible action sequence. Over the timescale of a single trial in a task, belief updates correspond to inference (i.e., updating beliefs about hidden states), while, over longer timescales, belief updating gives rise to *learning* (i.e., updating beliefs about the parameters of the generative model in order to increase predictive accuracy).

The inference procedure in these models is based upon a likelihood mapping between a set of factorised hidden states (s_τ) and observations (o_τ) at each timepoint τ, which encodes the probability of each observation given each possible hidden state, $p(o_\tau|s_\tau)$. Inference further depends on a model of the transitions between hidden states that would occur under each possible course of action or policy (π), $p(s_{\tau+1}|s_\tau,\pi)$. The likelihood mapping is specified by a set of matrices, denoted in Fig. 1 by the letter **A**, each of which describes the mapping between all hidden state factors and a distinct outcome modality. As mentioned above, to allow for interactions between hidden states and the observations they generate, they are factorised into distinct sets of outcome modalities, each with their own **A** matrix. The transition probabilities are encoded by a set of matrices denoted by the letter **B** (at least one matrix per hidden state factor), which describe the probability of a current state

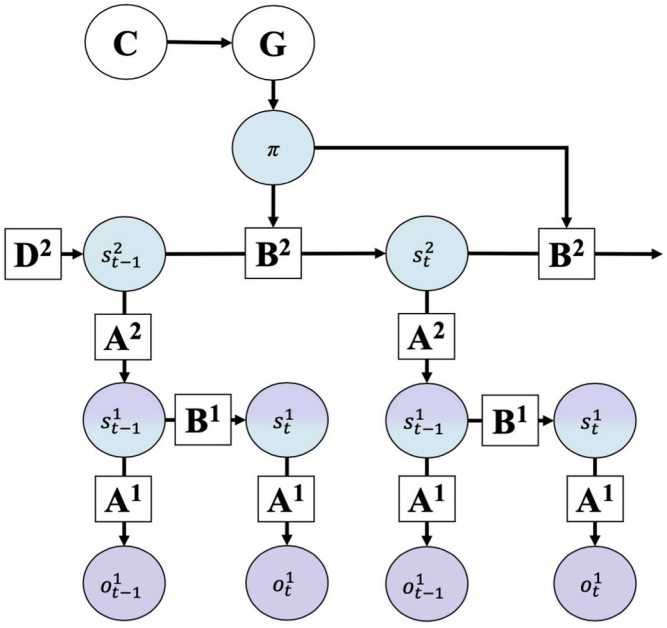

Fig. 1. A graphical depiction of a hierarchical POMDP. Arrows show the dependencies between variables (circles). Observations (purple) depend on hidden states (green), and state transitions are (partially) determined by policies (π). This representation also highlights the role of the vectors/matrices (squares) in determining the conditional dependencies between variables. Observations are generated by hidden states described by the matrix **A.** The **B** matrix determines state transitions, which function as empirical priors. The **D** vector serves as the prior for initial states. When the **B** matrix is under the control of the agent, these state transitions (actions) depend upon the policy. The probability that a particular policy will be selected depends on the expected free energy (**G**) of the policy, which is (partially) dependent on the prior preferences specified by **C**. In hierarchical models such as this, beliefs at the second level provide priors for inference at the first level. In turn, posterior beliefs over hidden states at the first level function as observations for the second level (represented by the shading from green to purple). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

conditional on a choice of policy, past states, and expected future states. In active inference, actions are modelled as the direct control of state transitions by the agent. Each agent is equipped with a policy space that describes a set of allowable action sequences, where each allowable transition (i.e., action) at each time point is assigned a distinct **B** matrix. An agent's prior preferences for particular observations, $p(o_\tau|C)$, are described by a set of matrices denoted by the letter **C** (one per outcome modality), which quantify the degree to which agents prefer, or are averse to, particular observations at each time point. Finally, prior beliefs about initial states, $p(s_{\tau=1})$, are determined by a set of vectors denoted by the letter **D** (one per hidden state factor), which function analogously to the **B** matrices but for the hidden states at the first time point. The probability distributions encoded in **A**, **B**, **C** and **D** are each categorical distributions with Dirichlet priors. For details on the free energy functionals and update equations, we refer readers to **Appendix 1**; for more in-depth walkthroughs, see (Da Costa et al., 2020; Sajid et al., 2021; Smith et al., 2022).

In hierarchical models, such as the model described in the following section, hidden states at the first level serve as observations at the second level (see Fig. 1). This affords inferences about deep temporal structure. For example, as shown in previous simulations of reading, the first level of a model can be used to infer single words from visual input, while the second level can be used to infer the narrative meaning entailed by sequences of words – where the latter process requires evidence to be accumulated over longer temporal scales (Friston et al., 2017).

A key feature of active inference is that it comes equipped with a detailed, domain-general neural process theory describing a proposed mapping between neurobiology and the equations governing belief updating (see Fig. 2 for a description of the relevant equations). Here we focus exclusively on the elements of the theory that map the update equations to electrophysiology. In this theory, posterior beliefs over each hidden state are mapped to firing rates in distinct neuronal populations. The average membrane potential of these neuronal populations, which controls their respective firing rates, is based on a depolarisation variable (v). The value of v is updated (denoted by '←') based on a state prediction error term (ε): $v \leftarrow v + \varepsilon$ (the equation for this state prediction error is shown in Fig. 2). This update entails that more surprising changes in beliefs about states (i.e., after a new observation) will generate greater changes in v. Depolarisation levels then update firing rates by applying a softmax (normalised exponential; σ) function to beliefs over states, $s \leftarrow \sigma(v)$, which transforms v into a probability distribution. The use of the softmax function (which is simply a generalisation of the sigmoid logistic function to vector inputs) to simulate average firing rates is based on the assumption made in mean-field models of neural dynamics that the average firing rate of a population can be treated as a sigmoid function of the average membrane potential (Breakspear, 2017; Da Costa et al., 2021). ERPs and local field potentials are modelled as the time derivative (i.e., rate of change) of the normalised firing rate. It is worth highlighting the face validity of this setup. Because the depolarisation variable is not normalised, it can take both positive and negative values, like voltage; in contrast, after being normalised by the softmax function, it is bounded between zero and one, like a normalised firing rate.

2.2. A generative model of conscious access

To model conscious access, both with and without subjective report, we simulated a phenomenal masking paradigm with a delay period between the presentation of the stimulus and the collection of a subjective report in the report condition, and an extended ISI in the no-report condition. We introduced the delay period into the otherwise typical phenomenal masking paradigm to explicitly link the neural consequences of increased working memory demands to the functional pay offs facilitated by goal-directed working memory maintenance (i.e., the elevated neural activity associated with choosing to maintain a stimulus in working memory in order to successfully complete a task).

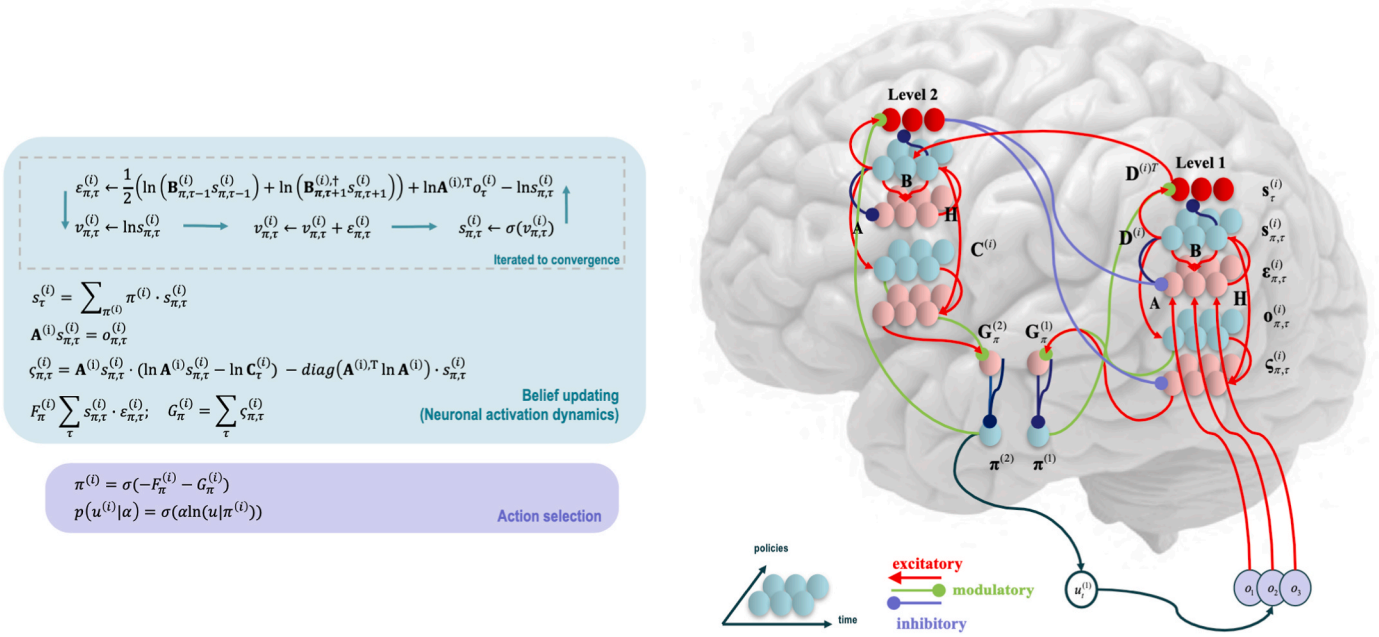

Fig. 2. Neural process theory associated with active inference. The **left** portion of this figure shows the update equations and free energy functionals. Heuristically, state prediction errors $\boldsymbol{\varepsilon}$ score the (log) difference between the generative model and the approximate posterior after receiving an observation. Outcome prediction errors $\boldsymbol{\varsigma}$ encode beliefs about the value of each policy (i.e., higher outcome prediction errors for a given policy roughly correspond to lower probabilities of observing preferred outcomes under that policy, as well as less informative observations expected under that policy). Further below are expressions for marginal free energy **F** and expected free energy **G**, expressed in terms of the above-mentioned (state and outcome) prediction errors. Also shown are the update equations for states, policies, Bayesian model averages for states weighted by policies ($s_\tau^{(i)}$), and the depolarisation variable (v), as well as selection of actions ($u_t^{(i)}$). Subscripts denote dependence on policies (π) and time (τ). Superscripts (i) denote hierarchical level. The **right** panel provides a schematic of message passing between cell populations that could potentially implement these updates. Red units encode Bayesian model averages, cyan units encode expectations over hidden states, and pink units encode state and outcome prediction errors. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

This aspect of the task was inspired by previous unconscious working memory paradigms (King et al., 2016; Soto et al., 2011; see Fig. 3). We also note that, even when asked to report experience directly after stimulus presentation, it is plausible to assume some (if minimal) working memory demand associated with holding that prior percept in mind to construct a report. As such, our results below do not depend on the presence of extended delay periods associated with traditional working memory tasks. They only depend on the assumption that, in relevant task trials, agents choose to engage in the (relatively) more effortful cognitive processes associated with generating reports after stimulus presentation.

Our simulated task began by presenting the agent with a blank screen. At the second timepoint, the agent was presented with either a left- or right-oriented Gabor patch (the target stimulus) or with another blank screen. In the report condition, the agent was required to maintain the stimulus over a delay and to either report whether they had seen a

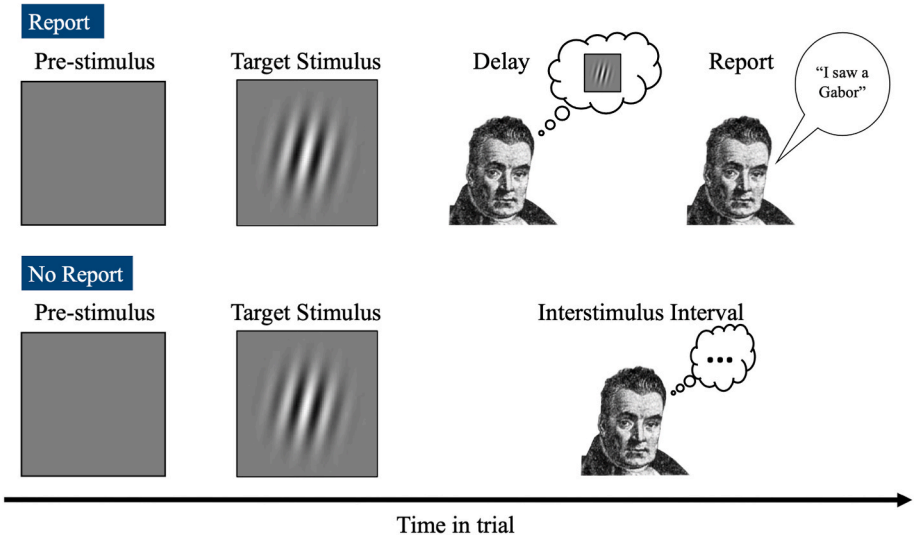

Fig. 3. Depiction of the simulated task. In the report condition, the agent was presented with a Gabor patch, followed by a delay and then a subsequent report phase. In the no-report condition, the Gabor patch was presented in identical fashion, but the agent did not anticipate the need to report its experience and therefore did not maintain the percept in working memory.

Gabor patch or perform a two-alternative forced-choice task. In the no-report condition, the agent had no task demands and was only required to passively perceive the stimulus (see Fig. 3). Inspired by the no-report paradigm used by Cohen and colleagues (Cohen et al., 2020), we presented the Gabor patch at two different thresholds – one well above the threshold for report and the other well below the threshold for report (more on this below) – with the aim of simulating the electrophysiological correlates of conscious access with and without the cognitive demands of subjective reports.

To simulate the task, we constructed a generative model consisting of the essential features of the paradigm specified in terms of the **A**, **B**, and **C** matrices, and **D** vectors described above. To capture the hierarchical relationship between visual cortices and frontoparietal regions, we implemented a two-level deep temporal model (Friston et al., 2017). To generate the relevant behaviour and simulate neuronal dynamics, we inverted the model using the standard marginal message passing scheme used in active inference (Parr et al., 2019). Here we provide a verbal description of the generative model which should be sufficient for conceptual understanding of the simulations. For readers seeking to reproduce the simulations, the matrix form of the generative model is described in full in the **Supplementary materials**.

The first level of the model contained two hidden state factors: "visual stimulus" (three possible states: "blank screen", "left-oriented Gabor patch", "right-oriented Gabor patch") and "attention allocation" (two possible states: "high sensory precision", "low sensory precision"); see Fig. 4. Factors at this first level were intended to correspond (in a minimal sense) to sensory processing within the visual system and saliency maps within posterior parietal cortex, respectively.

The second, temporally deep level of the model had four hidden state factors: "stimulus sequence" (three states: "left Gabor sequence", "right Gabor sequence", "blank sequence"), "task phase" (six states: "1" … "6"), working memory goal (three states: "null", "don't maintain", "maintain"), and report (three states: "null", "not seen", "seen"). Factors at this level were, broadly speaking, intended to correspond (in a minimal sense) to the frontoparietal or "executive control" network (Thomas Yeo et al., 2011), where each state factor might be thought of as a distinct network hub.

In our previous model (Whyte and Smith, 2021), perceptual contents were automatically gated into the temporally deep "working memory" level (independent of policies). Agents then selected subjective report polices based on the information encoded in working memory. In contrast, we here modelled working memory maintenance as itself being a type of policy selection (i.e., in addition to verbal report policies). This was motivated by previous work casting working memory gating as a kind of 'cognitive action' (Limanowski and Friston, 2018), implemented in parallel loops connecting PFC regions with the basal ganglia (Hazy et al., 2007; O'Reilly and Frank, 2006). The idea here was to cast report and no-report conditions in terms of the incentives motivating the selection of working memory policies – based on a prior body of literature showing that the allocation of working memory resources is a motivated process that depends upon task incentives (Westbrook and Braver, 2016). To do so, we modelled report vs no-report conditions by first giving our agents a (slight) preference to avoid the mental effort (cognitive demands) associated with maintaining items in working memory. This aversion was overcome in the report condition by giving the agent the additional (greater) preference for giving correct reports, while this preference was removed in the no-report condition. Thus, the goals of the agent governing policy selection were determined by the task instructions provided in each condition (i.e., by adding vs. removing the preference to report correctly in the respective conditions).

To model the neural and functional effects of working memory gating, available choices (policies) regarding what to maintain and report controlled both 1) the precision of the second-level **A** matrix that mapped first-level "stimulus states" to the second level of the model (i.e., corresponding to the gating of information from visual cortex into working memory), and 2) the precision of the second-level **B** matrix (corresponding to the voluntary maintenance of items in working memory). A high precision in the second-level **A** matrix entailed a strong influence of first-level states on second-level states. A high precision in the second-level **B** matrix entailed that working memory states remained stable over time (i.e., states only transitioned to themselves with high probability). We manipulated the precision of the **A** and **B** matrices by passing them through a softmax function equipped with precision parameters ζ and ω, respectively (i.e., inverse temperature parameters). Under a trial where the agent expected to report its experience, the

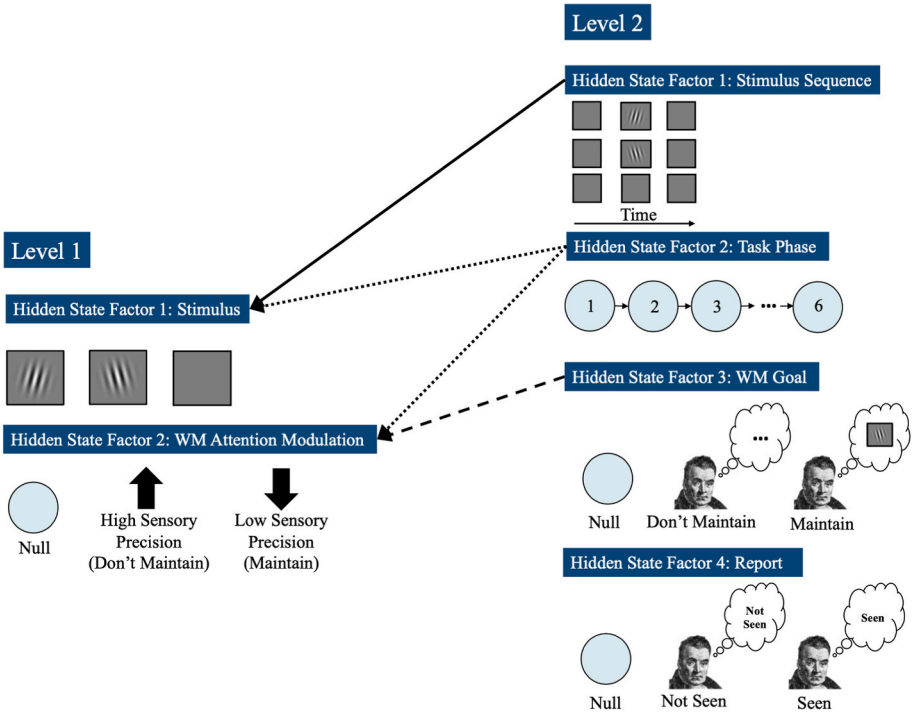

Fig. 4. Depiction of generative model structure. The first level of the model had two state factors corresponding to 1) the stimulus, and 2) working memory (WM)-dependent attentional states. The second level of the model had four hidden state factors, corresponding to 1) the stimulus sequence (e.g., that the stimulus appeared and then disappeared), 2) task phase, 3) the goal to maintain the stimulus in working memory, and 4) the choice of what to report. The second-level goal states (which governed whether to maintain or not maintain a stimulus in working memory) modulated attentional states at the first level. Namely, attention to external stimuli (i.e., the precision of the first-level state-outcome mapping) was attenuated when the agent maintained items in working memory, as a means of preventing interference from further sensory input. First-level stimulus states thus continued to represent items in working memory (i.e., after stimulus removal) due to top-down influence from the second level.

associated report policies (where either 'seen' or 'unseen' could be reported at the end of a trial) first caused the second-level **A** matrix mapping to become relatively precise ($\zeta = 1$ vs. $\zeta = 0.5$), implementing a form of goal-directed attention that lead to rapid belief updating at the second (working memory) level of the model upon stimulus presentation. These policies also increased the precision of the second-level **B** matrix ($\omega = 1.5$ vs. $\omega = 0.5$) such that working memory states remained stable over time after the stimulus had been encoded (for similar previous approaches to modelling working memory using active inference, see Parr and Friston, 2017; Smith et al., 2019). Finally, these policies down-weighted the sensory signal at the first-level of the model during the delay period of the task (i.e., reduced the precision of the first-level **A** matrix, by setting appropriate priors on first-level attentional states). Specifically, the "working memory goal" hidden state factor mapped to the first-level "attention allocation" state factor, which modulated the precision of the first-level **A** matrix (i.e., the first-level **A** matrix had different precisions under different "attention allocation" states). This had the effect that when the model was in a "maintain" state at the second level, first-level **A** matrix precision was lowered (for more details on these precision values, see **Supplementary materials**). This minimized any possible interference that could be caused by new sensory stimuli conflicting with the contents of working memory, and had the effect that first-level perceptual states continued to mirror the contents of second-level working-memory states (consistent with empirical studies showing maintained visual cortex activation during working memory maintenance (Albers et al., 2013; Serences et al., 2009)).

In contrast, when the agent did not expect to report its experience, the associated no-report policy caused the second-level **A** matrix mapping to become relatively imprecise ($\zeta = 0.5$), implementing a form of passive/diffuse attention that led to slower and less confident belief-updating at the second level of the model upon stimulus presentation. The no-report policy simultaneously reduced the precision of the second-level **B** matrix ($\omega = 0.5$), such that beliefs quickly decayed away over the delay period (i.e., no working memory maintenance). First-level sensory signals were not down-weighted under this policy (i.e., attention at the first level remained in the default, high-precision state), as there was no potential for interference with working memory demands.

The policy space was thus composed of three distinct policies – "no report", "report unseen", and "report seen" – each corresponding to a sequence of controllable state transitions in both the second-level "working memory goal" and "report" state factors. Under the "no report" policy, the "working memory goal" factor transitioned into the "don't maintain" state and stayed there (i.e., corresponding to imprecise transitions and thus decay of working memory contents), and the "report" state factor stayed in the "null" state throughout the trial. Under the "report unseen" policy, the "working memory goal" factor transitioned into the "maintain" state and stayed there (i.e., corresponding to a precise transition mapping from states back to themselves, and thus maintenance of working memory contents), and the "report" state factor transitioned from the "null" state to the "report unseen" state when the agent was asked to give a report at the end of the trial. The "report seen" policy was identical to the "report unseen" policy, except that the agent transitioned from the "null" state to the "report seen" state when asked to give a report at the last time step.

To model forced-choice behaviour, the agent was instead asked to report whether the stimulus was oriented to the left or to the right. For these simulations, the state-observation mapping was altered such that the agent would only observe feedback that it was correct when the reported orientation matched the true orientation. This contrasts with the self-report simulations described above, in which the agent was asked to state whether or not the stimulus was seen (i.e., where the mapping from states to observations entailed that reporting "seen" was correct whenever the stimulus was present, independent of whether it was oriented to the left or to the right).

In summary, the second level of the model allowed generation of reports based on combining information about task phase, goals, and stimulus sequence, with the aim of capturing the hypothesized role of PFC in conscious access. Although the hidden state factors representing this information are specified in abstract mathematical terms, we believe that three key features of this computational architecture licence the broad strokes anatomical conclusions we wish to draw. First, the lower level of the model tracks moment-to-moment changes in the stimulus, whereas the higher level tracks the sequence of states at the lower level and so necessarily evolves at a slower timescale. This separation of timescales between the lower and higher levels of the model mirrors empirical findings showing that prefrontal areas have a slower intrinsic timescale than sensory regions (Murray et al., 2014; Wolff et al., 2022). Second, the higher level of the model simulates multiple neural response patterns observed in PFC (see results below). This includes neural activity associated with goal states, task phase, and maintenance over the delay period, the stability of which is modulated by the precision of the transition probabilities at the higher level (see Kapoor et al., 2018 for evidence that distinct neural populations in PFC encode information about task phase and conscious content). In previous work, this feature of active inference has been used to model the relationship between delay period activity and recurrent glutamatergic connections in layer III of PFC (Parr et al., 2020). Third, similar to other cognitive models of working memory (e.g. Manohar et al., 2019), the maintenance of stimulus-related activity at the lower "sensory" level of the model relies on activity being fed back from the higher "working memory" level, where the magnitude of the feedback depends on the strength of higher-level delay period activity.

3. Results

3.1. Conscious access with and without report

For both the report and no-report conditions, we first simulated curves that related report frequencies, and forced-choice performance, to the posterior probability over states at the second level of the model. We did this by iteratively reducing the first-level **A** matrix precision using an additional parameter ς (i.e., this parameter interacted with the influence of the "attentional allocation" state to produce final precision values). This type of precision parameter was also included in our previous work (Whyte and Smith, 2021) and can be understood to correspond to stimulus strength manipulations, such as contrast or presentation time (see **Supplementary materials** for details of how this was implemented). We iteratively reduced precision by starting at the lowest value where the model still reported the stimulus 100% of the time ($\varsigma = .22$ for report simulations, and $\varsigma = .2$ for forced-choice simulations) and stopped when report frequency dropped to 0% (Fig. 5a). At each level of precision, we simulated 100 trials. This allowed us to study the hypothetical report frequencies of the model in the no report condition based on the posterior confidence threshold obtained in the report condition. Specifically, because we had access to the posterior probability at the second level, in both the report and no-report conditions, we could ask, *counterfactually*, whether the agent *would* have reported seeing the Gabor *if* we had asked it to do so.

We used this approach to address the primary aim of our simulations, which was to explain why the neural correlates of conscious access change as a function of reporting demands. As reviewed above, while traditional (report) paradigms in visual consciousness research have found strong relationships between self-reports, the P3b, and elevated PFC activation, when reporting demands are removed, the P3b vanishes entirely and PFC activity is greatly reduced. The next two sub-sections describe the simulated firing rates and ERPs of a set of simulations based on the no-report paradigm of Cohen and colleagues (Cohen et al., 2020). Analogous to the experimental procedure used in that study, we presented the agent with stimuli both well above report threshold ($\varsigma = .5$), and well below report threshold ($\varsigma = .05$), in both report and no-report conditions, leading to four conditions; *above threshold/report*,

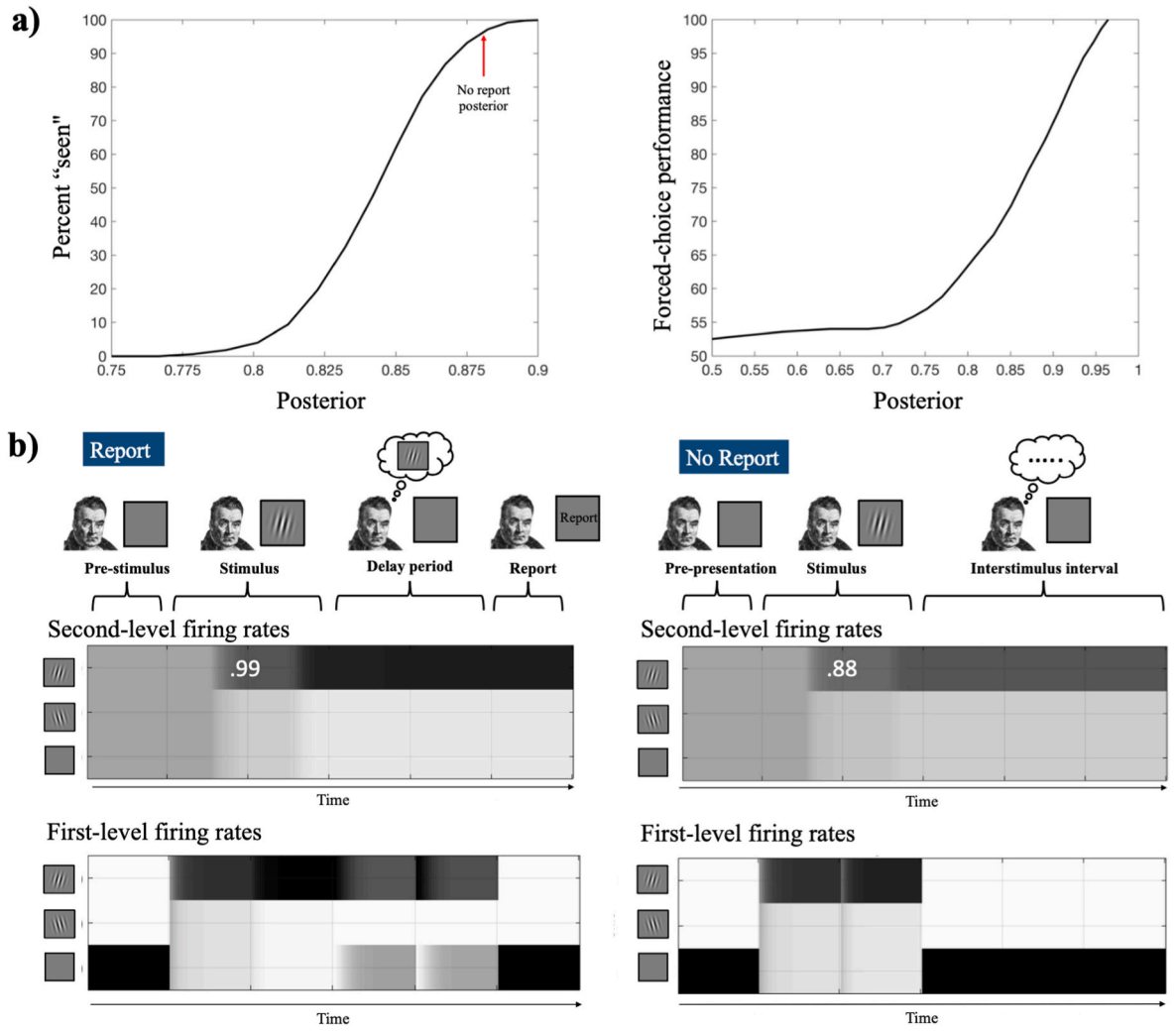

Fig. 5. The left plot in the top panel **(a)** shows the percentage of trials across 100 simulations in which agents reported seeing a stimulus given different posterior probabilities over second-level states. The top-right plot illustrates how posterior beliefs were related to accuracy in an analogous set of forced-choice simulations (see main text). The bottom panel **(b)** shows the normalised firing rates (posterior probabilities) for report and no-report trials. Rows represent each possible stimulus percept (hidden state), columns indicate discrete time-steps (where each discrete time step consists of 16 iterations of gradient decent on variational free energy with respect to states), and darker colours indicate higher firing rates (higher probabilities). These results illustrate that, even though second-level firing rates (encoding posteriors over stimulus sequence states) in the no-report condition are lower than in the report condition, these firing rates (and associated posterior probabilities) are still above the threshold for (close to 100%) successful reporting in the report condition (indicated by the red arrow; posterior probability = .88). Thus, despite observed reductions in PFC activation in no-report studies (as is also the case in these simulations), this suggests that such findings need not indicate a lack of prefrontal involvement when stimuli are seen. That is, the simulated prefrontal firing rates (and associated posterior probabilities) in the no-report condition are still sufficiently high at stimulus presentation that the agent – based on behaviour in the reporting condition – would still have been successful in self-reporting if asked to do so (i.e., consistent with a theory in which prefrontal involvement is essential to access). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

above threshold/no-report, *below threshold/report*, and *below threshold/no-report*.

3.1.1. Simulated firing rates

In studies using fMRI (Bisenius et al., 2015), and more direct measures of neural activity such as intracranial EEG (Gaillard et al., 2009), subjective reports of conscious access are associated with a large increase in activity throughout the frontoparietal network (even when contrasted with conditions where participants still generate a report of not seeing a stimulus; c.f. Michel and Morales, 2020). However, as reviewed in the introduction, when reporting demands are removed prefrontal activity drops below the level of significance in fMRI (Frassle et al., 2014), and it is greatly reduced in magnitude (although still present) when measured by more direct means (Noy et al., 2015) (e.g. electrocortitcography, ECoG).

The simulated firing rates of our model, in each of the four conditions described above, speak directly to this set of results (see Fig. 5b). In line with traditional report paradigms, in the *above threshold/report* condition the second-level firing rates for the presented stimulus were substantially higher when the stimulus was presented above the threshold for report, with a posterior probability at the end of the stimulus presentation of 0.99 for the presented stimulus (a right-oriented Gabor). In contrast, when we removed the reporting demands – leading the precision of the messages sent from the first to the second level of the model to be reduced (i.e., corresponding to a reduction in the strength of afferent connectivity between visual cortex and frontoparietal regions) – the firing rates at the second level of the model during the stimulus presentation period were attenuated. Crucially, however, the peak posterior probability for the presented stimulus (again a right-oriented Gabor) over the presentation period was 0.88, which we know from our report curves (Fig. 5a) is at the threshold for approximately 100% reportability.

This shows that the reduction in PFC activity that accompanies the removal of reporting demands can be explained by the corresponding reduction in the precision of the messages being passed between sensory cortex and PFC. In other words, removing reporting demands reduced the strength of effective connectivity between sensory cortex and PFC.

We now turn to the results of the two below-threshold conditions (Fig. 6). In the *below threshold/report* condition, although stimulus precision was too low to be reported as visible, it was still high enough to have stimulus information be propagated to the second level of the model during the stimulus presentation period. After the stimulus was removed, the model assigned most of the probability mass to the "blank" stimulus state (as it was confident in not having seen the stimulus). Importantly, however, as can be seen in the firing rates (Fig. 6), there was still a greater posterior probability assigned to the presented stimulus state (right Gabor) than the other possible stimulus state (left Gabor). So when the model was given a forced choice between the left and right orientation, its performance was still slightly above chance 52% (to see this, compare the second-level posterior probability for the presented stimulus shown in Fig. 6 to the report curves in Fig. 5). Finally, in the *below threshold/no-report* condition, the model was well below the posterior confidence threshold for both report and forced-choice performance.

Although the primary explanatory target of our model was the change in the neural correlates of consciousness as a function of reporting demands, it is worth highlighting that the results in this section are also very much in line with the results of King and colleagues (King et al., 2016), who used an unconscious working memory paradigm very similar to the task performed by our simulated agent. Specifically, they found that they could decode the presence and orientation of the target stimulus (also a Gabor) throughout a delay period – even when the stimulus was reported unseen. Like our simulated agent, they found that participants' reports of visibility correlated with forced-choice performance. Crucially, and also like our simulated agent, even when the participants reported the stimulus as invisible, they still performed marginally above chance in a forced-choice task. To see this effect in our results, compare the forced-choice accuracy curve to the subjective report curve. Below a posterior probability of .78, the model no longer reported the stimulus as seen, but it still performs well above chance in the forced-choice task. In addition, analogous to how report percentage is tied to the second-level firing rate in our model (i.e., to second-level posterior probability), King and colleagues found that decoding accuracy during the delay period correlated with participants' visibility ratings, suggesting a common neural substrate for report and maintenance. That is, if maintenance and visibility relied on independent substrates, then decoding should have been identical across visibility levels, but this is not what was found. This therefore provides additional face validity to our model structure, as simulated neural activity in the second level of our model would show a similar pattern of decodability in trials where the stimulus is not reported as seen.

3.1.2. Simulated event related potentials

In the previous sub-section, we saw how, with a relatively simple model, we can account for the results of both no-report paradigms and a prominent study on unconscious working memory. In this sub-section, we examine the ERPs produced by our model in each of the four conditions described above, and relate them to previous empirical findings in the literature. We will first review some of these empirical findings in more detail to motivate the importance of the questions we address.

When conscious access is measured via trial-by-trial subjective report, the amplitude of the P3b, a relatively late frontocentral ERP, correlates tightly with stimulus visibility. At one point, GNW theorists considered the P3b to be a signature of sensory contents becoming conscious. The idea was that when perceptual contents gain access to the frontoparietal network identified with the global workspace, only a fraction of the inputs are depolarised, while the majority of competing inputs are inhibited. This response pattern was argued to explain the frontocentral positivity that characterises the P3b (Dehaene, 2014). However, this result has come under increasing scrutiny, leading to the now accepted view that the P3b reflects the task demands associated with reports of conscious access, and not conscious access itself (Cohen et al., 2020; Förster et al., 2020; Sergent et al., 2021). The first experiment to convincingly dissociate conscious access and the P3b was conducted by Pitts and colleagues (Pitts et al., 2014), who, as described in the introduction, used a no-report variant of an inattentional blindness

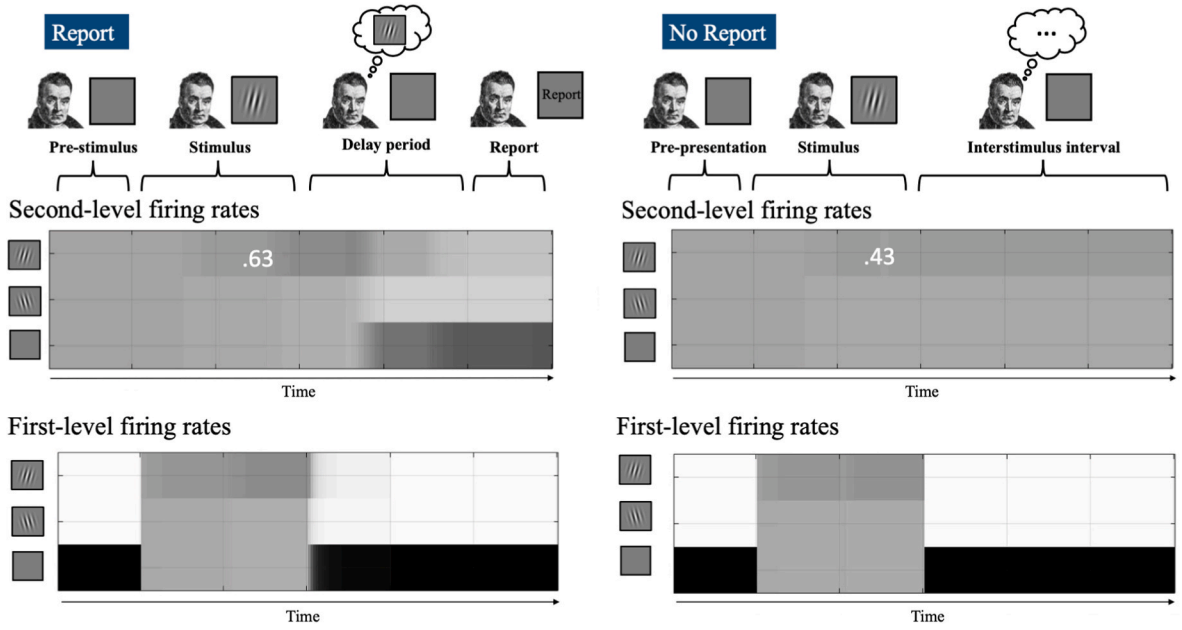

Fig. 6. First- and second-level firing rates when stimuli were presented below visibility threshold in report (left) and no-report (right) conditions. As can be seen, second-level posteriors were higher in the report condition, and still slightly favoured the correct stimulus over the incorrect stimulus at the report phase – providing a basis for above-chance forced-choice performance. Recall that darker shades in these plots indicate higher firing rates and that each row corresponds to a possible hidden (perceptual) state. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

paradigm that collected reports after each block instead of trial-by-trail. They found that the amplitude of the P3b correlated with task relevance, not awareness. Yet, because reports were collected block-by-block, and attention was at best only diffusely present until the stimulus was made task relevant, the results are open to the objection that participants may have missed the occurrence of the critical stimulus on a substantial number of trials, explaining the lack of P3b. This objection was addressed by a study conducted by Cohen and colleagues (Cohen et al., 2020), whose paradigm inspired the task performed here by our simulated agent. They found that the results of Pitts and colleagues (Pitts et al., 2014) generalised to a standard masking paradigm. Specifically, they presented the stimulus well above the threshold for visibility, and well below the threshold for visibility, under both report and no-report conditions. Crucially, in the no-report condition they had participants perform an incidental memory task, which they performed with a high degree of accuracy, but only for stimuli presented above the threshold for visibility – thereby reducing the plausibility of the objection that the unmasked stimuli may not have been experienced on no-report blocks. Just like Pitts and colleagues (Pitts et al., 2014), Cohen and colleagues (Cohen et al., 2020) found that the P3b correlated with reporting demands, but not conscious access (i.e., as with our results in Fig. 7 below). Unlike the results of Pitts and colleagues (Pitts et al., 2014), the absence of the P3b in the no-report condition cannot be parsimoniously explained by a diffuse allocation of attention, since the participants were not performing a separate attention-demanding task like they were in the Pitts and colleagues (Pitts et al., 2014) study. With respect to the series of results just reviewed, the purpose of the model presented in this paper was therefore to see if the working memory gating hypothesis,

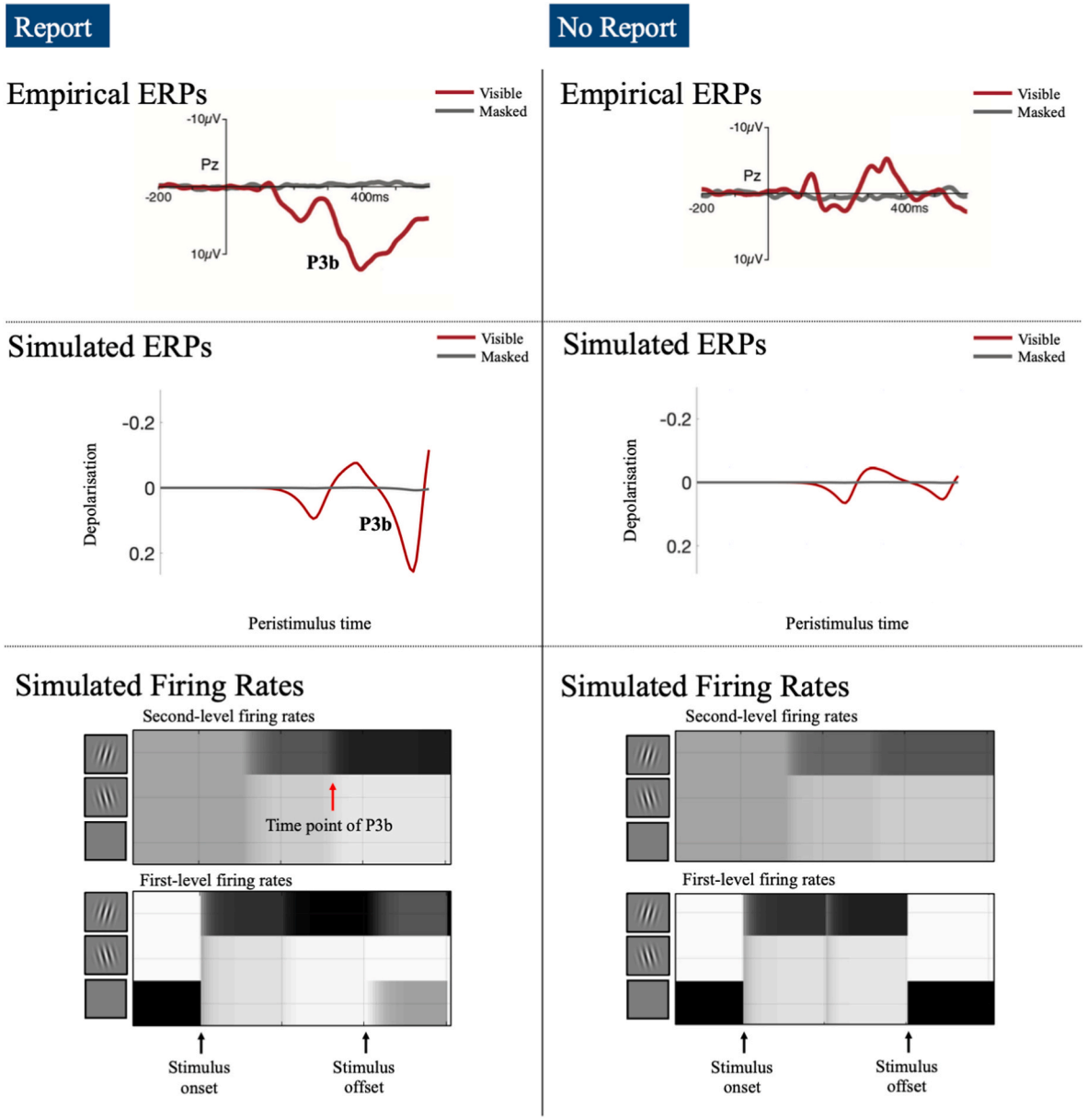

Fig. 7. This figure illustrates how the cognitive demands of reporting generate rapidly increasing firing rates at the second level of the model – and the resultantly large P3b-like ERPs in the report condition (i.e., based on the neural process theory associated with active inference, in which ERPs reflect rates of change in posterior beliefs over states and associated firing rates). In contrast, rates of change in firing rates are lower in the no-report condition – leading to the absence of P3b-like ERPs. This reproduces and offers an explanation for the results of Cohen and colleagues (Cohen et al., 2020). Thus, despite the persistence of PFC activity (associated with conscious access) in the simulations in Fig. 5, these latter results simultaneously account for the dissociation between the P3b and conscious access. We remind the reader that darker shades in these plots indicate higher firing rates, time-in-trial progresses from left to right, and each row corresponds to a possible hidden (perceptual) state. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

which could parsimoniously explain the univariate fMRI and ECoG no-report results described above, might also be capable of explaining the lack of the P3b in a standard masking paradigm once reporting demands are removed.

Despite the simplicity of our model, the simulated ERPs in each condition display a striking similarity to the results of Cohen and colleagues (Cohen et al., 2020). Fig. 7 shows the simulated ERPs for each of the four conditions alongside the empirical results. Specifically, we plotted the temporal derivative of the normalised firing rate (posterior over states at each epoch of gradient decent; see **Appendix 1**) described above – the proposed basis of ERPs under the active inference process theory – with respect to states at the second level of the model for each of the four conditions (*above threshold/report*, *below threshold/report*, *above threshold/no-report*, and *below threshold/no-report*). Reproducing the empirical results, our simulations show that a P3b-like ERP is present in the report condition but vanishes in the absence of reporting demands. Crucially, we know from our previous examination of the second-level posterior beliefs that the model is still above the threshold for near 100% visibility in both the report and no-report conditions. Inspection of the firing rates associated with the report and no-report conditions (bottom panel of Fig. 7) makes the reason for the difference in ERPs apparent. In the report condition, the messages passed between the first and second level of the model are precise and lead to a rapid rate of belief updating at the second level of the model (i.e., corresponding to a large rate of change in the posterior over states), generating an ERP resembling the P3b. In contrast, in the no-report condition the precision of the messages passed between levels was greatly reduced, leading second-level posterior beliefs to update in a more gradual manner and therefore generating a (substantially) smaller ERP. Importantly, this interpretation of the no-report results suggested by our model leads to a strong empirical prediction. Namely, if a dynamic causal model (DCM; Friston et al., 2003; Kiebel et al., 2008) with nodes for visual and lateral frontal cortices were fitted to EEG data from this or a similar no-report paradigm (Cohen et al., 2020; Sergent et al., 2021), we should see a modulation of bidirectional connectivity as a function of reporting demands. Specifically, the strength of bidirectional connectivity between frontal and visual cortices should be reduced in the absence of report.

4. Discussion

The debate between cognitive theorists (Brown et al., 2019; Dehaene and Changeux, 2011; Mashour et al., 2020) and perceptual theorists (Lamme, 2006; Oizumi et al., 2014) about the involvement of prefrontal cortex in consciousness predates the introduction of no-report paradigms. But with the experimental contrast between conscious and unconscious conditions depending, until recently, on the collection of subjective reports, evidence from both neuroimaging and invasive electrophysiology in non-human primates primarily supported cognitive theories (Bisenius et al., 2015; Panagiotaropoulos et al., 2012). With the introduction of no-report methods (Tsuchiya et al., 2015), however, much of this evidence came under scrutiny and new results seemed to confirm the central arguments of perceptual theorists – namely, that consciousness is, at bottom, a perceptual phenomenon, that PFC (at most) modulates conscious access, and that its activation in neuroimaging studies is a result of the cognitive demands of report generation (Boly et al., 2017; Tsuchiya et al., 2015). As reviewed in the introduction, the evidential pendulum has now begun to swing back in the other direction. Namely, when no-report paradigms have been used in conjunction with invasive recordings and more sophisticated analytic techniques, supporting evidence for cognitive theories has emerged (Dwarakanath et al., 2020; Kapoor et al., 2020; Sergent et al., 2021). This recent body of work led to the question that we set out to answer in this paper. Namely, given the now compelling evidence for the involvement of PFC in consciousness, what is the difference in *cognitive* processing that explains why the neural correlates of consciousness change as a function of reporting demands?

Here we advanced a model of conscious access that casts the engagement of cognitive resources associated with reporting as a variety of 'cognitive action' – here corresponding to goal-directed adjustments in the effective connectivity within and between frontal and visual cortices (Limanowski and Friston, 2018). Under our model, conscious access is simply a matter of having a high enough posterior probability over states at a temporally deep level of representation. This allows an agent to have access to its own first-level perceptual representations such that it could report them – and counterfactually so in no-report conditions. That is, simulated PFC activity crosses the threshold for reportability at stimulus presentation in no-report tasks, despite remaining lower than in report tasks. Building on our previous work (Hohwy, 2013; Smith, 2016, 2017; Whyte, 2019; Whyte and Smith, 2021), we propose that conscious access occurs when sensory states have a high enough precision that inferences about these states at temporally deep levels of the cortical hierarchy pass the posterior probability threshold for reporting (i.e., where reporting here stands in for other goal-directed uses of information enabled by temporally deep processing). The content of our moment-to-moment experience is therefore not encoded at temporally deep prefrontal levels of the model (i.e., whose content here corresponds to longer-timescale perceptual sequences). Rather it is the process of first-level perceptual representations being integrated into a temporally deep representation that makes their contents conscious. In other words, perceptual experience corresponds to the moment-to-moment updates to the deeper prefrontal level, based on the impact of the messages passed up to prefrontal regions from sensory cortices.

Crucially, if reporting requires a type of 'cognitive action' in which sensory content is gated into, and maintained in, working memory, then this carries important functional implications. Specifically, it follows that imposing reporting demands will require an agent to effect goal-directed increases in the precision of the messages being passed between sensory and prefrontal cortices (as well as within prefrontal cortices) – thereby altering the neural correlates of conscious access. If the precision of these messages instead remains low in no-report conditions, and the rate/magnitude of belief updates in prefrontal cortex is therefore reduced, then this can account for the reduced prefrontal activity and smaller P3b ERPs seen empirically (Cohen et al., 2020; Pitts et al., 2014). Conversely, when agents are required to report their experience, the resulting increase in the precision of messages passed between sensory and prefrontal cortices will *increase* the rate of belief updating, resulting in increased PFC activity and a large P3b. Indeed, this is exactly what was observed in a recent simultaneous EEG-fMRI experiment (Dellert et al., 2021), which had participants perform a modified version of the Pitts and colleagues (Pitts et al., 2014) inattentional blindness task. When the stimulus was conscious, but not task relevant, there was strong activation of visual regions, and a large N170, but only weak prefrontal activation and no P3b. In contrast, when the stimulus was conscious, and task relevant, there was strong prefrontal activation, and a large P3b (Dellert et al., 2021).

A direct prediction of our account is that the frequency of perceptual transitions in binocular rivalry, which are known to be slowed in the absence of attention (Paffen et al., 2006), should likewise be slowed in the absence of reporting demands. That is, under our model, and the active inference process theory on which it is based, both attention and working memory gating modulate the precision of afferent sensory information sent from lower to higher levels of processing. We would predict, therefore, that reduced working memory demands, like the absence of attention, should slow the speed of transitions in rivalry. This is indeed what was found by Frässle and colleagues (Frassle et al., 2014), who reported that in the passive (i.e., no-report) condition of their binocular rivalry paradigm, perceptual transitions (when measured by changes in pupil size) were slower than in the report condition.

Throughout this paper we have referred to a "cognitive theory" as any theory which posits that PFC plays a necessary functional role in determining the contents of consciousness. There are of course

important differences between cognitive theories. Most prominently, higher order theory (HOT) proposes that PFC houses a mechanism for higher-order representations, where lower-order representations become conscious when they are, in some way, the target of such higher-order representations (Brown et al., 2019). In contrast, global neuronal workspace (GNW) theory posits that PFC, among a network of other structures, is part of a "global workspace" that houses the contents of consciousness. On this view, contents become conscious when the workspace is "ignited", corresponding to selectively increased effective connectivity between regions allowing the GNW network to gain access to locally represented information (i.e., that would otherwise only be represented unconsciously). The signature of this ignition process is a non-linear bifurcation-like phenomenon that can be detected in neural activity (Joglekar et al., 2017; Mashour et al., 2020; Sergent et al., 2021). This of course raises the question of whether the model and simulations presented in this paper provide support for HOT or the GNW. Reminiscent of HOT, conscious access in our model depends on a second (higher) level that represents information about the sensory contents at the first level. However, this is not a simple re-representation. Instead, the second level infers the sequential unfolding of states at the first level, and it is the temporal depth of the second level – not the fact that it re-represents information about the first level – that allows the model to generate reports. We therefore see our model as more closely aligned with the GNW. Indeed, we have previously proposed (Hohwy, 2013; Whyte, 2019), and later shown through simulation (Whyte and Smith, 2021), that many of the classic behavioural and neural results cited in support of GNW theory (including ignition dynamics) arise naturally in a two-level active inference model of conscious access/report generation. Also reminiscent of the GNW, our model requires several independent state factors at the higher level that must jointly gain access to first-level contents. These different state factors could be seen as different hubs in a GNW-like network, only some of which engage in the above-mentioned form of re-representation. Thus, our model can be seen as capturing and integrating aspects of HOT, GNW theory, and predictive processing – highlighting similarities that may not otherwise be apparent.

We follow Hohwy and Seth (Hohwy and Seth, 2020; Seth and Hohwy, 2021; also see Vilas et al., 2021) in interpreting the present active inference model as a theory *for* consciousness science rather than a theory *of* consciousness. That is, instead of constructing a theory of consciousness *per se*, our approach appeals to active inference as a general theory of brain function that can be used to construct generative models of representative experimental tasks in consciousness research. This approach allowed us to show in simulations how inferential dynamics in the model naturally reproduce and provide mechanistic explanations for the neurophysiological signatures known to accompany conscious access. Since the model was based on the structure of an empirical task, and did not assume a specific cognitive theory, we interpret our results as supporting any cognitive theory that embraces working-memory gating as a necessary component in the mechanism(s) of report generation.

To conclude, we have presented a computational model that demonstrates how results in no-report paradigms need not pose a problem for cognitive theories of consciousness. In fact, the results of these paradigms emerge naturally when the engagement of cognitive resources needed for reporting is modelled as a variety of cognitive action under temporally deep active inference. In this model, which is consistent with cognitive theories, one can reproduce and explain available empirical results while also demonstrating how prefrontal engagement remains necessary and associated with conscious contents. It is important to acknowledge that this model and associated simulations are vastly oversimplified, and future work will need to both extend the scope of these simulations and test their empirical predictions. This notwithstanding, the general computational architecture and dynamics of the model and simulation results we have shown appear to provide important and potentially generalizable insights about the way current results can be explained in a unifying and theoretically informative manner.

CRediT authorship contribution statement

Christopher J. Whyte: Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing. **Jakob Hohwy:** Writing – review & editing. **Ryan Smith:** Conceptualization, Formal analysis, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Shay Tobin, Catriona Scrivener, and the audience at the Meta Lab's Consciousness Club seminar series for discussion and feedback on the topic of this manuscript. C.W. is supported by the University of Cambridge Harding Distinguished Postgraduate Scholars Programme. J. H. is supported by the Australian Research Council DP160102770 and the Three Springs Foundation. R.S. is funded by the William K. Warren Foundation and the National Institute of General Medical Sciences (P20GM121312).

Appendix 1. Objective functions and update equations

Here the generative model is a partially observable Markov decision process allowing the joint distribution to be factorised into the product of a prior over the initial state $p(s_1)$, prior over policies $p(\pi)$, likelihood $p(o_\tau|s_\tau)$, and transition probability $p(s_\tau|s_{\tau-1},\pi)$.

$$p(o_{1:T}, s_{1:T}, \pi) = p(s_1)p(\pi)\prod_{\tau=1}^{T} p(o_\tau|s_\tau)\prod_{\tau=2}^{T} p(s_\tau|s_{\tau-1},\pi) \tag{A.1}$$

$$p(o_{1:T}, s_{1:T}|\pi) = p(s_1)\prod_{\tau=1}^{T} p(o_\tau|s_\tau)\prod_{\tau=2}^{T} p(s_\tau|s_{\tau-1},\pi) \tag{L2}$$

$$= s_1 \cdot \mathbf{D}\prod_{\tau=1}^{T} o_\tau \cdot \mathbf{A} s_\tau \prod_{\tau=2}^{T} s_\tau \cdot \mathbf{B}_{\pi,\tau} s_{\tau-1} \tag{L3}$$

The second line of A.1 conditions the generative model on π as inference is performed under each available policy. Line three then shows the matrix form of the generative model conditioned on π, replacing each categorical distribution with matrices whose columns contain the parameters of the distributions. That is, $p(s_1) = s_1 \cdot \mathbf{D}$, $p(o_\tau|s_\tau) = o_\tau \cdot \mathbf{A} s_\tau$, and $p(s_\tau|s_{\tau-1},\pi) = s_\tau \cdot \mathbf{B}_{\pi,\tau} s_{\tau-1}$. Where o_τ and $s_{\tau-1}$ are vectors of zeros with a one placed in the element corresponding to the state or observation of interest. These vectors select the relevant element of the **A** and **B** matrices corresponding to a

specific state-outcome pair or current state-previous state pair.

To perform state estimation under this generative model in a tractable manner we need to introduce an objective function which can be optimised via gradient decent. Here we use marginal free energy which is defined as follows:

$$F_{\pi,\tau} = s_{\pi,\tau} \cdot \left(\ln s_{\pi,\tau} - \frac{1}{2} \left(\ln \mathbf{B}_{\pi,\tau-1} s_{\pi,\tau-1} + \ln \mathbf{B}^{\dagger}_{\pi,\tau} s_{\pi,\tau+1} \right) - \ln \mathbf{A}^{\mathrm{T}} o_{\tau} \right) \tag{A.2}$$

Here, $\mathbf{B}^{\dagger}_{\pi,\tau}$ denotes the transpose of $\mathbf{B}_{\pi,\tau}$ with normalised columns, and when $\tau = 1$, $\mathbf{B}_{\pi,\tau-1} s_{\pi,\tau-1}$ is replaced by $\mathbf{D}$. For the derivation and motivation behind marginal free energy, see Parr et al. (2019).

With marginal free energy serving as our objective function, we take the gradient of marginal free energy with respect to states for each of its arguments to obtain the following:

$$\nabla_{s_{\pi,\tau}} F_{\pi} = \ln s_{\pi,\tau} + \overleftarrow{1} - \frac{1}{2} \left(\ln \mathbf{B}_{\pi,\tau-1} s_{\pi,\tau-1} + \ln \mathbf{B}^{\dagger}_{\pi,\tau} s_{\pi,\tau+1} \right) - \ln \mathbf{A}^{\mathrm{T}} o_{\tau} \tag{A.3}$$

$$-\nabla_{s_{\pi,\tau}} F_{\pi} = \frac{1}{2} \left(\ln \mathbf{B}_{\pi,\tau-1} s_{\pi,\tau-1} + \ln \mathbf{B}^{\dagger}_{\pi,\tau} s_{\pi,\tau+1} \right) + \ln \mathbf{A}^{\mathrm{T}} o_{\tau} - \ln s_{\pi,\tau} \tag{L.2}$$

Line two multiplies both sides by negative one and drops the vector of ones as it is constant across elements of the gradient. It is this negative (marginal) free energy gradient that we define as state prediction error $-\nabla_{s_{\pi,\tau}} F_{\pi} = \varepsilon_{\pi,\tau}$. This makes conceptual sense. The negative free energy gradient is the (log) difference between the generative model after having received an observation and our approximate posterior over states.

Finally, we define a new 'depolarisation' variable $v_{\pi,\tau} = \ln s_{\pi,\tau}$, which represents the membrane potential/voltage of the neuronal population encoding the posterior over states. Using our state prediction error and membrane potential variables we then write down a set of equations that perform a gradient decent on marginal free energy with respect to states under each policy and can be iterated until convergence.

$$v_{\pi,\tau} \leftarrow \ln s_{\pi,\tau} \tag{A.4}$$

$$\varepsilon_{\pi,\tau} \leftarrow \frac{1}{2} \left(\ln \mathbf{B}_{\pi,\tau-1} s_{\pi,\tau-1} + \ln \mathbf{B}^{\dagger}_{\pi,\tau} s_{\pi,\tau+1} \right) + \ln \mathbf{A}^{\mathrm{T}} o_{\tau} - v_{\pi,\tau} \tag{L.2}$$

$$v_{\pi,\tau} \leftarrow v_{\pi,\tau} + \varepsilon_{\pi,\tau} \tag{L.3}$$

$$s_{\pi,\tau} \leftarrow \sigma\left(v_{\pi,\tau}\right) \tag{L.4}$$

Line one of equation A.4 initialises the depolarisation variable as the log of the approximate posterior over states. Line two calculates the value of the negative free energy gradient which is then used to update the depolarisation variable ($v_{\pi,\tau}$) in the direction of steepest decent in line three. Finally, in line four, the depolarisation variable is normalised to give an updated value for the approximate posterior over states.

The gradient decent on marginal free energy serves a dual purpose: 1) it optimises the approximate posterior distribution over states such that it approximates the true posterior over states, and 2) it furnishes us with a model of neuronal dynamics that, as we highlighted in the main text, has a degree of face validity. Namely, because the membrane potential (log posterior over states) is not normalised it can take both positive and negative values like voltage, and after the depolarisation variable is transformed into an approximate posterior by passing it through a softmax function, it has a value that is bounded between zero and one, like a normalised firing rate.

Turning now to action selection. We would like our agent to select actions that will minimize marginal free energy in the future. This is, however, not straightforwardly possible since marginal free energy depends upon observations that by definition have not yet occurred. The solution to this problem is to introduce a new quantity, expected free energy (denoted G_{π}), which, unlike marginal free energy, treats observations as random variables that enter into the expectation operator.

$$G_{\pi} = \mathrm{E}_{q(o,s|\pi)}[\ln q(s|\pi) - \ln p(o,s|\pi)] \tag{A.5}$$

$$\approx \mathrm{D}_{\mathrm{KL}}(q(o|\pi)||p(o|C)) + \mathrm{E}_{q(s|\pi)}[\mathrm{H}[p(o|s)]] \tag{L2}$$

$$= \sum_{\tau} \left(\mathbf{A} s_{\pi,\tau} \cdot \left(\ln \mathbf{A} s_{\pi,\tau} - \ln \mathbf{C}_{\tau} \right) - diag\left(\mathbf{A}^{\mathrm{T}} \ln \mathbf{A} \right) \cdot s_{\pi,\tau} \right) \tag{L3}$$

Equation (A.5) starts with the definition of expected free energy for a generic generative model. Line two shows the standard decomposition of expected free energy into risk plus ambiguity. Risk is the KL divergence between the observations expected under a policy and the preference distribution $\mathrm{D}_{\mathrm{KL}}(q(o|\pi)||p(o|C))$, where C encodes the agent's preferences over observations. To minimize risk, agents must select policies expected to generate the observations that are most preferred. Ambiguity is the expected entropy of the likelihood. To minimize ambiguity, agent's must select policies that minimize the entropy of the likelihood $\mathrm{E}_{q(s|\pi)}[\mathrm{H}[p(o|s)]]$. That is, they must select actions that minimize the uncertainty of the mapping between states and observations. To minimize expected free energy as a whole, agents balance between minimizing risk and minimizing ambiguity, solving the exploration-exploitation dilemma. Finally, line 3 shows the matrix form of the risk + ambiguity decomposition, replacing the distributions with matrices/vectors whose columns contain the elements of the distributions. For a step-by-step derivation of this decomposition, see the appendix of Smith et al. (2022).

The posterior over policies is then a softmax function of both the expected free energy under each policy, and the marginal free energy under each policy.

$$q(\pi) = \sigma(-F - G) \tag{A.6}$$

Including both marginal free energy and expected free energy in the term for the posterior over policies is not always necessary. For policies that only look one time step ahead we only need expected free energy. We include both here because the simulations in this paper are chiefly concerned with the selection of deep policies that require the agent to infer the policy it is following over the entire time horizon of each trial (while in each epoch of the trial). The inclusion of marginal free energy means that as each trial progresses and new observations are received at each epoch, policies that

are inconsistent with present observations will become implausible. For example, when our agent holds an item in memory over the delay, first-level observations are generated that are inconsistent with the "no-report" policy, rendering it implausible.

Supplementary materials

Supplementary materials to this article can be found online at https://doi.org/10.1016/j.crneur.2022.100036.

References

Albers, A.M., Kok, P., Toni, I., Dijkerman, H.C., de Lange, F.P., 2013. Shared representations for working memory and mental imagery in early visual cortex. Curr. Biol. 23 (15), 1427–1431. https://doi.org/10.1016/j.cub.2013.05.065.

Bisenius, S., Trapp, S., Neumann, J., Schroeter, M.L., 2015. Identifying neural correlates of visual consciousness with ALE meta-analyses. Neuroimage 122, 177–187. https://doi.org/10.1016/j.neuroimage.2015.07.070.

Boly, M., Massimini, M., Tsychiya, 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* [Preprint]. Neuroscience. https://doi.org/10.1101/118273.

Breakspear, M., 2017. Dynamic models of large-scale brain activity. Nat. Neurosci. 20 (3), 340–352. https://doi.org/10.1038/nn.4497.

Brown, R., Lau, H., LeDoux, J., 2019. The Misunderstood Higher-Order Approach to Consciousness. https://doi.org/10.31234/osf.io/xpy8h [Preprint]. PsyArXiv.

Cohen, M.A., Ortego, K., Kyroudis, A., Pitts, M., 2020. Distinguishing the neural correlates of perceptual awareness and postperceptual processing. J. Neurosci. 40 (25), 4925–4935. https://doi.org/10.1523/JNEUROSCI.0120-20.2020.

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., Friston, K., 2020. Active inference on discrete state-spaces: a synthesis. J. Math. Psychol. 99, 102447. https://doi.org/10.1016/j.jmp.2020.102447.

Da Costa, L., Parr, T., Sengupta, B., Friston, K., 2021. Neural dynamics under active inference: plausibility and efficiency of information processing. Entropy 23 (4), 454. https://doi.org/10.3390/e23040454.

Dehaene, S., 2014. Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts. Penguin Publishing Group.

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.

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. J. Neurosci. 41 (37), 7864–7875. https://doi.org/10.1523/JNEUROSCI.2799-20.2021.

Dwarakanath, A., Kapoor, V., Werner, J., Safavi, S., Fedorov, L.A., Logothetis, N.K., Panagiotaropoulos, T.I., 2020. *Prefrontal state fluctuations control access to consciousness* [Preprint]. Neuroscience. https://doi.org/10.1101/2020.01.29.924928.

Förster, J., Koivisto, M., Revonsuo, A., 2020. ERP and MEG correlates of visual consciousness: the second decade. Conscious. Cognit. 80, 102917. https://doi.org/10.1016/j.concog.2020.102917.

Frassle, S., Sommer, J., Jansen, A., Naber, M., Einhauser, W., 2014. Binocular rivalry: frontal activity relates to introspection and action but not to perception. J. Neurosci. 34 (5), 1738–1747. https://doi.org/10.1523/JNEUROSCI.4403-13.2014.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O Doherty, J., Pezzulo, G., 2016. Active inference and learning. Neurosci. Biobehav. Rev. 68, 862–879. https://doi.org/10.1016/j.neubiorev.2016.06.022.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G., 2017. Active inference: a process theory. Neural Comput. 29 (1), 1–49. https://doi.org/10.1162/NECO_a_00912.

Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. Neuroimage 19 (4), 1273–1302. https://doi.org/10.1016/S1053-8119(03)00202-7.

Friston, K.J., Rosch, R., Parr, T., Price, C., Bowman, H., 2017. Deep temporal models and active inference. Neurosci. Biobehav. Rev. 77, 388–402. https://doi.org/10.1016/j.neubiorev.2017.04.009.

Gaillard, R., Dehaene, S., Adam, C., Clémenceau, S., Hasboun, D., Baulac, M., Cohen, L., Naccache, L., 2009. Converging intracranial markers of conscious access. PLoS Biol. 7 (3), e1000061 https://doi.org/10.1371/journal.pbio.1000061.

Gelbard-Sagiv, H., Mudrik, L., Hill, M.R., Koch, C., Fried, I., 2018. Human single neuron activity precedes emergence of conscious perception. Nat. Commun. 9 (1), 2057. https://doi.org/10.1038/s41467-018-03749-0.

Hazy, T.E., Frank, M.J., O'Reilly, R.C., 2007. Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. Phil. Trans.: Biol. Sci. 362 (1485), 1601–1613.

Hohwy, J., 2013. The Predictive Mind. Oxford University Press.

Hohwy, J., Seth, A., 2020. Predictive processing as a systematic basis for identifying the neural correlates of consciousness. Phil. Mind Sci. 1 (II) https://doi.org/10.33735/phimisci.2020.II.64. Article II.

Joglekar, M.R., Mejias, J.F., Yang, G.R., Wang, X.-J., 2017. *Inter-areal balanced amplification enhances signal propagation in a large-scale circuit model of the primate cortex* [Preprint]. Neuroscience. https://doi.org/10.1101/186007.

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. Communicat. Biol. 1 (1), 215. https://doi.org/10.1038/s42003-018-0225-1.

Kapoor, V., Dwarakanath, A., Safavi, S., Werner, J., Besserve, M., Panagiotaropoulos, T. I., Logothetis, N.K., 2020. *Decoding the contents of consciousness from prefrontal ensembles* [Preprint]. Neuroscience. https://doi.org/10.1101/2020.01.28.921841.

Kiebel, S.J., Garrido, M.I., Moran, R.J., Friston, K.J., 2008. Dynamic causal modelling for EEG and MEG. Cognitive Neurodynamics 2 (2), 121–136. https://doi.org/10.1007/s11571-008-9038-0.

King, J.-R., Pescetelli, N., Dehaene, S., 2016. Brain mechanisms underlying the brief maintenance of seen and unseen sensory information. Neuron 92 (5), 1122–1134. https://doi.org/10.1016/j.neuron.2016.10.051.

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

Limanowski, J., Friston, K., 2018. 'Seeing the dark': grounding phenomenal transparency and opacity in precision estimation for active inference. Front. Psychol. 9, 643. https://doi.org/10.3389/fpsyg.2018.00643.

Manohar, S.G., Zokaei, N., Fallon, S.J., Vogels, T.P., Husain, M., 2019. Neural mechanisms of attending to items in working memory. Neurosci. Biobehav. Rev. 101, 1–12. https://doi.org/10.1016/j.neubiorev.2019.03.017.

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.

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

Noy, N., Bickel, S., Zion-Golumbic, E., Harel, M., Golan, T., Davidesco, I., Schevon, C.A., McKhann, G.M., Goodman, R.R., Schroeder, C.E., Mehta, A.D., Malach, R., 2015. Ignition's glow: ultra-fast spread of global cortical activity accompanying local "ignitions" in visual cortex during conscious visual perception. Conscious. Cognit. 35, 206–224. https://doi.org/10.1016/j.concog.2015.03.006.

Oizumi, M., Albantakis, L., Tononi, G., 2014. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. PLoS Comput. Biol. 10 (5), e1003588 https://doi.org/10.1371/journal.pcbi.1003588.

O'Reilly, R.C., Frank, M.J., 2006. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural Comput. 18 (2), 283–328. https://doi.org/10.1162/089976606775093909.

Paffen, C.L.E., Alais, D., Verstraten, F.A.J., 2006. Attention speeds binocular rivalry. Psychol. Sci. 17 (9), 752–756. https://doi.org/10.1111/j.1467-9280.2006.01777.x.

Panagiotaropoulos, T.I., Deco, G., Kapoor, V., Logothetis, N.K., 2012. Neuronal discharges and gamma oscillations explicitly reflect visual consciousness in the lateral prefrontal cortex. Neuron 74 (5), 924–935. https://doi.org/10.1016/j.neuron.2012.04.013.

Parr, T., Friston, K.J., 2017. Working memory, attention, and salience in active inference. Sci. Rep. 7 (1), 14678. https://doi.org/10.1038/s41598-017-15249-0.

Parr, T., Markovic, D., Kiebel, S.J., Friston, K.J., 2019. Neuronal message passing using Mean-field, Bethe, and Marginal approximations. Sci. Rep. 9 (1), 1889. https://doi.org/10.1038/s41598-018-38246-3.

Parr, T., Rikhye, R.V., Halassa, M.M., Friston, K.J., 2020. Prefrontal computation as active inference. Cerebr. Cortex 30 (2), 682–695. https://doi.org/10.1093/cercor/bhz118.

Pitts, M.A., Padwal, J., Fennelly, D., Martínez, A., Hillyard, S.A., 2014. Gamma band activity and the P3 reflect post-perceptual processes, not visual awareness. Neuroimage 101, 337–350. https://doi.org/10.1016/j.neuroimage.2014.07.024.

Prinz, J., 2012. The Conscious Brain. OUP, USA.

Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A.R., Schulz, J.B., Fox, P.T., Eickhoff, S.B., 2012. Modelling neural correlates of working memory: a coordinate-based meta-analysis. Neuroimage 60 (1), 830–846. https://doi.org/10.1016/j.neuroimage.2011.11.050.

Sajid, N., Ball, P.J., Parr, T., Friston, K.J., 2021. Active inference: demystified and compared. Neural Comput. 33 (3), 674–712. https://doi.org/10.1162/neco_a_01357.

Salti, M., Bar-Haim, Y., Lamy, D., 2012. The P3 component of the ERP reflects conscious perception, not confidence. Conscious. Cognit. 21 (2), 961–968. https://doi.org/10.1016/j.concog.2012.01.012.

Schlossmacher, I., Dellert, T., Pitts, M., Bruchmann, M., Straube, T., 2020. Differential effects of awareness and task relevance on early and late ERPs in a No-report visual oddball paradigm. J. Neurosci. 40 (14), 2906–2913. https://doi.org/10.1523/JNEUROSCI.2077-19.2020.

Serences, J.T., Ester, E.F., Vogel, E.K., Awh, E., 2009. Stimulus-specific delay activity in human primary visual cortex. Psychol. Sci. 20 (2), 207–214. https://doi.org/10.1111/j.1467-9280.2009.02276.x.

Sergent, C., Baillet, S., Dehaene, S., 2005. Timing of the brain events underlying access to consciousness during the attentional blink. Nat. Neurosci. 8 (10), 1391–1400. https://doi.org/10.1038/nn1549.

Sergent, C., Corazzol, M., Labouret, G., Stockart, F., Wexler, M., King, J.-R., Meyniel, F., Pressnitzer, D., 2021. Bifurcation in brain dynamics reveals a signature of conscious

processing independent of report. Nat. Commun. 12 (1), 1149. https://doi.org/10.1038/s41467-021-21393-z.
Seth, A.K., Hohwy, J., 2021. Predictive processing as an empirical theory *for* consciousness science. Cognit. Neurosci. 12 (2), 89–90. https://doi.org/10.1080/17588928.2020.1838467.
Smith, R., 2016. The relationship between consciousness, understanding, and rationality. Phil. Psychol. 29 (7), 943–957. https://doi.org/10.1080/09515089.2016.1172700.
Smith, R., 2017. A neuro-cognitive defense of the unified self. Conscious. Cognit. 48, 21–39. https://doi.org/10.1016/j.concog.2016.10.007.
Smith, R., Friston, K.J., Whyte, C.J., 2022. A step-by-step tutorial on active inference and its application to empirical data. J. Math. Psychol. 107, 102632. https://doi.org/10.1016/j.jmp.2021.102632.
Smith, R., Lane, R.D., Parr, T., Friston, K.J., 2019. *Neurocomputational mechanisms underlying emotional awareness: Insights afforded by deep active inference and their potential clinical relevance* [Preprint]. Neuroscience. https://doi.org/10.1101/681288.
Soto, D., Mäntylä, T., Silvanto, J., 2011. Working memory without consciousness. Curr. Biol. 21 (22), R912–R913. https://doi.org/10.1016/j.cub.2011.09.049.
Thomas Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., Fischl, B., Liu, H., Buckner, R.L., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. 106 (3), 1125–1165. https://doi.org/10.1152/jn.00338.2011.
Tononi, G., Boly, M., Massimini, M., Koch, C., 2016. Integrated information theory: from consciousness to its physical substrate. Nat. Rev. Neurosci. 17 (7), 450–461. https://doi.org/10.1038/nrn.2016.44.
Tsuchiya, N., Wilke, M., Frässle, S., Lamme, V.A.F., 2015. No-report paradigms: extracting the true neural correlates of consciousness. Trends Cognit. Sci. 19 (12), 757–770. https://doi.org/10.1016/j.tics.2015.10.002.
Vilas, M.G., Auksztulewicz, R., Melloni, L., 2021. Active inference as a computational framework for consciousness. Rev. Philo. Psychol. https://doi.org/10.1007/s13164-021-00579-w.
Westbrook, A., Braver, T.S., 2016. Dopamine does double duty in motivating cognitive effort. Neuron 89 (4), 695–710. https://doi.org/10.1016/j.neuron.2015.12.029.
Whyte, C.J., 2019. Integrating the global neuronal workspace into the framework of predictive processing: towards a working hypothesis. Conscious. Cognit. 73, 102763. https://doi.org/10.1016/j.concog.2019.102763.
Whyte, C.J., Smith, R., 2021. The predictive global neuronal workspace: a formal active inference model of visual consciousness. Prog. Neurobiol. 199, 101918. https://doi.org/10.1016/j.pneurobio.2020.101918.
Wolff, A., Berberian, N., Golesorkhi, M., Gomez-Pilar, J., Zilio, F., Northoff, G., 2022. Intrinsic neural timescales: temporal integration and segregation. Trends Cognit. Sci. 26 (2), 159–173. https://doi.org/10.1016/j.tics.2021.11.007.
Ye, M., Lyu, Y., Sclodnick, B., Sun, H.-J., 2019. The P3 reflects awareness and can Be modulated by confidence. Front. Neurosci. 13, 510. https://doi.org/10.3389/fnins.2019.00510.