A predictive processing account of card sorting: Fast proactive and reactive frontoparietal cortical dynamics during inference and learning of perceptual categories

Francisco Barceló

Laboratory of Neuropsychology, University of the Balearic Islands, Mallorca, Spain

Correspondence: f.barcelo@uib.es

Abstract

For decades a common assumption in Cognitive Neuroscience has been that prefrontal executive control is mainly engaged during target detection (Posner & Petersen, 1990, 3: 25–42, *Ann Rev Neurosci.*). More recently, predictive processing theories of frontal function under the Bayesian brain hypothesis emphasize a key role of proactive control for anticipatory action selection (i.e., planning as active inference). Here, we review evidence of fast and widespread electroencephalographic (EEG) and magnetoencephalographic (MEG) fronto-temporo-parietal cortical activations elicited by feedback cues and target cards in the Wisconsin Card Sorting Test (WCST). This evidence is best interpreted when considering negative and positive feedback as predictive cues (i.e., sensory outcomes) for proactively updating beliefs about unknown perceptual categories. Such predictive cues inform posterior beliefs about high-level hidden categories governing subsequent response selection at target onset. Quite remarkably, these new views concur with Don Stuss' early findings concerning two broad classes of P300 cortical responses evoked by feedback cues and target cards in a computerized WCST analogue. Stuss' discussion of those P300 responses —in terms of the resolution of uncertainty about response (policy) selection, and the subject's expectancies for future perceptual or motor activities and their timing— were prescient of current predictive processing and active (Bayesian) inference theories. From these new premises, a domain-general frontoparietal cortical network is rapidly engaged during two temporarily distinct stages of inference and learning of perceptual categories that underwrite goal-directed card sorting behavior, and they each engage prefrontal executive functions in fundamentally distinct ways.

Keywords: Executive functions, Higher level cognition, Neuropsychology, P300, Prefrontal cortex

Acknowledgements

The author wishes to express his gratitude to Karl Friston for his many edits, clarifying comments, and valuable suggestions on an earlier version of this manuscript. Thanks also to the reviewers and editor of this Special Issue for their thoughtful feedback. Funding support for this work to FB was provided by the Spanish Ministry of Science and Innovation (PID2019-106045GB-I00).

Preprint: In press, Journal of Cognitive Neuroscience (mitpressjournals.org/loi/jocn)

"Perhaps the most widely accepted measure to show executive functional deficits is the sorting task" (Stuss & Benson, 1984, p. 18)

INTRODUCTION

In recent years there has been a paradigm shift in the cognitive neurosciences motivated by views of the brain as a prediction machine, whose working principle is to make active (Bayesian) inferences about the causes of its sensory inputs (Friston, 2005, 2010). In this paper, the implications of this new theory of cortical responses are examined for a deeper understanding of the Wisconsin card sorting test (WCST), one of the most distinctive tests of frontal lobe function (Milner, 1963). In doing so, we hope to showcase the potential of these new views to solve some paradoxes of the frontal lobe riddle—and to resolve long-lived controversies in the literature (Donchin & Coles, 1988; Stuss & Alexander, 2007; Teuber, 1964/2009). As with other shifts in paradigm, predictive processing has led many to update our beliefs and think differently about old findings (Barceló, Perianez, & Knight, 2002), recasting them in the light of the new ideas. In this spirit, we review electroencephalographic (EEG) and magnetoencephalographic (MEG) evidence about the fast neural dynamics underlying two processing stages during WCST performance, as indexed by two broad classes of scalp-recorded P300 cortical responses to informative feedback and target stimuli. These two classes of P3-like responses show distinct scalp topographies along a frontoparietal axis, with discrete contributions from frontal and posterior multimodal association cortices (Knight, 1997). Crucially, these P3-like responses can be explained under a common overarching principle of surprise minimization at either higher (frontal) or lower (nonfrontal) levels in cortical hierarchies (Friston, 2005). Hence, given the relevance of surprise minimization under active inference, the thread of this paper will revolve around the P300 family of cortical responses, one of the most widely used EEG indexes of cognition, which has been linked to surprise and uncertainty resolution ever since its discovery (Donchin, 1981; Sutton, Braren, Zubin, & John, 1965).

The new theory of cortical responses allows us to recast card sorting behavior simply in terms of two temporarily distinct information processing stages of inference and learning of perceptual categories ruling goal-directed action selection (Friston, 2005). Crucially, on this view perception and action are closely intertwined into perception-action cycles (Fuster, 2013), thus reinstating old ideas about reafference and corollary discharge in the neuropsychology of the frontal lobes (Luria, 1966; Teuber, 1964/2009). Further, the information processing demands placed on the subject being assessed do not rigidly depend on the relevant ('attend') or irrelevant ('ignore') task conditions as instructed by the examiner. Instead, demands are flexibly linked to the subject's internal model of the statistical structure of the task and its sensorimotor contingencies (O'Regan & Noe, 2001; Parr, Rikhye, Halassa, & Friston, 2019). These represent major departures from traditional serial processing schemes that consider behavioral responses as the final output of stimulus-feature competition and intermediate cognitive operations (Norman & Shallice, 1986), with little influence on earlier information processes (Barceló & Cooper, 2018b). As a result, predictions from the new theory differ from those of conventional schemes, and offer solid grounds to explain many paradoxical results. For example, that frontal damage does not always impair detection of relevant targets (Knight, 1997; cf., Posner & Petersen, 1990), or that 'irrelevant' distractors and ancillary feedback cues can both tax working memory capacity and engage frontal resources more than 'relevant' target stimuli (Barceló, Escera, Corral, & Periánez, 2006; Barceló & Knight, 2007a).

In the following sections, we describe how the new theory of cortical responses allows us to recast card sorting in terms of predictive action selection (i.e., planning as active inference; Botvinick & Toussaint, 2012; Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2017). In this context, Don Stuss extensive work with frontal lobe patients has been a source of inspiration for our own studies on the fast neural dynamics subserving WCST performance. There is, however, one early and less well-known study where he employed event-related potentials (ERP) to measure the cortical responses in a card sorting analogue of the WCST (Stuss & Picton, 1978). In retrospect, it is remarkable that Don and Terence's discussion of their P3-like responses—to informative feedback cues—in terms of the resolution of uncertainty about *correct response selection*, and as linked to the *contextual updating* in the subject's expectancies for *future perceptual or motor activity*, pre-empted current predictive processing and active inference theories

of perceptual categorization. These new views can elegantly subsume diverse frontal lobe functions such as energization, task setting and monitoring (Stuss & Alexander, 2007), but also executive attention, inhibitory control, working memory and decision making (Fuster, 2019; Stuss & Benson, 1984; Stuss, Shallice, Alexander, & Picton, 1995), all of which are deployed during WCST performance, as well as in many other complex forms of goal-directed behavior that are disrupted by frontal lobe lesions (Luria, 1966).

ACTIVE INFERENCE AND PREFRONTAL EXECUTIVE FUNCTIONS

The active inference and predictive processing frameworks (Clark, 2013; Friston, 2005, 2019; Friston et al., 2017; Hohwy, 2019), together with the revival of enactivism in cognitive science (O'Regan & Noe, 2001), underwrite action-oriented cognition with concepts such as sensorimotor contingencies and perception-action cycles for a full understanding of higher cognitive functions in humans (Fuster, 2013). According to these views, a common ruling principle underlies perception (i.e., sensory state estimation), action (i.e., response policy selection), and learning (i.e., perceptual and reinforcement learning), which is to minimize the same information theory quantity in our brains called the 'free energy'. The free energy principle posits that brains must minimize surprise when sampling sensory data given some internal generative model (Friston, 2010). A generative model is a probabilistic and formal rendition of traditional schemas (Norman & Shallice, 1986; Stuss et al., 1995) and neuronal models (Sokolov, 1963) that have been long used to explain brain and behavioral responses in the exchanges with our surroundings. The basic idea is appealing in its simplicity: the brain sets up a number of competing hypotheses or predictions about the causes of its sensory inputs, and then actively updates these predictions through action on the basis of bottom-up prediction errors. These errors result from a mismatch between what is predicted and what is actually observed, and can thus be thought of as 'surprise signals' being transmitted through ascending connections up the neural hierarchies (Friston, 2019). This recursive exchange of descending predictions and ascending prediction errors results in information transmission (or 'neuronal message passing') between high and low levels across the neural hierarchies. This information exchange evolves dynamically over time and terminates when the generative model is updated (referred to as 'belief updating'), and now encodes the belief that more accurately predicts the hidden cause of those sensations, thus minimizing surprise, as mandated by the free energy principle (Friston, 2005, 2010).

A complementary account of the imperative to minimise surprise (as scored by free energy) is apparent when one appreciates that free energy is also (a mathematical bound on) the log likelihood of sensory input, under our internal or generative model. In statistics, this (marginal) likelihood is known as model evidence, while in machine learning, it is known as evidence bound (Winn & Bishop, 2005). In short, the imperative is to minimise surprise, which is exactly the same as maximizing the evidence for our models of the world—sometimes referred to as self-evidencing (Hohwy, 2016). When considering the consequences of any action on the world, the imperative to minimise surprise becomes nuanced: in other words, the imperative becomes the minimization of uncertainty[1] (i.e., surprise expected following an action). This lends free energy minimization an epistemic aspect, in which most things that we do are in the service of resolving uncertainty about how our sensations are generated (K. Friston, personal communication, 2020).

The simplicity of this basic scheme is also appealing to many of us who have used the Wisconsin card sorting test (WCST; Fig. 1) as a neuropsychological tool for examining prefrontal executive functions in research and clinical contexts. As will be further explained below, this is an open-ended test where the subject (called 'agent' in active inference) is requested to sort a pack of cards without clear instructions about what is the correct course of action. In these settings, agents need to engage in active inference and set up hypotheses to predict which of several perceptual categories—that are 'hidden' in the sensory cues from the cards—will be most likely rewarded by the examiner. Then, the subject actively tests each of her predictions one after another to resolve her uncertainty and disambiguate among plausible alternatives. This is an active and recursive process, where actions generate informative prediction errors, until the

[1] Technically, in information theory, surprise is known as self-information and expected surprise corresponds to entropy. Entropy is a measure of uncertainty. This means that minimising expected surprise corresponds to resolving uncertainty (cf., Sutton et al., 1965).

examiner provides confirmatory 'correct' feedback about the rewarded perceptual category. In this first stage of perceptual inference, which normally takes several trials, subjects use actions in an epistemic or exploratory fashion to disambiguate the rewarded category hidden in the cards (Friston et al., 2017).

Crucially, in active inference, the feedback provided by the examiner has 'epistemic affordance'. In other words, choosing one card or another can resolve uncertainty about the contingencies currently in play, above and beyond the pragmatic value of choosing the correct card. In turn, reducing uncertainty about the card sorting contingencies enables some more confident inference about 'what should I do next?' In short, response selection becomes an integral constituent of recursive perception-action cycles (Friston et al., 2017; Fuster, 2013). Once the subject infers the correct sorting category, the examiner keeps on reinforcing the same category for the next ten cards or so, thus favoring a context where there is little uncertainty about the correct course of action. As the subject becomes increasingly confident about the reward contingencies, her precision[2] in selecting among competing perceptual categories and response policies increases leading to greater behavioral efficiency. At this point, the epistemic (i.e., exploratory) affordance of any policy gives way to the pragmatic (i.e., exploitative) affordance of securing rewards. This second stage reflects 'context learning', when the agent's actions primarily serve a pragmatic or exploitative function (Friston et al., 2016). Here, we contend that these two temporarily distinct stages of inference and learning can be readily identified during WCST performance, and they each engage prefrontal executive functions in fundamentally different ways. Note the correspondence with accounts of executive functions in terms of nonroutine and routine activities, respectively (Stuss et al., 1995).

In this article, existing evidence will be reappraised in the light of this new theory of cortical responses, in the hope of illustrating its potential to explain apparently contradictory findings and to resolve long-lasting dialectics in the literature. An introduction to probability theory and Bayesian inference is beyond the scope of this paper (see Doya & Ishii, 2007). However, it will be useful to clarify some key concepts from the outset. First, in active inference the causes of sensory inputs are called 'hidden' states or variables because they cannot be directly observed, and need to be inferred from the sensory exchanges between an agent and its environment, "which means computing the posterior probability of (unknown or hidden) causes, given observed outcomes" (Friston et al., 2017 p. 7). An example of a hidden variable is the various ways to sort a card depicting two blue circles (i.e., whether the 'correct' response depends upon the color, the form, or the number of elements in the card).

Second, there is a straightforward mathematical correspondence between Bayes theorem and the joint and conditional probabilities used to compute the mutual information between any two variables (Doya & Ishii, 2007). Hence, estimates of average Bayesian surprise being transmitted across frontal and posterior multimodal association cortices during WCST performance may also be expressed as the mutual information between hidden states and sensory outcomes (Friston et al., 2017; cf., Fig. 2). Similar formalisms have been used to estimate mental capacity limits in humans (Koechlin & Summerfield, 2007), and were one of the founding cornerstones of information processing views in cognitive neuroscience (G. A. Miller, 1956).

Third, while active inference is a powerful scheme to examine belief updating and surprise minimization at any level in the neural hierarchies, here we will focus on two broad classes of scalp-recorded P3-like cortical responses with distinct topographies overlying frontal and posterior multimodal association cortices, respectively (Knight, 1997). We will refer to these two classes of cortical responses as 'anterior' and 'posterior' P3-like responses, respectively, or P3a and P3b for short (Polich, 2007; Squires, Squires, & Hillyard, 1975). This anatomical distinction should suffice to match the research aims of many brain lesion studies (Milner, 1963). This also agrees with hierarchical models of prefrontal function that explicitly distinguish between high-level hidden states represented at prefrontal cortices, from low-level states at posterior association cortices (see Fig. 2; cf., Stuss, Picton, & Alexander, 2001).

Finally, the family of P3-like responses has long been regarded as a proxy for surprise minimization, that is, reduction in the uncertainty generated by surprising events (Sutton et al., 1965), but also of learning and memory consolidation (Donchin, 1981). Hence, P3-like responses offer sufficient topographical and

[2] Precision is an important quantity that can be read as the complement of uncertainty (i.e., the precision of a Gaussian probability density is the inverse of its variance or dispersion).

temporal resolution to assess high- and low-level belief updating at frontal and posterior multimodal association cortices, during the inference and learning of perceptual categories in card sorting.

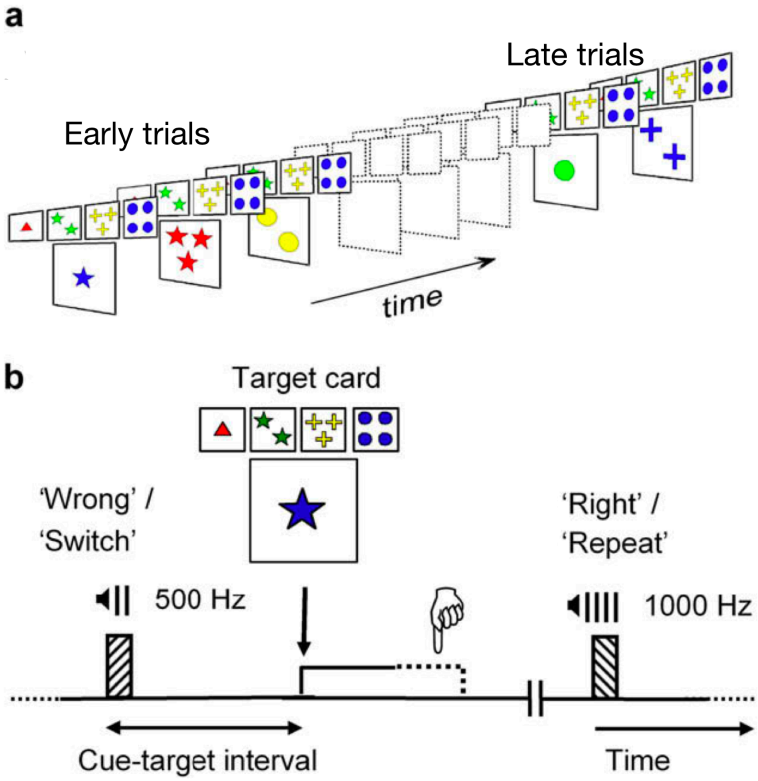

Figure 1. Computerized WCST version adapted for recording cortical responses. (a) Schematic of one card sorting series where early and late trials broadly map onto the stages of inference and learning of perceptual categories. (b) Schematic of one card sorting trial where simple tonal sounds can be instructed either as negative and positive feedback, or as 'switch' and 'repeat' cues informing about probabilistic updates in the policy for responding to the ensuing target card. Adapted from Nyhus and Barceló (2009), with permission.

A GENERATIVE MODEL OF CARD SORTING

In this section, card sorting will be framed in terms of active inference —that is, the updating of the subject's beliefs about how her sensations are caused during WCST performance. Card sorting can be seen as a paradigmatic example of goal-directed uncertainty resolution, whereby perception and action are jointly deployed to reduce contextual uncertainty. This resolution of uncertainty rests upon inferring the perceptual categories 'hidden' in the cards (in the parietal hierarchies), and then using this inference to select a response on the basis of inferred outcomes under each plausible choice (in prefrontal hierarchies). This entails solving the dual problem of (1) inferring some hitherto unknown sorting categories, and (2) learning the statistical structure of the task, that underwrites policy selection (Friston et al., 2017). For this, active inference rests on a generative model of observed sensory outcomes, which is just a context-sensitive, time dynamic and revisable hypothesis about how observed sensory outcomes are generated, while evaluating competing hypotheses about their hidden causes. Crucially, unlike other schemes, these generative models have a hierarchical structure and incorporate actions, responses or decisions. This means that sensory observations depend on actions (e.g., which card you select), which requires the generative model to entertain expectations about sensory outcomes under different action plans or response 'policies' (Friston et al., 2016; Friston et al., 2017).

A great deal of clinical and brain imaging studies on the WCST have been inspired by Milner's (1963) seminal work, and many of them have adopted similar testing materials and procedures. Milner used four keycards differing in color, form, and number: one red triangle, two green stars, three yellow crosses, and

four blue circles. These were placed in front of the patient, who was also given a pack of 128 choice cards that varied along these same dimensions (Fig. 1). The only instruction given to the patient was to match each choice card with one of the four keycards, following some hitherto undeclared sorting principle. Then the patient was to rely on the information provided by the examiner's positive ('right') or negative ('wrong') feedback—in order to discover the hidden sorting rule and try to get as many correct card-sorts as possible. Patients received no further information about how to proceed, and many other important task variables were initially unknown to the patient. Crucially, one such hidden variable was the examiner's policy for reinforcing each of the perceptual categories in the cards, which are the color, the number and the type of forms in the card. Also Milner arbitrarily used the following (hidden) sequence: color - form - number - color - form - number. Likewise, and unknown to the patient, Milner changed the reinforced sorting category after ten consecutive correct sorts—and did so without giving any warning to the patient (Milner, 1963).

Test administration terminated when the patient successfully completed the six sorting categories, or when all 128 choice cards had been placed on top of the keycards. The main test scores were the total number of categories achieved, and the number of perseverative and non-perseverative (i.e., 'set-loss') errors. A perseverative error was scored when a card was sorted using the previously reinforced—but now irrelevant—sorting category, or when there was a continued tendency to respond using one and the same category. All other errors were considered as non-perseverative (Milner, 1963). Note that successful performance in this test requires that patients efficiently infer and learn some critical hidden variables ruling their (sensorimotor) information exchanges with their testing environment.

Over the years, this original version of the WCST has been adapted in many different ways to best examine the componential structure of the information processes underlying brain and behavioral responses in healthy and clinical populations. For instance, Nelson (1976) removed the choice cards that shared more than one attribute with the keycards, thus eliminating response ambiguity and simplifying the scoring of errors. This version consists of two packs of 24 cards each, categories are scored with only six consecutive correct sorts, and patients are told when to change the category, but the actual criterion is not declared. In other versions, patients are informed beforehand of the three sorting categories (Stuss et al., 2000). Clearly, these modifications facilitate inference about sorting categories. In fact, there are nearly as many subtle adaptations of the original test as published studies, and it would be difficult to review them all here. A more cost-effective approach is to examine the basic elements of a generative model of perceptual categorization during card sorting (Rigoli, Pezzulo, Dolan, & Friston, 2017), and see how these elements help us to interpret the frontal and posterior P3-like responses evoked during the two temporarily distinct stages of inference and learning. Put simply, a generative model defines sensory inputs as a function of the underlying hidden causes (i.e., states or variables) in the environment (Friston, 2005, 2010):

$$e = g(v, \theta), \tag{1.1}$$

where e represents the sensory evidence (e.g., exteroceptive, proprioceptive and interoceptive sensory inputs), and $g(v, \theta)$ is a nonlinear probabilistic function that generates sensory inputs from their hidden causes. These causes are represented by a vector v, which is just a list of the unknown or hidden variables in a typical testing situation (i.e., v = variable type and number of sorting categories, variable length of trial sequences, variable inter-stimulus intervals, variable orientation and color of visual features in the cards, variable pitch of 'right' and 'wrong' feedback sounds, variable room luminance, etc.). The generative model is further elaborated to incorporate a hierarchical structure and temporal dynamics, meaning that some of those hidden variables will be disambiguated early by belief updating at low levels in the cortical hierarchy (e.g., primary sensory cortices), whereas other variables will entail a longer-lasting exchange of predictions and prediction errors at higher levels in the hierarchy (e.g., prefrontal and posterior association cortices; Fig. 2), until belief updating is complete and surprise is minimized. The parameters (θ) in the model encode the contingencies or relationships among those variables, and need to be learned through practice (e.g., that the sorting category changes after 10 correct sorts).

From the extant literature, one might predict that prefrontal cortices will be mostly engaged when resolving the uncertainty about the abstract sorting categories, and also in setting up the corresponding response policy (i.e., *task setting*, namely, the sensorimotor mapping between attributes in the card and response selection as made explicit by the keycards; cf., Stuss & Alexander, 2007; Stuss et al., 1995). Further, from conventional views on the anatomy of the executive attention system, one might assume maximal prefrontal engagement when the target card is on display for its appraisal (cf., Posner & Petersen, 1990). In turn, one might be less inclined to expect frontal involvement in response to changes in the sensory attributes of the cards and the feedback, and might dismiss other variables (i.e., room luminance) as largely irrelevant for evaluating frontal lobe functions.

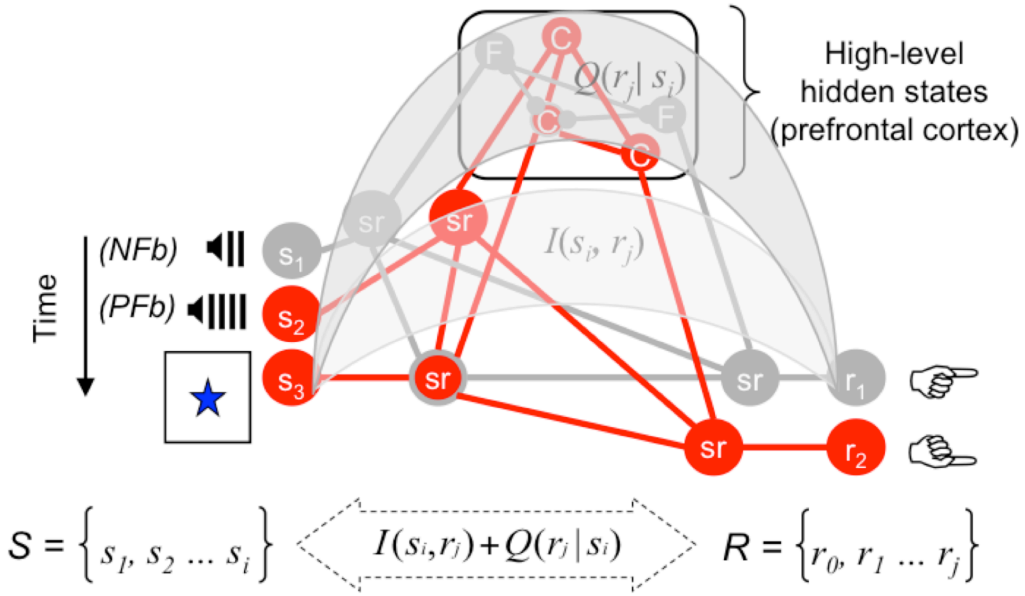

Figure 2. Hierarchical model of prefrontal function (adopted from E. K. Miller & Cohen, 2001). High- and low-level belief updates at frontal and posterior multimodal association cortices can be modeled as the mutual information between hidden states and sensory outcomes transmitted through higher [$Q(r_j|s_i)$] and lower [$I(s_i, r_j)$] levels in the neural hierarchies, respectively (Barceló & Cooper, 2018b). The quantity $Q(r_j|s_i)$ corresponds to the psychological notion of cognitive control (Koechlin & Summerfield, 2007). For simplicity, only two perceptual categories, color (C) and form (F), are illustrated, together with only three stimuli and two motor responses from the pool of all stimuli {S} and responses {R} in our WCST analogue. Red indicates active units or pathways. Small circles represent conflict between two antagonistic high-level units. Negative feedback (*NFb*) cues are very informative because they inform about updates in the hidden perceptual category and corresponding response policy selection. In turn, positive feedback (*PFb*) cues and target cards are comparatively less informative events for response selection. The bidirectional arrow captures the concept of 'reafference', or feedback-feedforward connectivity (Stuss & Benson, 1984), through mutually informed exteroceptive, proprioceptive and interoceptive sensory outcomes (Friston, 2010). Adapted by permission from Nyhus & Barceló (2009).

These are some aspects where hierarchical generative models depart from traditional schemes because predictions and prediction errors—and how high up in the neural hierarchy these surprise signals are transmitted—depend on nonlinear interactions between those hidden variables (Friston et al., 2017). Very roughly, active inference relies on Bayes rule [$p(h|e) = p(e|h)p(h)/p(e)$] to formalise belief updating from a prior belief about a hypothesis, $p(h)$, to a posterior belief, $p(h|e)$, based on the likelihood that the available evidence was generated by that hypothesis, $p(e|h)$. The terms $p(h)$ and $p(h|e)$ are known as 'prior' and 'posterior' probabilities, respectively; and together with the actual evidence, $p(e)$, allow us to quantify Bayesian surprise (e.g., a prediction error[3]) as the divergence between the prior and posterior probabilities

[3] Technically, the Bayesian surprise can be associated with a precision weighted prediction error. In other words, the degree of belief updating in predictive processing schemes is determined by the magnitude of prediction errors weighted by their precision. This means a precise prediction error will have more influence on belief updating as it

(Hohwy, 2019). Below, we show how this simple rule may be iteratively applied on a trial-by-trial basis to narrow down the subject's hypotheses about the hidden sorting category based on evidence from the feedback—and, indeed, beliefs about what to do next, and the requisite task parameters that have to be learned.

To determine which aspects of the testing situation will generate Bayesian surprise signals large enough to engage prefrontal cortices rests on computing the Kullback–Leibler divergence between prior and posterior distributions (Itti & Baldi, 2009). This can be approximated by computing the mutual information between hidden variables and observed sensory outcomes; i.e., the information gained about unobservable hidden states from observable outcomes (Barceló & Cooper, 2018a; Friston et al., 2017; Fig. 2). These estimations of the informativeness of task events are far more fine-tuned than traditional stimulus taxonomies in terms of targets, distracters, feedback cues, and so on. However, to date very few studies have attempted to quantify the information gain of task events during card sorting (cf., Kopp & Lange, 2013; Nyhus & Barceló, 2009). As an alternative, below we present some intuitive examples to illustrate under which testing conditions Bayesian surprise signals (i.e., precision weighted prediction errors) can be large enough to engage prefrontal as opposed to posterior multimodal association cortices during the inference and learning of the sorting categories. For an accurate recording of behavioral and cortical responses, we will assume the examiner employs a computerized version of the WCST for testing the subject (see Fig. 1).

Inference about the Sorting Category.

To infer the sorting category, the subject needs to find out about the response policy (course of action) being rewarded by the examiner. For this, the subject needs to resolve her uncertainty about the current context (i.e., the 'hidden' correct sorting category) in order to know what to do next; and her actions should therefore fulfill exploratory or epistemic imperatives (Friston et al., 2017). The P3a and P3b responses associated with this type of belief updating index rapid evidence accumulation over a scale of milliseconds (Friston, 2005), and engage prefrontal and posterior multimodal association cortices (Barceló et al., 2002; Stuss & Picton, 1978), as part of a more widely distributed neural network (Friston et al., 2017; Parr et al., 2019). Now, let us examine an example of how this inferential process entails the planning of a sequence of perception-action cycles over several card sorting trials.

To start with, the subject needs to rely on her prior knowledge on how to sort things, including cards, together with the scarce instructions from the examiner that there is some 'correct' way of sorting the cards. Then she makes a first educated guess (based upon prior beliefs about the nature of these kinds of tasks) for 'number' and matches the first choice card with the first keycard by pressing button 1 on the keypad. For this the subject has visually scanned the scene, sampling the evidence, and has correctly categorized three perceptual dimensions in the card. This also allows her to estimate a prior probability $p(\mathrm{h}) = 1/3$ for number. She also knows that the probability of a positive feedback is $p(\mathrm{e}) = 1/2$, although the likelihood of this first hypothesis fitting the evidence from feedback, $p(\mathrm{e}|\mathrm{h})$, is rather low. In any case, she must wait to hear the informative first negative feedback to update her belief and discard (inhibit) ‘number’ as a hypothesis. In doing so, her brain computes a posterior probability, $p(\mathrm{h}|\mathrm{e})$, in response to ascending prediction errors sufficiently precise to engage prefrontal cortices (Barceló et al., 2002; Stuss & Picton, 1978; see Fig. 3, left panel). Now the posterior of this first card sort becomes the prior for the next trial, and the perception-action cycle starts again.

In fact, the subject decides to bet for 'form' just after hearing the first negative feedback, and before the second card is on display. Hence, she quickly presses button 3 for form when she sees the second card. This action is like a hypothesis-testing experiment (Friston, Adams, Perrinet, & Breakspear, 2012; Parr, Rees, & Friston, 2018), but it does not disambiguate the correct sorting category yet. Instead, it generates another very informative outcome: a second negative feedback, which also sends off ascending prediction errors to prefrontal cortices to update posterior beliefs (Fig. 4). By now, an attentive and efficient subject, who has kept track of all previous sensorimotor contingencies, will be fairly certain that there is just one remaining category to choose from. Thus, when the third card is displayed, the subject sorts it by color on

---

ascends the cortical hierarchy. In turn, this is usually interpreted in terms of a larger evoked response as measured electrophysiologically (K. Friston, personal communication, 2020).

the fourth pile. Again, this new action does not resolve her residual uncertainty. Only the ensuing first positive feedback confirms the subject's prediction, although this is still likely to generate ascending surprise signals large enough to reach prefrontal cortices (Barceló et al., 2002; Kopp & Lange, 2013; Li, Wang, Du, & Cao, 2018). It is normally the second positive feedback that fully matches the subject's predictions about the hidden sorting category, thus terminating perceptual inference. This moment is marked by the phenomenon of repetition suppression[4] of frontal P3a responses to the second and subsequent positive feedbacks in the card sorting series (Fig. 4). This is reminiscent of the extinction of cortical orienting responses (Sokolov, 1963), and it ensues from the suppression of prediction errors about the newly disclosed perceptual category by top-down predictions from higher cortical regions (Friston, 2005). Note that this 'planning as active inference' fits well with the role of prefrontal cortex in the manipulation and input gating (i.e., updating) of information in working memory (Badre & Nee, 2018; Fuster, 2013). Note also the correspondence between this stage of inference and frontal lobe processes such as 'energization' and 'task setting' (Shallice, Stuss, Picton, Alexander, & Gillingham, 2008; Stuss & Alexander, 2007).

For subsequent cycles of perceptual inference, the first negative feedback will normally evoke larger surprise signals than the second negative and subsequent feedbacks (Barceló et al., 2002). This gradual reduction depends not only on progressive belief updating, and hence, a lesser magnitude of prediction errors for these conditions; but also on interactions with other hidden variables. For instance, while the first negative feedback is unpredictable (i.e., the subject does not know the length of trial series yet), the temporal onset of the next feedback can be easily predicted following the first card sort. Further, a first positive feedback can trigger prediction errors of similar or even larger magnitude than a second negative feedback. This happens when the subject tests the correct category after the first negative feedback, and due to an interaction with a sensory change in the feedback cue from negative to positive (Barceló et al., 2002; Kopp & Lange, 2013; Lange, Seer, & Kopp, 2017; Periáñez & Barceló, 2009). Crucially, the engagement of prefrontal cortices does not depend on the type of feedback as delivered by the examiner, but on the magnitude (precision) of prediction errors during perceptual inference. In other words, it is not the sensory evidence *per se* that determines belief updating—it is the degree to which that evidence calls for a revision of posterior beliefs (that are quintessentially time and context dependent).

This scheme suggests that feedback stimuli are very informative predictive cues that resolve the uncertainty about ongoing predictions—and guide the planning of subsequent actions towards the inference of the hidden sorting category (Botvinick & Toussaint, 2012). In contrast, the target cards are merely the workbench, where those hypotheses are tested through action (Friston et al., 2017), and they are comparatively less salient and informative than feedback cues for inferring the sorting category (Barceló & Cooper, 2018a, 2018b). Note that in this scheme, proprioceptive inputs from actions need to be integrated with exteroceptive (i.e., visual, auditory) inputs from the cards and feedback cues, as they all inform the inference of the same hidden variable. Frontoparietal multimodal association cortices are candidate structures for this multisensory integration (Andersen & Cui, 2009), as they generate top-down predictions in the form of corollary discharges in order to suppress exteroceptive and proprioceptive prediction errors (Friston, Shiner, et al., 2012; Fuster, 2013; Teuber, 1964/2009). In sum, there is a direct correspondence between this iterative belief updating process—to infer the sorting category—and planning as active inference, whereby an agent plans a desired future state of affairs as the joint probability over the available perceptual categories, her goal-directed actions, and the looked-after reward from the examiner (Botvinick & Toussaint, 2012; Rigoli et al., 2017).

Learning of the Sorting Category.

Context learning (that is, the learning of the sorting category) proceeds after inference is completed by high-level belief updating, and once the subject is confident about the course of action being rewarded by the examiner. Now the subject still needs to resolve the uncertainty about the parameters in the generative model. These are the values of hidden quantities that do not change over time, such as the stimulus-response mappings, the length of trial sequences, the inter-stimulus intervals, and so on. This type of

[4] The conventional term 'habituation' is not fully appropriate since it normally refers to a type of non-associative learning, whereas P3-like responses are best described in terms of classical Hebbian associative learning in the cortical hierarchies (Friston, 2010).

uncertainty reduction resolves the subject's ignorance about the probabilistic structure of the task, and enables her actions to fulfill an exploitative or pragmatic function (Friston et al., 2017). In one sense, resolving ignorance about quantities that change slowly calls on the same principles as resolving uncertainty about quantities that change quickly. In what follows, we will explore the notion that resolving uncertainty about model parameters necessarily renders inferences about hidden perceptual categories progressively more precise, and therefore, more evident in belief updating—and electrophysiologically (K. Friston, personal communication, 2020).

The parietal P3b responses associated with this type of long-term belief updating reflect evidence accumulation evolving over slower time scales of seconds or longer (Friston, 2005), and engage posterior multimodal association cortices (Barceló et al., 2002; Fig. 3, right panel), as part of a widely distributed neural network (Friston et al., 2017; Parr et al., 2019). Critically, the same imperative of minimizing surprise also applies to this second stage of learning. Let us examine how this type of uncertainty resolution entails the accumulation of evidence through iterative perception-action cycles over several trials under the same sorting category.

Once the subject has inferred the first sorting category, this variable is parameterized with a fixed value (e.g., the stimulus-response mapping for 'color'), thus providing a stable context for fulfilling the goal rewarded by the examiner. With each new card sort under the correct sorting category, the visual scanning of the scene becomes more and more efficient as the subject learns the spatial arrangement of the keycards. This learning speeds up button presses and improves behavioral efficiency as the subject practices the corresponding stimulus-response mappings (Barceló, Muñoz-Céspedes, Pozo, & Rubia, 2000; Barceló et al., 2002; Figs. 4a,b), and also gradually learns the parameters of her generative model (e.g., length of trial sequences, inter-stimulus intervals, visual attributes in the cards, etc.). Note that residual uncertainty about all these task parameters does not prevent the subject from securing the rewarded behavioral goal by repeating the same response policy. Importantly, now the subject's actions simply confirm predictions about the ongoing sorting category, which results in repetition suppression of P3a and P3b responses to the positive feedback, and in repetition enhancement of P3b responses to the target cards (Fig. 4a,b). While the former phenomenon reflects suppression of prediction errors to the positive feedback, the enhancement of target P3b responses reflects a gradual increase in the precision of predictions at higher levels in the cortical hierarchy, as more and more evidence about task parameters is accumulated within a card sorting series and in successive series (Auksztulewicz & Friston, 2016).

Thus, an efficient subject will become gradually more confident about the task parameters as she successfully infers and learns more sorting categories. Task parameters are encoded at different levels in the neural hierarchy in the form of probability distributions represented by their mean and precision (the inverse of variance). The higher the precision the narrower the distribution is around the mean (Friston et al., 2017; Hohwy, 2019). This may eventually lead the subject to become quite confident that, for example, there are only three sorting categories to choose from, that the category will change every 10 correct sorts, that there is a constant delay between the feedback and the card, and so on. In other words, the precision of posterior expectations about many hidden variables will gradually increase over the course of testing (Auksztulewicz & Friston, 2016). In spite of this, the subject cannot predict which of the many target cards will follow the feedback, nor which button press will have to be selected to close the perception-action cycle under the current sorting category. Hence, top-down predictions cannot suppress prediction errors to the onset of target cards, which is why these evoke target P3b responses over posterior association cortices. Crucially, this explains why these target P3b responses never 'habituate' (Donchin, 1981), and instead show repetition enhancement, indexing a gradual increase in the precision of predictions about many task parameters being learned along the card sorting series (Auksztulewicz & Friston, 2016; Figs. 4 and 5).

Note that this stage of context learning may proceed with less belief updating in frontal cortices during the maintenance of information in working memory. This is consistent with evidence that target P3b responses are largely preserved following prefrontal damage (Barceló & Knight, 2007a; Knight, 1997), given that output gating of information from working memory relies more on posterior association cortices and related subcortical structures (Badre & Nee, 2018; Stuss & Alexander, 2000). Note also the correspondence of this second stage of learning with processes such as 'monitoring' (Shallice et al., 2008; Stuss & Alexander, 2007; Stuss et al., 1995).

Dynamic balance between Inference and Learning

From the foregoing description, one might be tempted to associate the first stage of inference to the negative feedback early in the card sorting series, and the second stage of learning to the target cards later on in the series (Figs. 3 and 4). However, this is another aspect where active inference departs from conventional theories of frontal lobe function. Crucially, in active inference the dynamic balance between inference (information exploration) and learning (information exploitation) does not depend solely on the type of stimulus or task condition as defined by the examiner (cf., Posner & Petersen, 1990). Instead, this balance depends on the magnitude of the ascending surprise signals that will ultimately result in either high-level or low-level belief updating—engaging either prefrontal or posterior multimodal association cortices, respectively (Maisto, Friston, & Pezzulo, 2019).

In other words, whether or not an agent's generative model will undergo high-level belief updating at prefrontal cortices does not depend solely on the type of stimulus that generated the sensory inputs, such as a target card, a distracter stimulus, a positive or negative feedback, and so on. As will be shown below, any of these stimuli can potentially generate both anterior P3a and posterior P3b responses, indexing the inference and learning of perceptual categories, respectively. Hence, the critical question is what is the threshold magnitude of surprise signals above which they will engage prefrontal cortices? In active inference the magnitude of those prediction errors, and how high up in the neural hierarchy they penetrate, hinge upon dynamic trial-by-trial updates in the subject's generative model (Figs. 4 and 5).

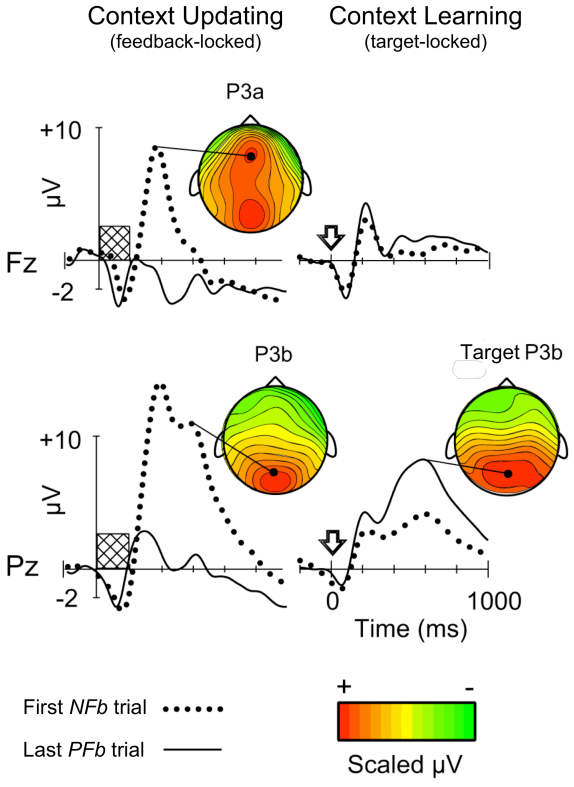

Figure 3. Cortical responses to feedback cues and WCST target cards. Grand-averaged ERPs time-locked to feedback cues (shaded rectangle) and target cards (wide arrow) are displayed for first negative feedback (*NFb*) trials and last positive feedback (*PFb*) trials in a card sorting series, at mid-frontal (Fz) and mid-parietal (Pz) regions. Voltages are in microvolts (μV). Scalp potential maps are shown for mean P3a and P3b responses to first *NFb* cues, and for mean target P3b responses to the last correct target card in the series. The color scale is in normalized units. Early *NFb* trials foster perceptual inference, whereas late *PFb* trials foster context learning. Adapted by permission from Barceló, Periáñez and Knight (2002).

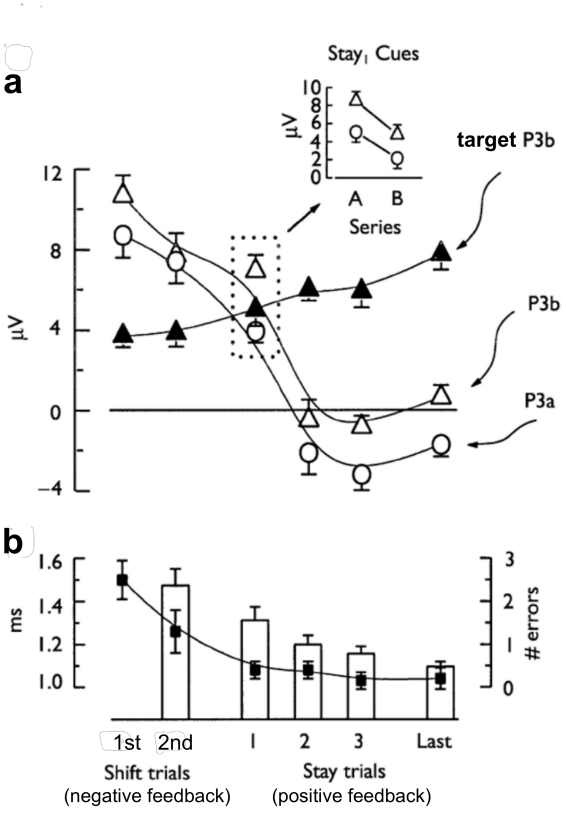

Figure 4. Cortical and behavioral responses to negative and positive feedback trials in a card sorting series. (a) Group-averaged mean ±*SEM* amplitudes of feedback-locked P3a and P3b, and target-locked P3b responses plotted across negative feedback (shift) and positive feedback (stay) trials. Mean P3a and P3b amplitudes were measured from mid-frontal (Fz) and mid-parietal (Pz) scalp regions, respectively. (b) Mean ±*SEM* reaction times from efficiently completed WCST series without errors (solid squares), and mean ±*SEM* number of set-loss errors from failed series (bars) are shown during the inference and learning of the sorting category. Adapted by permission from Barceló, Periáñez and Knight (2002).

These belief updates do not depend solely on the sensory evidence, like the mean stimulus probability (Duncan-Johnson & Donchin, 1977), but they are a function of the likelihood of sensory outcomes given their expected hidden causes, and the prior probability of those causes (Friston et al., 2017). Further, the dynamic balance between inference and learning also depends on the interactions between concurrent hidden variables during WCST performance, such as the number of sorting categories, the length of the card sorting series, inter-stimulus intervals, and so on. All of these factors determine the precision of posterior beliefs. For example, a precise belief that the sorting rule changes on every 10th trial is very different from an imprecise belief that the rule might change at some point in the future. How these factors determine the precision of ascending surprise signals, and whether there may be an information threshold above which these signals may potentially engage prefrontal cortices will be addressed next.

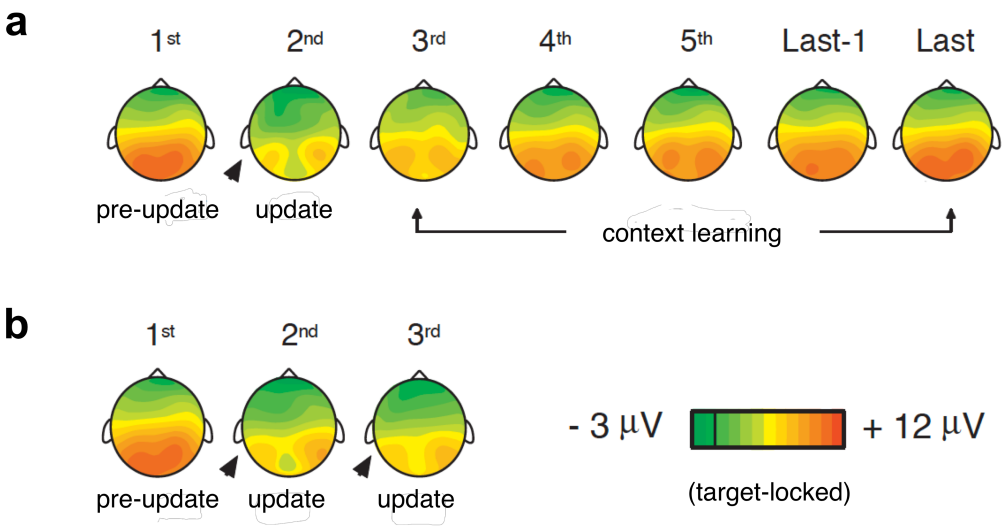

Figure 5. Voltage maps of mean target P3b amplitudes during the inference and learning of perceptual categories. Arrowheads mark target trials preceded by a negative feedback (i.e., update trials). (a) Series with only one negative feedback trial. (b) Series with two negative feedback trials (post-update trials not shown). Card sorting series (a) and (b) evoked similar repetition enhancement of target P3b responses after the first positive feedback. Note the large belief update to the positive feedback in the third trial of series (a) indexing the correct inference of the sorting category. Adapted by permission from Barceló et al. (2000).

COGNITIVE FLEXIBILITY IN ACTIVE INFERENCE

The WCST has long been considered as a gold standard for the assessment of cognitive flexibility (Diamond, 2013; E. K. Miller & Cohen, 2001), which may be compromised mostly in two ways: one, when repeating a sorting category following negative feedback, which is scored as a perseverative error; and second, when switching the sorting category following a positive feedback, which is scored as a set-loss error. In her seminal study, Milner (1963) concluded that patients with frontal lesions were more prone to perseverative errors than patients with nonfrontal lesions. Even though the *specificity* of WCST error scores as markers of frontal lobe function has been questioned on several grounds (Barceló & Knight, 2002; Nyhus & Barceló, 2009), their *sensitivity* to frontal damage has also been confirmed many times over the years (Demakis, 2003; Glascher, Adolphs, & Tranel, 2019; Stuss et al., 2000), thus lending support to Milner's seminal findings. Many metabolic brain imaging studies have examined the neural correlates of cognitive flexibility using WCST analogues (Buchsbaum, Greer, Chang, & Berman, 2005; Konishi et al., 1998; Wang, Cao, Cai, Gao, & Li, 2015). Notably, lesion and brain imaging studies do not inform about the fast neural dynamics of inference and learning of perceptual categories, nor about their subtle trade-off during WCST performance. Here, we contend that knowledge about the fast dynamics of

inference and learning of the sorting categories can improve our understanding of specific disruptions in information processes underlying the commission of WCST errors (Barceló, 1999; Kopp & Lange, 2013; Lange, Seer, & Kopp, 2017; Lange, Seer, Muller-Vahl, & Kopp, 2017).

In the literature on executive functions, the stages of inference and learning of perceptual categories can be broadly assimilated within the cognitive processes of task-set switching and task-set maintenance, respectively (Barceló, 2003; Diamond, 2013; Gajewski, Ferdinand, Kray, & Falkenstein, 2018; Glascher et al., 2019; Shallice et al., 2008). From traditional views, the information processing demands in a cognitive task are typically linked to nominal stimulus taxonomies as defined by the testing conditions and experimental procedures (i.e., a target card, a distractor, a negative feedback, switch and repeat cues, etc.). For instance, a larger frontal engagement for processing target stimuli is generally assumed (Posner & Petersen, 1990). In turn, the subject's motor responses and the ensuing feedback signals are often considered as the end-product of earlier stimulus-feature competition and intermediate cognitive operations (Barceló & Cooper, 2018b). In sharp contrast, in active inference proprioceptive inputs from actions are an integral part in the recursive exchanges of information to disambiguate the hidden causes of sensations (see Fig. 2). Hence, in active inference, cognitive flexibility depends on a dynamic balance between inference and learning (Maisto et al., 2019). Let us now consider some intuitive examples to illustrate how the brain can rapidly alternate between context learning and perceptual inference, even midtrial, in the relatively well structured conditions of WCST performance.

Whereas perceptual inference can be mainly linked to the negative feedback, a positive feedback can also trigger large prediction errors that elicit frontal P3a responses, especially when there is residual uncertainty about the hidden category (Cunillera et al., 2012; Kopp & Lange, 2013). Likewise, in a volatile context with frequently changing sorting categories, target cards may also elicit an anterior P3a indexing perceptual inference (Adrover-Roig & Barceló, 2010; Barceló, Sanz, Molina, & Rubia, 1997; Kopp, Lange, Howe, & Wessel, 2014). This can be best seen in the first target following a predictive cue (Barceló et al., 2006; Brydges & Barceló, 2018), or with WCST analogues that employ more than three sorting categories and short trial sequences, which forces the subject to adopt an epistemic response policy that prevails over context learning (Kopp et al., 2014; Stuss & Picton, 1978). Such an epistemic policy can be fostered with WCST versions that use ambiguous choice cards that share more than one stimulus dimension with the keycards (Milner, 1963; Stuss et al., 2000). This is because ambiguous choice cards reduce the efficacy of the subject's actions to infer the correct sorting category (Fig. 1a).

Moreover, nonlinear interactions between hidden variables, such as the number of sorting categories and correct trial length, may explain some adjustments in the timing of perceptual inference. For instance, if midway through the testing session the subject learns and becomes confident (i.e., forms precise posterior beliefs) about the constant length of trial sequences, then eventually she may correctly anticipate an upcoming change in category. In this case, high-level belief updating will be endogenously generated following the last correct card sort (Barceló et al., 2000). An extreme example is when a mischievous subject decides to sort the cards in the pile that matches none of the perceptual categories (Barceló et al., 2000). Here, one would not expect to see a frontal P3a to the continual negative feedbacks, because they are perfectly predicted by the subject. In turn, target cards will continue to elicit parietal P3b responses, as their contents cannot be predicted by the subject. In fact, always sorting on the non-matching pile provides a stable context for learning under the same response policy, even though the examiner might find the subject's strategy highly surprising. This speaks to the importance of the subject's generative model to estimate trial-by-trial changes in the *subjective surprise* associated with a sensory outcome given its hidden causes (cf., Duncan-Johnson & Donchin, 1977; Friston et al., 2017; Itti & Baldi, 2009). These examples illustrate that the dynamic balance between inference and learning depends on the precision of prediction errors. This can vary from trial to trial as a function of the subject's prior beliefs, the sensorimotor contingencies between task events, and the interactions between concurrent hidden variables.

In healthy subjects, cognitive flexibility hinges on an adaptive and context-sensitive balance between perceptual inference and context learning, or between the exploration of novel information versus the exploitation of familiar information (Ebitz, Sleezer, Jedema, Bradberry, & Hayden, 2019). Next, we use some more examples to show how a well-practiced subject that has been efficiently sorting cards for a while, will quickly revert her pragmatic policy into an epistemic policy in the face of novel, unexpected,

or surprising information. For instance, if unknown to the subject the examiner slips a grey card among the regular color cards, or suddenly rings a bell, these events will trigger perceptual inference and the accompanying frontal P3a responses—indexing large ascending surprise signals, very much like those elicited by the negative feedback (Barceló et al., 2006; Wessel & Aron, 2017).

The subject might even stop sorting and turn towards the examiner asking for help on how to sort grey cards (Wessel et al., 2016). Similar P3a responses are elicited if the keyboard button gets stuck and does not lead to the expected outcome and the closure of the perception-action cycle. Further, a very gradual change in room luminance may go unnoticed; but if the room lights are suddenly switched off, again the subject will likely stop sorting and turn her head around looking to disambiguate the hidden cause of sudden darkness. All these are common examples of epistemic affordances that elicit orienting responses to salient sensory cues (Sokolov, 1963). In other words, the very small likelihood of any such unexpected sensory outcomes, under the current hypothesis, generates precise prediction errors that engage prefrontal cortices (Barceló & Knight, 2007a; Friedman, Cycowicz, & Gaeta, 2001). Ignoring such surprising changes in the environment, and to perseverate sorting the cards, could be considered as a clear sign of cognitive inflexibility[5]. In other words, perceptual inference of unexpected and highly informative distracters introduces a change in context that interrupts ongoing behavior, and leads to a reevaluation of current goals and goal-directed action selection (Maisto et al., 2019; Rigoli et al., 2017).

Finally, active inference could potentially offer a parsimonious explanation for both perseverative and set-loss errors as lapses in control due to transient information overload, either during inference of the sorting category or during context learning, respectively (Dehais et al., 2019). A perseverative error could thus result from a highly precise prior belief that has accumulated over several trials under the previous sorting category (Friston et al., 2016). Conversely, a set-loss error could be due to the low precision of (or confidence in) posterior beliefs for the newly acquired sorting category in the face of distracting sensory inputs. From these premises one would predict perseverative card sorts to evoke large posterior P3b responses typical of context learning; and conversely, card sorts with set-loss errors to evoke frontal P3a responses typical of perceptual inference. Indeed, this was the main pattern of results found when comparing P3-like responses to target WCST cards resulting in either type of error (Barceló, 1999). More recently, the amplitude of both P3a and P3b responses to negative feedback cues have been shown to be inversely correlated with the number of perseverative errors in Parkinson disease patients (Lange, Seer, & Kopp, 2017; Lange et al., 2016). These findings are consistent with the implication of the domain-general frontoparietal control network in the regulation of both perseverative and set-loss errors (Dehais et al., 2019). In sum, perseverative and set-loss errors could be seen as transient information overloads in the subject's exchanges with her testing environment either during perceptual inference or during context learning, respectively (Ebitz et al., 2019; Friston et al., 2017).

WHEN NEUROPSYCHOLOGY MEETS ELECTROPHYSIOLOGY

Neuropsychologists have traditionally relied on structural and functional brain imaging techniques to explore the neural substrates of cognition (Stuss & Alexander, 2007). In turn, electrophysiological techniques offer split-second temporal resolution into the fast neural dynamics of the inference and learning stages of perceptual categorization, as well as sufficient topographical specificity to discern anterior P3a from posterior P3b responses (Knight, 1997). In a series of ERP studies, we used a computerized version of the WCST to examine the neural dynamics underlying the inference and learning of perceptual categories. One common theme in these studies was that negative and positive feedback stimuli were regarded as predictive cues that proactively prompted for a switch, or repetition, in the ongoing sorting category in anticipation of the next card sort (Barceló, 2003; Barceló et al., 2006; Barceló et al., 2002; Garcia-Garcia, Barceló, Clemente, & Escera, 2011; Periáñez & Barceló, 2009). In one early study, Barceló, Periáñez and Knight (2002) used the 24 choice cards that can be unambiguously matched with the four keycards based on just one perceptual dimension (cf., Nelson, 1976). These 24 choice cards were semi-randomly arranged into 18 series, each ruled by a different sorting category. Each series

[5] In active inference, there is an intimate relationship between cognitive flexibility and the deployment of executive attention. This follows because attention is thought to be mediated by setting the precision of various beliefs at different levels in the cortical hierarchies. Please see Parr & Friston (2017), and Parr et al (2019) for details (K. Friston, personal communication, 2020).

contained between six and eight target cards so as to provide a stable context for learning, and also to prevent the anticipation of the next series. A fixed cue-target interval of 1600 msec allowed sufficient preparation time—and minimized the effects of temporal uncertainty upon target detection. Target cards were displayed for 1500 msec, or until a response was issued, and inter-trial intervals between button presses and the next feedback varied randomly between 1500-2000 msec (Barceló, 2003; Fig. 1b).

With these task parameters, healthy subjects can comfortably sort the cards with high accuracy rates. The correct sorting category was initially unknown to the subject, and changed randomly from one series to the next, so subjects were forced into a perceptual inference stage, and they had to rely on the negative and positive feedback to disambiguate the hidden sorting category. Our findings revealed two broad classes of P3-like responses (Figs. 3 and 4): A frontal-central P3a and a posterior P3b that were time-locked to negative feedback stimuli early in the card sorting series, during perceptual inference, and showed rapid repetition suppression (Fig. 3, left panel). Then, a posterior P3b time-locked to the target cards that showed repetition enhancement during context learning later on in the card sorting series, whenever the context remained stable for a sufficient number of trials (Fig. 3, right panel). In spite of major differences in testing materials and procedures, similar findings have been consistently reported by many authors using WCST analogues (Adrover-Roig & Barceló, 2010; Cunillera et al., 2012; Kopp & Lange, 2013; Kopp et al., 2014; Kopp, Tabeling, Moschner, & Wessel, 2006; Lange, Seer, & Kopp, 2017; Lange et al., 2016; Lange, Seer, Muller-Vahl, et al., 2017; Li et al., 2018; Vila-Ballo et al., 2015).

For some time, we thought that our treatment was an original and intuitive, if unconventional, way to interpret these two classes of P3-like responses to the feedback cues and target cards. In those days, the prevailing ideas about the type of P3-like responses elicited by task relevant (target) and task irrelevant (distracter) stimuli in cognitive tasks—together with mainstream ideas about the anatomical substrates of the anterior and posterior attention systems—clashed with the interpretation of our findings. In short, we were reporting a gradually enhanced parietal P3b upon detection of 'relevant' target WCST cards, whose processing was then assumed to be mediated by the anterior executive system (cf., Posner & Petersen, 1990). Moreover, our frontal-central P3a to ancillary negative feedback cues resembled novelty P3 potentials to task irrelevant distracters in simple oddball tasks (Friedman et al., 2001; Polich, 2007). As the reader can imagine, it was not an easy enterprise to convince reviewers about the soundness of our evidence, in the face of prevailing hypotheses about frontal lobe function and the putative roles of these two classes of P3-like responses (Barceló et al., 2002; Donchin & Coles, 1988).

It was in this historical context when we were fortunate enough to meet Don Stuss for the first time: he had kindly accepted our invitation as keynote speaker in a conference we organized in Mallorca in 2004. In his talk, entitled "Reflections on the value of frontal lobe tests: suggestions for the future", he offered an overview of his work on brain lesioned patients assessed with several traditional tests of frontal function (such as the WCST, TMT, Stroop and Verbal Fluency tasks). In their well-known patient study on the WCST, Stuss and collaborators (2000) largely replicated and extended Milner's (1963) seminal findings: apart from confirming the sensitivity of the WCST to left and right dorsolateral prefrontal lesions, they also reported that a group with lesions in superior medial frontal cortex was the most impaired on nearly every measure. In the ensuing discussion, while admitting that such a multifactorial test as the WCST is unlikely to be sensitive only to frontal lobe lesions, he advised the analysis of the component processes involved in order to understand why and how lesions in different brain regions could lead to different types of errors in this test. In his opinion, the WCST could still provide a reasonable index of frontal lobe functioning, if there was control of some of the nonfrontal processes involved in WCST performance (Stuss et al., 2000). When do these nonfrontal processes intervene, and how they can be isolated from frontal processes remain pressing questions nowadays, and thus the contents of his talk are still very much up-to-date. In the poster session after the conference, Don showed a special interest in one of our ERP studies on the WCST. Initially, we thought the reason for his interest was that the study was co-authored by his friend and colleague Bob Knight (Stuss & Knight, 2013). Then, pointing to the figure with the P3-like responses elicited by the negative feedback (Fig. 3, left panel), he asked this rather cryptic question with his warm and friendly smile:

- *So why do you label those two peaks as P3a and P3b?*

This question revealed in Don a much deeper knowledge in cognitive electrophysiology than the typical neuropsychologist has, and that was how we found out about his earlier ERP study using an analogue of the WCST as part of his PhD thesis (Stuss & Picton, 1978). In this study, Don and Terence reported P3-like responses to negative feedback cues and target cards showing a similar timing, morphology, and scalp topography to those reported in our 2002 study. Even though they labeled their feedback-locked peaks as P3 and P4 respectively, we both agreed that these P3-like responses were related to the resolution of uncertainty about the choice of *upcoming motor responses* in the next card sort, and that they indexed functionally different neural operations each (Barceló et al., 2002; Stuss & Picton, 1978). Likewise, we both agreed that the growing parietal target P3b to the cards indexed rehearsal and overlearning of the same sorting rule over several correct card sorts, even though in their study they did not find a significant enhancement in their target P3b amplitudes probably due to their short trial series (cf., Stuss & Picton, 1978). At the time, neither of us could call on active inference to help us interpret those two classes of P3-like responses. Also none of us had quantified the high- and low-level Bayesian surprise associated with our feedback cues and target cards (Barceló & Cooper, 2018b; cf., Fig. 2), nor how this surprise changes dynamically on a trial-by-trial basis as a function of interactions between hidden variables such as the length of trial series, or the transition probabilities between the sorting categories (Maheu, Dehaene, & Meyniel, 2019).

Looked at from the perspective of active inference, these two ERP studies on the WCST offer some valuable insights into prefrontal executive functions. First, there is instant neural activation across a widely distributed fronto-temporo-parietal cortical network in response to the feedback cues, with a lesser frontal involvement in response to the ensuing target cards. Second, the feedback-locked P3a and P3b responses seem to index two distinct predictive processing operations in preparation for the next card sort, given their distinct timing, scalp topographies, and response to task variables. The candidate operations are (a) the inference of (or resolution of uncertainty about) the hidden sorting category, and (b) the updating of stimulus-response mappings in preparation for the next card sort, respectively (cf., Barceló et al., 2002; Stuss & Picton, 1978). In particular, the frontal P3a could be seen as an index of high-level surprise minimization during perceptual categorization (cf., Rigoli et al., 2017). This agrees with views of P3a as a cortical orienting response to novel percepts (Polich, 2007; Sokolov, 1963). Third, the parietal P3b to the target cards seems to index low-level surprise minimization over posterior association cortices (as subjects cannot predict the contents of the card), and it shows repetition enhancement as the subject learns the task context over several trials (Fig. 5). Finally, these P3-like responses could be regarded as domain-general signatures of perceptual inference and learning in many cognitive tasks (Friston, 2005).

In the following sections, we will answer Don's question in the light of evidence supporting a functional and anatomical dissociation between the stages of perceptual inference and context learning during card sorting. Most studies used analogues of the WCST, or related task switching paradigms (Kopp & Lange, 2013; Kopp et al., 2006), and they normally measured P3-like responses to both feedback cues and target cards. As shown below, the active inference framework offers a radically different, although much more flexible, dynamic and richer account of P3a and P3b responses than conventional views in terms of task-irrelevant ('ignore') and task-relevant ('attend') stimulus conditions (Donchin & Coles, 1988).

FAST NEURAL DYNAMICS OF INFERENCE AND LEARNING OF THE SORTING CATEGORIES

From conventional views on the anatomy of executive attention (Posner & Petersen, 1990), we initially searched for an ERP correlate of frontal function in response to the target cards (Barceló, 1999; Barceló et al., 2000; Barceló & Rubia, 1998; Barceló et al., 1997). These studies followed prior work by Mates et al (1991), who measured several slow cortical potentials, including P300, to WCST choice-cards, keycards, and the feedback cues. They compared early and late correct trials in each card sorting series, as well as early incorrect trials that lead to the inference of the category. Quite unexpectedly at the time, target P3b responses were neither sensitive to task conditions nor discriminated between healthy controls and schizophrenic patients. In contrast, negative and positive feedback cues early in the WCST series elicited large frontal P3a and parietal P3b responses that were much reduced in the patients. These effects were interpreted in terms of impaired context updating, and "inability to use feedback information to modify subsequent behavior" (Mattes et al., 1991, p. 203). Thus, even though these authors considered the feedback "at the end of the trial", they also assumed it could modify subsequent behavior. Importantly,

their frontal-central P3a to the feedback was regarded as an index of context updating, in spite of prevailing views linking context updating to the 'task relevant' and parietally distributed target P3b (Donchin & Coles, 1988).

In order to further explore earlier findings of conspicuous parietal target P3b responses to the WCST cards (Barceló & Rubia, 1998; Barceló et al., 1997), the study of Barceló et al. (2000) compared early and late card sorts within the WCST series with two control conditions. They found that the gradual build-up in target P3b amplitude along the WCST series was related to the learning of the sorting category after its endogenous inference following negative feedback (see Fig. 5). In turn, when the category was explicitly disclosed with the first card, or when subjects always sorted cards on the non-matching pile (thus generating continual but fully predictable negative feedbacks), both early and late target cards elicited a full-blown parietal P3b. These findings agree with the hypothesis that the repetition enhancement of target P3b reflects a gradual increase in the precision of endogenous predictions as the subject becomes gradually more confident about the ongoing sorting category, and more evidence about task parameters is accumulated along the card sorting series (Auksztulewicz & Friston, 2016). Importantly, the repetition enhancement of target P3b may go easily unnoticed when all target trials are averaged together (Mattes et al., 1991), or when the sorting category changes frequently, as a volatile context hampers learning (Kopp et al., 2014; Lange, Seer, Muller-Vahl, et al., 2017; Periáñez & Barceló, 2009; Stuss & Picton, 1978). Hence, the repetition enhancement of target P3b is best observed when ERP waveforms are obtained from long card sorting series without any errors (cf., Barceló et al., 2000; Fig. 5).

The insensitivity of target P3b to cognitive variables triggering high-level belief updating, and even to prefrontal damage (Barceló & Knight, 2007a; Knight, 1997), eventually led most researchers to focus on the feedback-locked P3a and P3b responses (Barceló, 2003; Cunillera et al., 2012; Garcia-Garcia et al., 2011; Kopp & Lange, 2013; Kopp et al., 2014; Kopp et al., 2006; Lange, Seer, & Kopp, 2017; Lange et al., 2016; Lange, Seer, Muller-Vahl, et al., 2017; Li et al., 2018; Periáñez & Barceló, 2009; Stuss & Picton, 1978; Vila-Ballo et al., 2015). In general, these studies are consistent with the hypothesis that P3a indexes the inference of the sorting category (i.e., high-level belief updating), whereas P3b indexes the updating of stimulus-response mappings in preparation for the next card sort. Note that there is a consistent delay between the peak latency of P3a (aprox. 250-350 msec) and P3b (aprox. 400-800 msec)[6]. This speaks of two temporally distinct stages of perceptual inference and context learning, with inference preceding and being a prerequisite for learning (Barceló & Cooper, 2018b; Friston, 2005). Accordingly, frontal P3a responses would index surprise minimization over hidden perceptual categories (e.g., response policies), whereas parietal P3b responses would index surprise minimization over task parameters (e.g., stimulus-response mappings) (Friston et al., 2017).

Source localization and MEG studies suggest that the early P3a aspect recruits activity from a distributed fronto-temporo-parietal network with key frontal nodes at the inferior frontal gyrus, middle frontal gyrus, anterior insula and anterior cingulate cortices (Bayless, Gaetz, Cheyne, & Taylor, 2006; Diaz-Blancat, Garcia-Prieto, Maestú, & Barceló, 2018; Periáñez et al., 2004), as well as the hippocampus for retrieval of the new sorting category (Knight, 1996), and the basal ganglia for inhibition of the old response policy and the selection of a new one (Lange et al., 2016; Wessel et al., 2016). In turn, the later P3b aspect recruits activity from the temporoparietal junction, with key nodes at the supramarginal gyrus, superior temporal gyrus, inferior parietal lobe and precuneus (Bayless et al., 2006; Diaz-Blancat et al., 2018; Periáñez et al., 2004). These structures map well onto the functional anatomy of active inference (Friston et al., 2017), and are broadly consistent with metabolic brain imaging studies of the WCST (Buchsbaum et al., 2005; Konishi et al., 1998; Monchi, Petrides, Petre, Worsley, & Dagher, 2001). In sum, the evidence supports the implication of a domain-general frontoparietal network in the elicitation of feedback-locked P3a and P3b responses during perceptual inference, compatible with that engaged by

[6] The earlier latency of P3a has been attributed to bidirectional hyperdirect prefrontotectal pathways (Barceló & Knight, 2007a, 2007b; Wessel et al., 2016). These hyperdirect pathways convey fast prior information about the spatio-temporal context of incoming sensory signals ascending through regular geniculocortical pathways. In other words, prefrontal cortices hold prior beliefs about whether the incoming sensory evidence belongs to the immediate perceptual context —and is thus part of active working memory—, or else, whether it is something completely unexpected that calls for new perceptual inference and the elicitation of an orienting response (Sokolov, 1963). For a review of frontal connectivity with tectal, thalamic, and other brainstem structures, see Stuss & Benson (1984).

surprising distracters in simpler categorization tasks, such as oddball tasks (Barceló et al., 2006; Bledowski et al., 2004).

Methodological, procedural, and technical differences between ERP studies can explain even gross differences in the timing, morphology, and scalp topography of feedback-locked P3-like responses. This is because the magnitude of ascending surprise signals, and hence, the relative engagement of frontal and posterior multimodal association cortices—during actively inferring the sorting category—critically depends on the inference and learning of other states and task parameters, such as the number and length of sorting categories, inter-stimulus intervals, and so on. In spite of this, there are remarkable consistencies across studies. For instance, when switch and repeat pre-cues are interspersed with feedback post-cues, similar P3a and P3b responses are elicited by the switch cues, the negative feedback, and the first positive feedback because all these stimuli help resolve the subject's uncertainty about the same hidden sorting category (Cunillera et al., 2012; Kopp & Lange, 2013; Vila-Ballo et al., 2015).

In general, the larger the number of sorting categories to be disambiguated, the larger and more frontally distributed the P3a responses (Barceló et al., 2006; Barceló et al., 2002; Kopp & Lange, 2013). In turn, larger P3b responses are elicited with more complex tasks and stimulus-response mappings (Barceló & Cooper, 2018a; Kopp & Lange, 2013). However, frontal P3a and parietal P3b may also be elicited by pre-cues and post-cues that merely prompt for a change in the hand used for sorting the card (Kopp et al., 2006). This evidence concurs with active inference, in that simple sensory cues may result in high-level belief updating whenever these resolve the subject's uncertainty about binary response policies conveying maximal sensorimotor ambiguity (Friston, 2010; Kopp & Lange, 2013). This speaks of the importance of sensorimotor contingencies and reafference to fully account for feedback-locked P3a and P3b responses (cf., Fig. 2), over and above the surprise conveyed by exteroceptive stimuli alone. Remarkably, Stuss and Picton also interpreted their frontal P3 to negative feedback cues in terms of context updating and the resolution of the subject's uncertainty about the correctness of her *motor responses*. And regarding their parietal P4, they pointed out that "a feedforward or 'corollary' discharge [...] output may be generated to modify the expectancies for future perceptual or motor activity" (Stuss & Picton, 1978; p. 157). Such an interpretation anticipated modern views of planning as active inference (Botvinick & Toussaint, 2012; Friston et al., 2017).

SURPRISE MINIMIZATION AND THE P300 IN CARD SORTING

In response to Don's question, we labeled our feedback-locked P3a and P3b after early work on the P300 to infrequent sounds in 'attend' conditions (Squires et al., 1975). In fact, there has been much variability in the labeling of the early and late aspects of feedback-locked P3-like responses in studies with WCST analogues: P3 and P4 (Stuss & Picton, 1978), P3a and P3b (Barceló et al., 2002; Kopp & Lange, 2013), early and late P3 (Cunillera et al., 2012; Periáñez & Barceló, 2009), early and late novelty P3 (Barceló et al., 2006; Garcia-Garcia et al., 2011), P3a and Sustained potential (Kopp et al., 2006); P3a and Sustained parietal positivity (Kopp et al., 2014; Lange et al., 2016), and P3a and Posterior switch positivity (Lange, Seer, & Kopp, 2017). In some studies, the early and late P3-like aspects are mingled, and they are referred to as P3 and P300 in spite of their large intensities over frontal scalp regions (Mattes et al., 1991; Vila-Ballo et al., 2015). To add up to this chaotic taxonomy, gross differences in the timing, morphology and scalp distribution of P3-like responses across studies could raise doubts about their comparability under a strict definition of ERP component (Luck & Kappenman, 2012).

In spite of these nominal discrepancies, most authors consistently interpreted these frontal P3a and posterior P3b to feedback and switch cues in WCST analogues as indexing an updating in the contents of working memory, in reference to the 'context updating' hypothesis of P300 (Donchin & Coles, 1988). Paradoxically, though, such an account challenges long held views of context updating as linked to 'relevant' target stimuli that elicit P3b potentials with parietal maxima (Donchin, 1981). Here a crucial question is: when is the task context updated? In response to the feedback cues, during perceptual inference? Or in response to the target stimuli, during context learning? In line with other authors, we also agree with Stuss and Picton (1978) in that the context updating P300 does engage frontal cortices in response to negative feedback during perceptual categorization. Crucially, this is because predictive feedback cues are more informative and salient than target cards for ruling complex goal-directed card sorting behavior (Barceló & Cooper, 2018a; 2018b; Figs. 2 and 6).

Hence, active inference, with its computational definition of surprise minimization, offers a solid integrative framework to dilute lexical disparities and resolve long-lasting controversies regarding the role of P3a and P3b responses in context updating in terms of high- and low-level belief updating at prefrontal and posterior multimodal association cortices, respectively. Further, these new views can also accommodate recent accounts of similar P3-like positivities in terms of evidence accumulation (O'Connell, Dockree, & Kelly, 2012). Such an overarching integrative framework is much needed for a theory-guided interpretation of frontal and nonfrontal P3-like responses, when examining higher cognitive functions with different testing procedures and stimulus materials (Friston, 2005; Luria, 1966; Parr et al., 2018; Stuss et al., 2000).

The integrative potential of active inference owes to the same principle of surprise minimization being applied to both inference and learning. Under these new views, interoceptive and proprioceptive inputs generated by card sorting actions need to be integrated with exteroceptive inputs from the visual cards and feedback sounds as they all inform the inference of the same perceptual category. Frontoparietal association cortices are candidate structures for this multisensory integration (Andersen & Cui, 2009), as they generate top-down predictions in the form of corollary discharges that suppress exteroceptive and proprioceptive prediction errors to minimise surprise at both prefrontal and posterior association cortices (Teuber, 1964/2009; cf., Fig. 2). From here, feedback-locked P3a and P3b responses can be seen as belief updates at prefrontal and posterior association cortices, respectively. The former resolves the uncertainty about the new sorting category, and the later resolves the uncertainty about the corresponding stimulus-response mappings in preparation for the next card sort. In both cases, these high- and low-level belief updates are informed by beliefs about a future desired outcome: the rewarded response policy (Friston et al., 2017). This concurs with Bayesian accounts of perceptual categorization (Rigoli et al., 2017), as well as with views of planning as active inference (Botvinick & Toussaint, 2012).

The same general principle of surprise minimization applies to context learning once the subject is confident about the correct sorting category. In this situation, the target P3b indexes low-level surprise minimization at posterior association cortices because the subject cannot predict the contents of the next target card, nor the corresponding button press to close the perception-action cycle. Crucially, this explains why these target P3b responses never 'habituate' (Donchin, 1981), and instead they show repetition enhancement, indexing a gradual increase in the precision of predictions about many task parameters being learned along the card sorting series (Auksztulewicz & Friston, 2016). Further, the key difference between feedback-locked P3b and target P3b is that the former contributes to the inference of the sorting category, and is informed by beliefs about the future (*prediction*); whereas the later indexes context learning and is informed by beliefs about the past (postdiction; Friston et al., 2017). This idea concurs with dual mode models that define cognitive control in terms of two temporarily distinct stages of proactive and reactive control (Barceló & Cooper, 2018a; Braver, 2012). Note that the existence of two functionally distinct types of P3b solves a long-lasting dialectic about two competing hypotheses of P3b elicitation in terms of context updating and context closure (Donchin & Coles, 1988; Verleger, 1988). Further, inference and learning each entail accumulation of evidence over different time scales (Friston, 2005). Hence, active inference can also accommodate evidence about a centroparietal P3-like positivity (CPP) that has been interpreted in terms of evidence accumulation during perceptual decision-making. The use of task designs that foster a stable context may explain why this CPP preferentially shows a centroparietal rather than a frontal-central scalp distribution (O'Connell et al., 2012).

In fact, the scalp-recorded target P3b has long been referred to as a 'late positive complex' consisting of manifold component operations (Polich, 2007; Sutton & Ruchkin, 1984). Recent single trial EEG decomposition has shown that target P3-like responses can be divided into stimulus-locked, response-locked and latency variable P3-like sub-components (Verleger, Grauhan, & Smigasiewicz, 2016), each putatively indexing fast cycles of belief updating in the subject's generative model in response to within trial changes in exteroceptive, proprioceptive and interoceptive sensory outcomes (Brydges & Barceló, 2018). Some of these target P3-like sub-components show a more frontal distribution on first target trials following predictive cues prompting for a switch in context (Barceló & Cooper, 2018a), consistent with the larger magnitudes of prediction errors during perceptual inference in volatile contexts. In these cases, the frontal target P3a aspect peaks 50-150 msec earlier than the ensuing parietal target P3b aspect (Brydges & Barceló, 2018; cf., Donchin, 1981, Figs. 15 and 16; Verleger et al., 2016), suggesting that the stage of perceptual inference always precedes the stage of context learning (Friston, 2005). These studies

also suggest that both high- and low-level belief updating may be endogenously generated, being time-locked neither to a stimulus nor a response (cf., Johnson & Donchin, 1985). Hence, a finer grained decomposition of target P3-like responses to various sources of sensory evidence during the stages of inference and learning of perceptual categories will foster our understanding of the complex workings of frontoparietal cortical networks subserving cognitive control (Brydges & Barceló, 2018; Friston, 2005).

Traditional views on the P300 portray a sharp dichotomy between anterior P3a and posterior P3b responses that is rigidly constrained by the 'ignore' versus 'attend' task conditions instructed by the examiner (Polich, 2007). In contrast, active inference defines perceptual categorization in a much flexible, time dynamic, and context-sensitive way towards the efficient control of complex forms of goal-directed behavior (Luria, 1966; Rigoli et al., 2017). One may assume that when certain information threshold is surpassed, surprise signals will result in high-level belief updates at prefrontal cortices. This threshold seems to be a function of the likelihood of sensory outcomes given their expected hidden causes, the prior probability of those causes, and their relative precision (Friston et al., 2017; Fig. 2). This new formal scheme allows for a graded engagement of prefrontal cortices during perceptual inference, and speaks to the importance of quantifying sensorimotor information for modeling task-averaged behavioral and brain responses and their trial-by-trial dynamics (Barceló & Cooper, 2018a; 2018b; cf., Figs. 2 and 6).

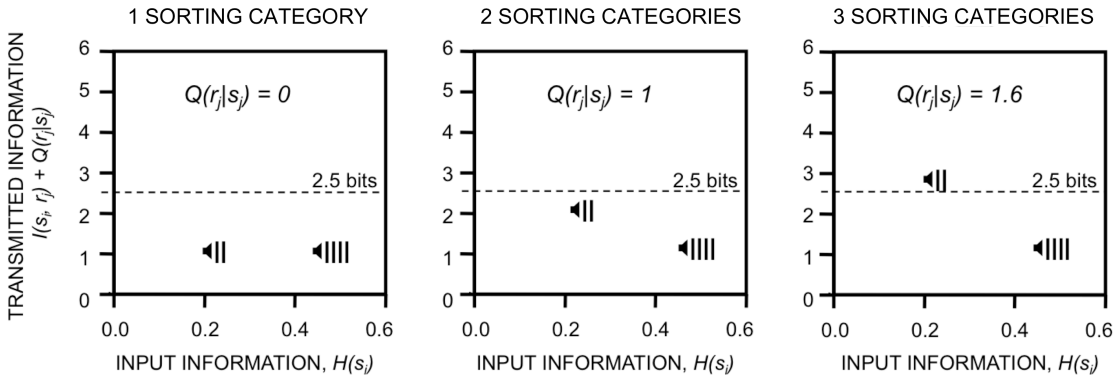

Figure 6. Estimations of information gain for predictive feedback cues as a function of the number of perceptual categories held in working memory. Estimates assume information transmission through higher [$Q(r_j|s_i)$] and lower [$I(s_i, r_j)$] levels in the neural hierarchies (cf., Fig. 2). The more perceptual categories associated with negative feedback cues (◀||), the larger the amount of information being conveyed by these cues for anticipatory response (policy) selection. Adapted from Nyhus and Barceló (2009), with permission.

To date, few studies on the WCST have attempted to quantify the sensorimotor information conveyed by task events about the hidden perceptual categories (Kopp & Lange, 2013; Nyhus & Barceló, 2009). This shall be necessary to model cortical responses in terms of information transmission across prefrontal and posterior association cortices during the inference and learning of the sorting categories (Fig. 2). For instance, using a task-switching version of the WCST, Kopp and Lange (2013) reported that frontal P3a amplitudes varied with the entropy of switch cues, whereas posterior P3b amplitudes correlated with the surprise of switch cues. However, such modeling approach was based on a simple neuronal model of the stimulus (Sokolov, 1963), and did not consider the subject's "expectancies for future perceptual or motor activity" (Stuss & Picton, 1978, p. 157).

Paradoxically, motor components and the reactions of skeletal muscles were originally seen as essential characteristics of the orienting reflex, defined as a "mechanism that facilitates the collection and transmission of information" (Sokolov, 1966; p. 351). Even though Sokolov (1966) also relied on Bayes theorem in his attempt to identify an information threshold for elicitation of the orienting reflex, he overlooked the importance of reafference between exteroceptive and proprioceptive sensory inputs for perceptual categorization (cf., Luria, 1966). In turn, when sensorimotor information is considered, then this threshold might well correspond with our capacity limits for processing information (G. A. Miller, 1956; Fig. 6). This idea was examined in a study that modeled the frontoparietal distribution of P3-like responses to predictive cues and target stimuli in a simplified task-switching WCST analogue (Barceló &

Cooper, 2018a). This study reported the largest frontal P3a responses to predictive cues in the most complex task condition conveying maximal sensorimotor information for anticipatory response (policy) selection during perceptual inference. In contrast, the largest reactive target P3b responses were obtained in the simplest task condition, consistent with maximal precision of predictions during context learning (Barceló & Cooper, 2018b).

In sum, the evidence reviewed here supports the notion of two broad classes of P3-like responses that index high- and low-level belief updating at prefrontal and posterior association cortices during the inference and learning of perceptual categories. Further modeling work will be needed to examine the existence of some information threshold in a subject's generative model for the elicitation of frontally distributed P3-like responses during perceptual categorization (Friston, 2005; Sokolov, 1966; see Fig 6).

CONCLUDING REMARKS

The Wisconsin card sorting test (WCST) has been one of the most distinctive tools for the assessment of frontal lobe functions in clinical and research contexts. Its extensive use has generated a very rich database of behavioral and brain imaging results, some of which seemed paradoxical or contradictory under traditional theories of executive attention and frontal lobe function. In this paper, we aimed to solve some of those paradoxes adopting a new theory of cortical responses inspired on the Bayesian brain hypothesis and the free energy principle (Friston, 2005, 2010). The new theory can explain changes in the frontoparietal scalp distribution of two broad classes of P300 cortical responses that had long been associated with uncertainty resolution (Sutton et al., 1965), and memory consolidation (Donchin, 1981) in many cognitive domains. The reviewed evidence suggests that when sorting cards there is instant activation across a distributed frontoparietal network (Barceló et al., 2002; Stuss & Picton, 1978), which jointly with other cortical and subcortical structures (Buchsbaum et al., 2005; Knight, 1996; Wessel et al., 2016), is differentially engaged during two temporally distinct stages: (1) during the inference of the hidden perceptual category; and (2) during learning of the task context or task parameters (Friston, 2005).

In retrospect, it is striking that Don and Terence's interpretation of their frontal-central P300 responses to the feedback cues in terms of context updating was prescient of modern views of planning as active inference. Much confusion in the literature owes to rigid attempts to define a unitary 'task relevant' P300 component based on its maximal parietal intensity (Donchin & Coles, 1988). In contrast, active inference offers a more flexible and richer picture of P3-like cortical responses as proxies of neural activity in a frontoparietal network that is differentially engaged during both inference and learning of perceptual categories. However, to date few studies have attempted to dissociate these two stages, which may have been inadvertently combined in the task-averaged ERP waveforms. This confound would foreseeably produce posterior P3b responses with parietal maxima indexing context learning, together with earlier frontal-central P3a responses with lesser intensities indexing perceptual inference in a proportion of the averaged trials (cf., Barceló et al., 2006, Fig. 2; Barceló et al., 1997, Fig. 1; Donchin, 1981, Figs. 4, 15, 16; Duncan-Johnson & Donchin, 1977, Figs. 1b,c ; Johnson & Donchin, 1985, Figs., 1, 2 ,3).

Future research could aim to undo potential confounds in the recording of P3-like responses during the stages of perceptual inference and learning in cognitive tasks. For this, it will be important to model Bayesian surprise, and to identify an information threshold above which there will be a gradually larger frontal involvement in the elicitation of anterior P300 responses during perceptual inference (Barceló & Cooper, 2018a; Sokolov, 1966). In this respect the Bayesian brain hypothesis reinstates pioneering information processing views in cognitive neuroscience (G. A. Miller, 1956, Fig. 6). A quantitative estimation of such a threshold would allow researchers to examine how this quantity depends on individual variables such as age, memory span, fluid intelligence, or brain damage. Further, an accurate report of P3-like responses requires their dissociation from other overlapping cortical responses, like negative expectancy waves (Gajewski et al., 2018; Luck & Kappenman, 2012; Stuss & Picton, 1978).

From these new premises, frontal and nonfrontal P3-like responses offer promising biomarkers of epistemic (exploratory) and pragmatic (exploitative) behavior during the inference and learning of perceptual categories in card sorting and, more generally, as basic tools in the neuropsychological assessment of higher cortical functions in humans (Luria, 1966; Parr et al., 2018; Stuss & Benson, 1984).

Competing interests

The author declares no competing interests.

REFERENCES

Adrover-Roig, D., & Barceló, F. (2010). Individual differences in aging and cognitive control modulate the neural indexes of context updating and maintenance during task switching. *Cortex, 46*, 434-450.

Andersen, R. A., & Cui, H. (2009). Intention, action planning, and decision making in parietal-frontal circuits. *Neuron, 63*(5), 568-583.

Auksztulewicz, R., & Friston, K. (2016). Repetition suppression and its contextual determinants in predictive coding. *Cortex, 80*, 125-140.

Badre, D., & Nee, D. E. (2018). Frontal Cortex and the Hierarchical Control of Behavior. *Trends in Cognitive Sciences, 22*(2), 170-188.

Barceló, F. (1999). Electrophysiological evidence of two different types of error in the Wisconsin Card Sorting Test. *Neuroreport, 10*(6), 1299-1303.

Barceló, F. (2003). The Madrid card sorting test (MCST): a task switching paradigm to study executive attention with event-related potentials. *Brain Research Protocols, 11*(1), 27-37.

Barceló, F., & Cooper, P. S. (2018a). An information theory account of late frontoparietal ERP positivities in cognitive control. *Psychophysiology, 55*, e12814.

Barceló, F., & Cooper, P. S. (2018b). Quantifying contextual information for cognitive control. *Frontiers in Psychology, 9*, 1693.

Barceló, F., Escera, C., Corral, M. J., & Periánez, J. A. (2006). Task switching and novelty processing activate a common neural network for cognitive control. *Journal of Cognitive Neuroscience, 18*(10), 1734-1748.

Barceló, F., & Knight, R. T. (2002). Both random and perseverative errors underlie WCST deficits in prefrontal patients. *Neuropsychologia, 40*(3), 349-356.

Barceló, F., & Knight, R. T. (2007a). An information-theoretical approach to contextual processing in the human brain: evidence from prefrontal lesions. *Cerebral Cortex, 17 Suppl 1*, i51-60.

Barceló, F., & Knight, R. T. (2007b). Theoretical sequelae of a chronic neglect and unawareness of prefrontotectal pathways in the human brain. *Behavioral and Brain Sciences, 30*(1), 83-85.

Barceló, F., Muñoz-Céspedes, J. M., Pozo, M. A., & Rubia, F. J. (2000). Attentional set shifting modulates the target P3b response in the Wisconsin card sorting test. *Neuropsychologia, 38*(10), 1342-1355.

Barceló, F., Perianez, J. A., & Knight, R. T. (2002). Think differently: a brain orienting response to task novelty. *NeuroReport, 13*(15), 1887-1892.

Barceló, F., & Rubia, F. J. (1998). Non-frontal P3b-like activity evoked by the Wisconsin Card Sorting Test. *NeuroReport, 9*(4), 747-751.

Barceló, F., Sanz, M., Molina, V., & Rubia, F. J. (1997). The Wisconsin Card Sorting Test and the assessment of frontal function: A validation study with event-related potentials. *Neuropsychologia, 35*(4), 399-408.

Bayless, S. J., Gaetz, W. C., Cheyne, D. O., & Taylor, M. J. (2006). Spatiotemporal analysis of feedback processing during a card sorting task using spatially filtered MEG. *Neuroscience Letters, 410*(1), 31-36.

Bledowski, C., Prvulovic, D., Hoechstetter, K., Scherg, M., Wibral, M., Goebel, R., et al. (2004). Localizing P300 generators in visual target and distractor processing: a combined event-related potential and functional magnetic resonance imaging study. *Journal of Neuroscience, 24*(42), 9353-9360.

Botvinick, M., & Toussaint, M. (2012). Planning as inference. *Trends in Cognitive Sciences, 16*(10), 485-488.

Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in Cognitive Sciences, 16*(2), 106-113.

Brydges, C. R., & Barceló, F. (2018). Functional dissociation of latency-variable, stimulus- and response-locked target P3 sub-components in task-switching. *Frontiers in Human Neuroscience, 12*, 60.
Buchsbaum, B. R., Greer, S., Chang, W. L., & Berman, K. F. (2005). Meta-analysis of neuroimaging studies of the Wisconsin card-sorting task and component processes. *Human Brain Mapping, 25*(1), 35-45.
Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181-204.
Cunillera, T., Fuentemilla, L., Perianez, J., Marco-Pallares, J., Kramer, U. M., Camara, E., et al. (2012). Brain oscillatory activity associated with task switching and feedback processing. *Cognitive, Affective, and Behavioral Neuroscience, 12*(1), 16-33.
Dehais, F., Hodgetts, H. M., Causse, M., Behrend, J., Durantin, G., & Tremblay, S. (2019). Momentary lapse of control: A cognitive continuum approach to understanding and mitigating perseveration in human error. *Neuroscience and Biobehavioral Reviews, 100*, 252-262.
Demakis, G. J. (2003). A meta-analytic review of the sensitivity of the Wisconsin Card Sorting Test to frontal and lateralized frontal brain damage. *Neuropsychology, 17*(2), 255-264.
Diamond, A. (2013). Executive functions. *Annual Review of Psychology, 64*, 135-168.
Diaz-Blancat, G., Garcia-Prieto, J., Maestú, F., & Barceló, F. (2018). Fast Neural Dynamics of Proactive Cognitive Control in a Task-Switching Analogue of the Wisconsin Card Sorting Test. *Brain Topography, 31*(3), 407-418.
Donchin, E. (1981). Surprise! Surprise? *Psychophysiology, 18*, 493-513.
Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences, 11*, 343-356.
Doya, K., & Ishii, K. (2007). A probability primer. In K. Doya, S. Ishii, A. Pouget & R. P. N. Rao (Eds.), *Bayesian brain: Probabilistic approaches to neural coding* (pp. 3-13). Cambridge, Massachussetts: The MIT Press.
Duncan-Johnson, C. C., & Donchin, E. (1977). On quantifying surprise: the variation of event-related potentials with subjective probability. *Psychophysiology, 14*(5), 456-467.
Ebitz, R. B., Sleezer, B. J., Jedema, H. P., Bradberry, C. W., & Hayden, B. Y. (2019). Tonic exploration governs both flexibility and lapses. *PLoS Computational Biology, 15*(11), e1007475.
Friedman, D., Cycowicz, Y. M., & Gaeta, H. (2001). The novelty P3: an event-related brain potential (ERP) sign of the brain's evaluation of novelty. *Neuroscience and Biobehavioral Reviews, 25*(4), 355-373.
Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. B. Biological Sciences, 360*(1456), 815-836.
Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews in Neuroscience, 11*(2), 127-138.
Friston, K. (2019). Waves of prediction. *Plos Biology, 17*(10), e3000426.
Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Frontiers in Psychology, 3*, 151.
Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience and Biobehavioral Reviews, 68*, 862-879.
Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active Inference: A Process Theory. *Neural Computation, 29*(1), 1-49.
Friston, K., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., et al. (2012). Dopamine, affordance and active inference. *PLoS Computational Biology, 8*(1), e1002327.
Fuster, J. M. (2013). Cognitive functions of the prefrontal cortex. In D. T. Stuss & R. T. Knight (Eds.), *Principles of frontal lobe function* (2ª ed., pp. 11-22). New York: Oxford University Press.
Fuster, J. M. (2019). The prefrontal cortex in the neurology clinic. In M. D'Esposito & J. H. Grafman (Eds.), *Handbook of Clinical Neurology. The Frontal Lobes* (3rd ed., Vol. 163, pp. 3-15). Amsterdam: Elsevier.
Gajewski, P. D., Ferdinand, N. K., Kray, J., & Falkenstein, M. (2018). Understanding Sources of Adult Age Differences in Task Switching: Evidence from Behavioral and ERP Studies. *Neuroscience and Biobehavioral Reviews, 92*, 255-275.
Garcia-Garcia, M., Barceló, F., Clemente, I. C., & Escera, C. (2011). COMT and ANKK1 gene–gene interaction modulates contextual updating of mental representations. *Neuroimage, 56*(3), 1641-1647.

Glascher, J., Adolphs, R., & Tranel, D. (2019). Model-based lesion mapping of cognitive control using the Wisconsin Card Sorting Test. *Nature Communications, 10*(1), 20.
Hohwy, J. (2016). The Self-Evidencing Brain. *Noûs, 50*(2), 259-285.
Hohwy, J. (2019). Prediction error minimization in the brain. In M. Sprevak & M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 159-172). Abingdon Oxon UK: Routledge.
Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research, 49*(10), 1295-1306.
Johnson, R., Jr., & Donchin, E. (1985). Second thoughts: multiple P300s elicited by a single stimulus. *Psychophysiology, 22*(2), 182-194.
Knight, R. T. (1996). Contribution of human hippocampal region to novelty detection. *Nature, 383*, 256-259.
Knight, R. T. (1997). A distributed cortical network for visual attention. *Journal of Cognitive Neuroscience, 9*, 75-91.
Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences, 11*(6), 229-235.
Konishi, S., Nakajima, K., Uchida, I., Kameyama, M., Nakahara, K., Sekihara, K., et al. (1998). Transient activation of inferior prefrontal cortex during cognitive set shifting. *Nature Neuroscience, 1*, 80-84.
Kopp, B., & Lange, F. (2013). Electrophysiological indicators of surprise and entropy in dynamic task-switching environments. *Frontiers in Human Neuroscience, 7*, 300.
Kopp, B., Lange, F., Howe, J., & Wessel, K. (2014). Age-related changes in neural recruitment for cognitive control. *Brain and Cognition, 85*, 209-219.
Kopp, B., Tabeling, S., Moschner, C., & Wessel, K. (2006). Fractionating the neural mechanisms of cognitive control. *Journal of Cognitive Neuroscience, 18*(6), 949-965.
Lange, F., Seer, C., & Kopp, B. (2017). Cognitive flexibility in neurological disorders: Cognitive components and event-related potentials. *Neuroscience and Biobehavioral Reviews, 83*, 496-507.
Lange, F., Seer, C., Loens, S., Wegner, F., Schrader, C., Dressler, D., et al. (2016). Neural mechanisms underlying cognitive inflexibility in Parkinson's disease. *Neuropsychologia, 93*(Pt A), 142-150.
Lange, F., Seer, C., Muller-Vahl, K., & Kopp, B. (2017). Cognitive flexibility and its electrophysiological correlates in Gilles de la Tourette syndrome. *Developmental Cognitive Neuroscience, 27*, 78-90.
Li, F., Wang, J., Du, B., & Cao, B. (2018). Electrophysiological Response to the Informative Value of Feedback Revealed in a Segmented Wisconsin Card Sorting Test. *Frontiers in Psychology, 9*, 57.
Luck, S. J., & Kappenman, E. S. (2012). *Handbook of event-related potential components*. Oxford: Oxford University Press.
Luria, A. R. (1966). *Higher cortical functions in man*. London: Tavistock Publications.
Maheu, M., Dehaene, S., & Meyniel, F. (2019). Brain signatures of a multiscale process of sequence learning in humans. *Elife, 8*.
Maisto, D., Friston, K., & Pezzulo, G. (2019). Caching mechanisms for habit formation in Active Inference. *Neurocomputing, 359*, 298-314.
Mattes, R., Cohen, R., Berg, P., Canavan, A. G. M., & Hopmann, G. (1991). Slow cortical potentials (SCPS) in schizophrenic patients during performance of the Wisconsin Card-sorting Test (WCST). *Neuropsychologia, 29*, 195-205.
Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*, 167-202.
Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97.
Milner, B. (1963). Effects of different brain lesions on card sorting. *Archives of Neurology, 9*, 100-110.
Monchi, O., Petrides, M., Petre, V., Worsley, K., & Dagher, A. (2001). Wisconsin Card Sorting revisited: distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *Journal of Neuroscience, 21*(19), 7733-7741.
Nelson, H. E. (1976). A modified card sorting test sensitive to frontal lobe defects. *Cortex, 12*, 313-324.
Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson , G. E. Schwartz & D. Shapiro (Eds.), *Consciousness and self-regulation* (Vol. 4). New York: Plenum.
Nyhus, E., & Barceló, F. (2009). The Wisconsin Card Sorting Test and the cognitive assessment of prefrontal executive functions: A critical update. *Brain and Cognition, 71*(3), 437-451.

O'Connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience, 15*(12), 1729-1735.
O'Regan, J. K., & Noe, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*(5), 939-973; discussion 973-1031.
Parr, T., & Friston, K. J. (2017). Working memory, attention, and salience in active inference. *Sci Rep, 7*(1), 14678.
Parr, T., Rees, G., & Friston, K. J. (2018). Computational Neuropsychology and Bayesian Inference. *Frontiers in Human Neuroscience, 12*, 61.
Parr, T., Rikhye, R. V., Halassa, M. M., & Friston, K. J. (2019). Prefrontal Computation as Active Inference. *Cerebral Cortex, 30*(2), 682-695.
Periáñez, J. A., & Barceló, F. (2009). Updating sensory versus task representations during task-switching: insights from cognitive brain potentials in humans. *Neuropsychologia, 47*(4), 1160-1172.
Periáñez, J. A., Maestu, F., Barcelo, F., Fernandez, A., Amo, C., & Ortiz Alonso, T. (2004). Spatiotemporal brain dynamics during preparatory set shifting: MEG evidence. *Neuroimage, 21*(2), 687-695.
Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology, 118*(10), 2128-2148.
Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13*, 25-42.
Rigoli, F., Pezzulo, G., Dolan, R., & Friston, K. (2017). A Goal-Directed Bayesian Framework for Categorization. *Frontiers in Psychology, 8*, 408.
Shallice, T., Stuss, D. T., Picton, T. W., Alexander, M. P., & Gillingham, S. (2008). Mapping task switching in frontal cortex through neuropsychological group studies. *Frontiers in Neuroscience, 2*(1), 79-85.
Sokolov, E. N. (1963). *Perception and the conditioned reflex*. Oxford: Pergamon Press.
Sokolov, E. N. (1966). Orienting reflex as information regulator. In A. N. Leontiev, A. R. Luria & A. A. Smirnov (Eds.), *Psychological Research in the U.S.S.R.* (pp. 334–360). Moscow: Progress Publishers.
Squires, N. K., Squires, K. C., & Hillyard, S. A. (1975). Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and Clinical Neurophysiology, 38*, 387-401.
Stuss, D. T., & Alexander, M. P. (2000). Executive functions and the frontal lobes: a conceptual view. *Psychological Research, 63*(3-4), 289-298.
Stuss, D. T., & Alexander, M. P. (2007). Is there a dysexecutive syndrome? *Philosophical Transactions of the Royal Society of London. B. Biological Sciences, 362*(1481), 901-915.
Stuss, D. T., & Benson, D. F. (1984). Neuropsychological studies of the frontal lobes. *Psychological Bulletin, 95*, 3-28.
Stuss, D. T., & Knight, R. T. (Eds.). (2013). *Principles of frontal lobe function* (2nd ed.). New York: Oxford University Press.
Stuss, D. T., Levine, B., Alexander, M. P., Hong, J., Palumbo, C., Hamer, L., et al. (2000). Wisconsin card sorting test performance in patients with focal frontal posterior brain damage: effects of lesion location and test structure on separable cognitive. *Neuropsychologia, 38*, 388-402.
Stuss, D. T., & Picton, T. W. (1978). Neurophysiological correlates of human concept formation. *Behavioral Biology, 23*(2), 135-162.
Stuss, D. T., Picton, T. W., & Alexander, D. C. (2001). Consciousness, self-awareness, and the frontal lobes. In S. P. Salloway, P. F. Malloy & J. D. Duffy (Eds.), *The Frontal Lobes and Neuropsychiatric Illness* (pp. 101-109). Washington: American Psychiatric Publishing, Inc.
Stuss, D. T., Shallice, T., Alexander, M. P., & Picton, T. W. (1995). A multidisciplinary approach to anterior attentional functions. *Ann N Y Acad Sci, 769*, 191-211.
Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science, 150*, 1187-1188.
Sutton, S., & Ruchkin, D. S. (1984). The late positive complex. Advances and new problems. *Annals of the New York Academy of Science, 425*, 1-23.
Teuber, H. L. (1964/2009). The riddle of frontal lobe function in man. *Neuropsychology Review, 19*(1), 25-46.

Verleger, R. (1988). Event-related potentials and cognition: A critique of the context updating hypothesis and an alternative interpretation of P3. *Behavioral and Brain Sciences, 11*, 343-356.

Verleger, R., Grauhan, N., & Smigasiewicz, K. (2016). Is P3 a strategic or a tactical component? Relationships of P3 sub-components to response times in oddball tasks with go, no-go and choice responses. *Neuroimage, 143*, 223-234.

Vila-Ballo, A., Cunillera, T., Rostan, C., Hdez-Lafuente, P., Fuentemilla, L., & Rodriguez-Fornells, A. (2015). Neurophysiological correlates of cognitive flexibility and feedback processing in violent juvenile offenders. *Brain Research, 1610*, 98-109.

Wang, J., Cao, B., Cai, X., Gao, H., & Li, F. (2015). Brain Activation of Negative Feedback in Rule Acquisition Revealed in a Segmented Wisconsin Card Sorting Test. *PLoS One, 10*(10), e0140731.

Wessel, J. R., & Aron, A. R. (2017). On the Globality of Motor Suppression: Unexpected Events and Their Influence on Behavior and Cognition. *Neuron, 93*(2), 259-280.

Wessel, J. R., Jenkinson, N., Brittain, J. S., Voets, S. H., Aziz, T. Z., & Aron, A. R. (2016). Surprise disrupts cognition via a fronto-basal ganglia suppressive mechanism. *Nature Communications, 7*, 11195.

Winn, J., & Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research, 6*, 661-694.

.