Representations of action monitoring and cognitive control by single neurons in the human brain

Thesis by

Zhongzheng Fu

In Partial Fulfillment of the Requirements for
the degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2019
(Defended December 10, 2018)

Zhongzheng Fu

ORCID: 0000-0002-2572-6284

Acknowledgements

I would like to thank my mentors Professors Ralph Adolphs and Ueli Rutishauser. I am truly grateful for their mentorship and generous support. They are the best advisors that I could have hoped for. Coming from a physics background, I entered the lab in 2013 with little experience in cognitive neuroscience and electrophysiology. Ralph and Ueli provide me with an excellent and welcoming environment and freedom to think and ask questions. No matter how stupid the questions turned out to be later, when I look back, they were always there to guide me through the twists and turns of scientific adventure with the utmost patience. My advisors, in turn, asked me questions. Their questions always propelled me to think more carefully about my hypotheses, models, and conclusions, and helped me identify blind spots and "acupuncture points" in my thinking and knowledge base, and helped me gradually become an independent thinker. I learned from them the importance of being rigorous, how to synthesize from a plethora of detailed rigorous findings a coherent big picture, and how to persevere through the hardship of scientific publishing and grant applications. I cherished every moment that I spent with my advisors in discovering, most importantly, the joy of doing science. Apart from nurturing my academic development, Ralph also spent an enormous amount of energy and time to organize the various events to pull me, and everybody else in the lab away from our desks and to embrace the beautiful So-Cal life. When I was going through the dark period of my very first paper rejection, Ralph offered me ticket to the LA Phil performance of Stabat Mater by Pergolesi. The concert served as a kind of catharsis and it gave me the inspiration and strength I needed to persist through the difficult times. We also had the opportunity to go on a rather strenuous hike in Catalina Island during which we had inspiring conversations about life philosophy that I would never forget.

I would like to thank Dr. Adam Mamelak for his mentorship, his insightful discussions and editing of numerous manuscripts. Adam is not only a dedicated and excellent neurosurgeon, but also a talented scientist that contributed a lot to my academic development. I would like to thank Dr. Jeffrey Chung and Chrystal Jeffrey for their dedicated assistance and facilitation for patient care and experiments.

I would like to thank my committee members: Richard Murray, John Allman and Joel Burdick for their support and advice.

I would like to thank the patients I was fortunate enough to work with over the years. They generously donated their time and energy to work with me on my experiments, even though they were in such a difficult situation. Without them, this thesis would not exist.

I would like to thank my friend and talented colleague, Shuo Wang. We collaborated in chamber music and he introduced Ralph and Ueli's work to me during our rehearsal breaks. It was from him that I first knew about the exciting and almost Caltech-exclusive opportunities of cognitive neuroscience and human single unit research at Caltech, and the amazing labs of Ralph and Ueli.

I would like to thank all of my talented colleagues in the Adolphs lab and Rutishauser lab for their helpful comments and collaboration over the years. Especially I would like to thank Remya Nair for her technical support. Under her efficient management, the high-power computing clusters in our lab became an indispensable resource for my research. Without her help, this thesis would have taken considerably longer time to complete due to all the heavy-lifting computation involved.

I would like to thank all my friends at Caltech and outside of Caltech for their companionship and advice, especially Drs. Sijia Dong, Yuqing Zhu, Jiang Li, Yingrui Chang, Pinglei Bao, Sai Sun, Lawrence Jiaqi Jin, Zhe An, Bo Xiong, Dean A. Nash, Yifan Li. I would like to thank all funding agencies that make our work possible.

Finally, I would like to thank my parents and dedicate this thesis to them. Without their care and support, I would not be where and who I am today.

Abstract

Cognitive control arises whenever a prepotent and often automatic response needs to be overcome by another response. Control is usually effortful and relies on monitoring processes that detect when control is needed and/or when it failed. Control is one of the most important aspects of human behavior in everyday life and is a critical component of executive function. In a series of three empirical chapters, I present results from invasive single-neuron recordings from the frontal cortex of neurosurgical human patients while they perform tasks requiring cognitive control. I show that a substantial proportion of neurons in the pre-supplementary motor area (pre-SMA), and in the dorsal anterior cingulate cortex (dACC), signal response errors shortly after they occurred, but well before onset of feedback. Here I demonstrate that these error neurons signal self-detected errors and that they were separate from neurons signaling conflict. The response of error neurons correlated trial-by-trial with the simultaneously recorded intracranial error-related negativity (iERN), thereby establishing a single-neuron correlate of this important scalp potential. iERN-error neuron synchrony in dACC, but not pre-SMA, predicted whether post-error slowing, which is a measure of control, occurred or not. Spike-field coherence between action potentials and local field potentials in specific frequency bands, and latency differences between the different brain regions, suggest a mechanistic model whereby information relevant to control is passed between sectors of the medial frontal cortex. Multiplexing of different ex-post monitoring signals by individual neurons further documents that control relies on multiple sources of information, which can be dynamically routed in the brain depending on task demands. These findings provide the most complete set of single-neuron data on how errors and conflict signals at the single neuron level contribute to cognitive controls in humans. They provide a first-single neuron correlate of an extensively utilized scalp EEG potential. Together, this work provides a strong complement to investigations of this topic using fMRI in humans, and using electrophysiology in monkeys, and suggests specific future directions.

Published Content and Contributions

Fu, Z., Wu, D. J., Ross, I., Chung, J. M., Mamelak, A. N., Adolphs, R. & Rutishauser, U. 2019. Single-Neuron Correlates of Error Monitoring and Post-Error Adjustments in Human Medial Frontal Cortex. Neuron, 101, 165-177. doi: 10.1016/j.neuron.2018.11.016

Z. Fu participated in the conception of the project, collected and analyzed data and participated in writing of the manuscript.

Table of Contents

Chapter 1. Introduction

1.1 General introduction

In this section, I will present a general background and some concepts that I will use to interpret the findings in the later part of the thesis. These concepts are intuitive, but they are at best very crude conceptual models of the plethora of neural mechanisms at work; they nevertheless provide a general readership with a starting point.

1.1.1 The malleable spectrum of automaticity

Learning is one of the most fundamental aspects of human life. Through learning, we acquire a diverse set of skills. Some skills have become so well-practiced that we can perform them with little attention and without worrying about intrusions from distraction. Some skills, however, remain difficult and require greater amount of mental effort to perform. One important example for the former is reading. Reading is a higher cognitive function that is incredibly complex (Friederici, 2017). Our brain needs to take in visual patterns of lines and dots that consist of letters, bind the low-level perceptual representations together in meaningful ways, and map these to memorized semantic representations; all of these processes happen within hundreds of milliseconds. In some cases, even if the exact spelling of a word is wrong, we can read it without noticing it is misspelled as long as it appears in the context of a sentence. By contrast, mental arithmetic tasks such as multiplying two numbers that are greater than 50 generally remain difficult and costly both in time and in mental effort even when the skill has been practiced since in grade school.

The distinction between these two intuitively different kinds of skills – like reading vs. doing complex arithmetic – has a long history in cognitive psychology that has attempted to map it onto a dichotomy of cognitive processing. Some of the earliest formalisms in cognitive psychology labeled the distinction "automatic" versus "controlled" (Schneider and Shiffrin, 1977, Shiffrin and Schneider, 1977) with a host of additional attributes that have been added over the years (Table 1). Not all of these come together as a package all of the time, and it is doubtful that there really are two clean "systems" of this kind in the brain in any strong sense. Nonetheless, they serve to characterize different types of processing and to introduce the topic of my thesis: given multiple (at least these two) types of processing, how does the brain prioritize one over the other? This is a large question whose answers range from winner-take-all mechanisms to arbitration between goal-directed and habit-based decision-making. The specific focus of my thesis is on the situation where an automatic process interferes with a deliberate process: the topic of cognitive control.

Tasks that are more deliberate are susceptible to interference from automatic responses. When we need to perform these deliberate tasks in order to achieve certain goals, we rely on cognitive control to protect against the interference from more automatic behaviors. In addition, when we are faced with a novel environment with a new set of goals to achieve, in which the learned skills may no longer apply in the new settings, our brain needs to engage cognitive control to adapt to the new cognitive demands and guard the learning process against intrusions of prior learned responses. Cognitive control is thus essential for functioning in nonstationary environments, and a key component of flexible, adaptive behavior. It is interesting to note that, across a number of tasks, cognitive control seems to correlate with brain volume. Animals with larger brains (specifically, logarithmic endocranial volume) perform better on tasks requiring cognitive control (MacLean et al., 2014). In vertebrates, cognitive control has long been linked to functions of the prefrontal cortex (Fuster, 2015, Passingham and Wise, 2012), which is the brain region on which my thesis will focus.

System 1	System 2
Automatic	Controlled
Heuristic	Systematic
Fast	Slow
Effortless	Effortful
Non-conscious	Conscious
Emotional	Rational
Implicit	Explicit
Reflexive	Reflective
Intuitive	Analytic
Parallel	Serial

Table 1.1 Representative attributes of System I and System II categories.

1.1.2 Action monitoring

In the pursuit of a goal, the intrusion of automatic responses can often lead to reduced quality of the performed action, and in some cases a total failure in performing the goal-required actions at all. It is thus indispensable for the brain to develop mechanisms to monitor the quality and outcomes of actions during goal-directed behaviors. Monitoring of outcomes can be achieved in two ways depending on the source of information utilized: whether it is internal or external to the subject. When actions required by the goals have not been previously learned, the agent relies heavily on external feedback information provided by the environment. This external feedback specifies the outcome of the actions. However, when these goal-directed actions have become very well learned after adequate exposure to the environment, the agent generates predictions of the outcomes and can thus rely on his or her internal models of the environment to monitor the actions and action outcomes. The internal model that the agent develops through interaction with the environment can generate predictions of actions and their outcomes, and a prediction error can thus be computed when the agent commits the action.

As mentioned above, another important aspect of performing goal-directed actions is the interference from automatic responses. The logic for monitoring action difficulty is simple: interference experienced when performing an action can provide crucial information for optimizing ongoing or subsequent actions. If an executed action is compromised by a concurrently activated automatic response such that it is performed with reduced quality and/or efficiency, it is important to engage cognitive control to optimize the next attempt at similar actions. This type of action monitoring differs fundamentally from the case of error monitoring: while errors can be determined externally (via explicit sensory feedback or reward manipulations), difficulty can only be determined internally and is thus an entirely subjective measure. This self-monitoring of actions is one typical type of metacognition. Metacognition, first used to label higher-level cognition by American psychologist John Flavell, is the ability to monitor and control internal cognitive processes. Self-monitoring of action outcome and action difficulty, which require access to ongoing action production and decision-making processes that occur covertly, qualifies as metacognition (Yeung and Summerfield, 2012, Metcalfe, 2008, Smith et al., 2003). In this thesis, I investigate the neural basis of both outcome (Chapter 2) and difficulty (Chapter 4) monitoring.

1.1.3 From monitoring to control

One of the most important consequences of monitoring an action and its outcomes is the subsequent optimization of similar actions, or reward probability in the future (Ullsperger et al., 2014). Cognitive control here refers to a diverse set of neural mechanisms that are sufficient and necessary for behavioral optimization (Botvinick et al., 2001, Ullsperger et al., 2014, Ridderinkhof et al., 2004). One fascinating aspect about cognitive control is that it can be applied to a huge variety of cognitive processes and is general-purpose in that sense. Yet it is also to some extent domain-specific: cognitive control can be improved in one task, but this improvement does not necessarily transfer to another task.

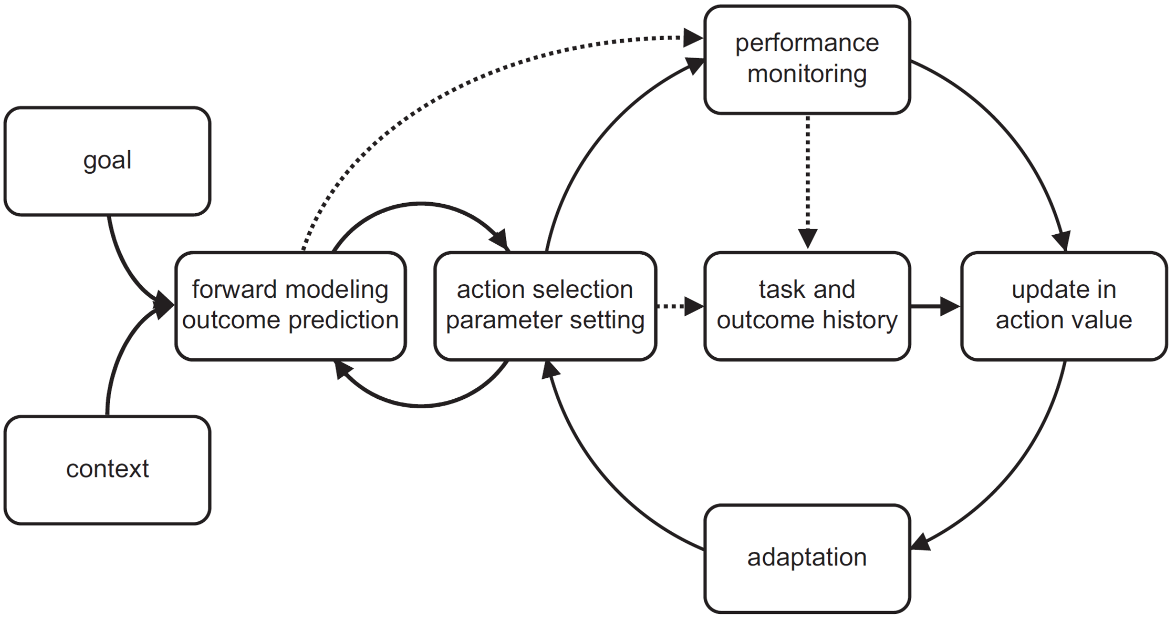

Figure 1.1 Schematic showing how different processes of the behavioral control feedback loop are related to each other. Adapted from (Ullsperger et al., 2014).

In the human psychology literature, the two most studied behavioral signatures that are thought to reflect cognitive control are post-error slowing (PES) and conflict adaptation effect. In post-error slowing, the subject delays the action on the next trial after making an error. The conflict adaptation effect, first described by (Gratton et al., 1992), refers to the fact that subjects tend to speed up on a trial where interference is experienced if it followed another trial with interference. Both behavioral signatures are an interaction effect between reaction times on current and previous trials. The conflict adaptation effect is taken as a signature of engagement of cognitive control: the conflict level in the stimuli is kept constant since both trials in question use stimuli that should cause interference, but cognitive control is engaged by the first interfering trial and continues on to resolve the interference on the next trial faster. The most straightforward explanation for the PES effect is that the occurrence of errors recruits cognitive control to slow down the motor system in service of responding with more caution and therefore higher accuracy in the next attempt. However, such singular accounts of PES as an adaptive behavior have been disputed. Studies has found that there are maladaptive components to PES: it includes an orienting response (Notebaert et al., 2009), as well as a reduction in sensitivity to sensory information (Purcell and Kiani, 2016).

The above brief introduction thus suggests specific components whereby intended actions are represented, the consequences of actions are monitored to check that they correspond to intentions, and control mechanisms can be engaged to optimize or correct actions as needed. This general scheme is based on a long history of first-principles reasoning, on engineering considerations in artificial systems, and on observations in biological systems. I turn next to the latter, and to the different types of tasks that have been used here.

1.2 Behavioral paradigms commonly used to study action monitoring and control

In this section, I will introduce the classical behavioral paradigms used to study performance monitoring and cognitive control in humans and animals. This by no means is an exhaustive review of all the existing tasks; the purpose here is to provide an inventory of the most common ones, including the ones I have used in my dissertation work.

1.2.1 Stroop task

The Stroop task is perhaps the most widely used task of all in humans. It is a cognitive behavioral task widely used to study attention and interference. This task is named after John Ridley Stroop, who systematically designed the task in its current form and analyzed the effect that was subsequently named after him (see below). Stroop was interested in understanding the interference between conflicting processes (MacLeod, 1991, Stroop, 1935). Since then, the Stroop effect has become one of the most studied behavioral effects in cognitive psychology.

Many variants of the Stroop task have been developed. In a typical modern version of the task, subjects are shown words whose meaning corresponds to colors, one by one. The font color that the word is printed in can be congruent or incongruent with its meaning.

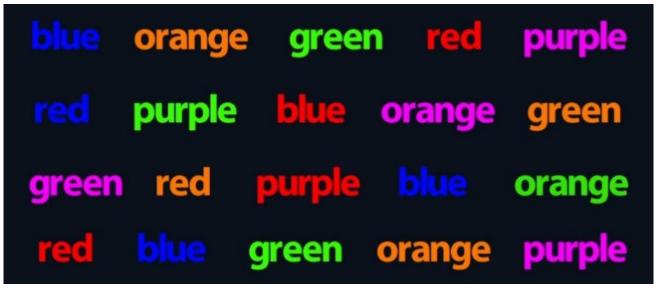
blue orange green red purple
red purple blue orange green
green red purple blue orange
red blue green orange purple

Figure 1.1 Typical stimuli for the Stroop task
The task instruction can be “name the color” or “name the word”. When the instruction is to “name the color”, reaction times are longer on word-color incongruent trials than on word-color congruent trials. Sources: study.com.

The subject is explicitly instructed to name the font color. The Stroop interference effect refers to the finding that subjects respond significantly slower on a word-color incongruent trial (e.g., ‘green’ written in red ink) than on a word-color congruent trial (e.g., ‘green’ written in green ink). In fact, the Stroop effect is one of the most reliable behavioral effects derived from psychometric tests(MacLeod, 1991). Over the years, it has found widespread application also in clinical settings. Given its robustness, it is perplexing that the detailed neural correlates of the Stroop effect remain largely unknown to this date.

1.2.2 Wisconsin card sorting task (WCST)

This task is also a very popular task among psychologists and clinicians and used to assess the functional integrity of the prefrontal cortex (Anderson et al., 1991, Gold et al., 1997, Glascher et al., 2012). It requires cognitive control, among many other cognitive processes such as working memory. The subject is instructed to match the card he or she has with one of the cards presented as possible choices. The symbols on the card have several properties, such as the shape, color, and quantity of a certain symbol, that each can be used to match. The rule for matching is not explicitly told to the subject; the subject is required to infer the rule by the feedback (often in the form of reinforcement). Successful performance of the task requires suppression of the previously rewarded rule in favor of an alternative rule, based on outcome history. This is different from the Stroop task where the response rule (which is to name the color) is constant.

1.2.2 Simon task

In the Simon task (Simon and Wolf, 1963), subjects respond to a visual stimulus (typically, a colored square or letter) by making a keypress or squeezing response on the left or right (stimulus identity-response mapping is pre-defined by the experimenter). The visual stimulus can appear either to the left or to the right of the central fixation mark. Although the stimulus location is irrelevant to the task, its spatial location (left or right) interferes with the task (mapping left or right button press to stimulus identity). The Simon effect describes the effect that subjects respond slower and make more errors when the stimulus appears on the location incongruent with the response. Like the Stroop task, the Simon task is a simple reaction-time interference task in which an automatic and prepotent visual-motor response interferes with a controlled response determined by task instructions. What differs from the Stroop task is that the prepotent response here is to respond to the sensory stimuli by the side they appear instead of reading a word.

1.2.3 Flanker task

Developed by Eriksen et al (Eriksen and Eriksen, 1974), the task can be performed with minimal instruction and does not involve word reading, similar to the Simon task. Here, the task is to report the direction of a target arrow when it is flanked by several arrows pointing to the direction that is either congruent of incongruent with the central target. Here, the flanker effect is evident when the central and flanker arrow directions do not match: subjects take more time to name the direction and make more errors. Here, the stimulus-response mapping is matched in terms of level of training between trials with or with interference: there is no inherent bias to press the left or right key. The flanker effect demonstrates that interference need not only exist between a prepotent and a controlled response, but also can manifest itself as a competition for attention.

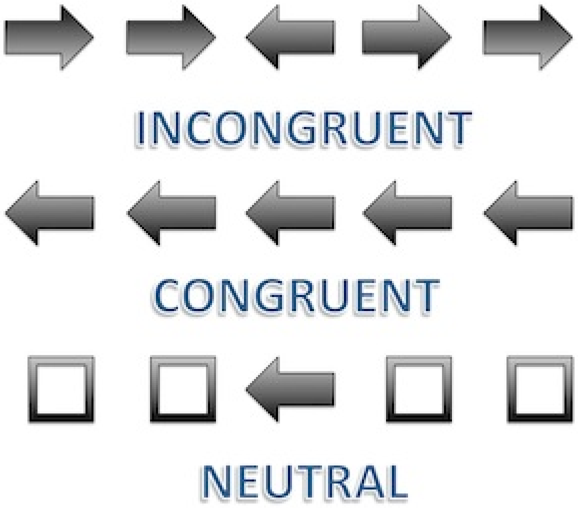

Figure 1.2 Typical stimuli for the Eriksen flanker task

1.2.4 Multi-source interference task (MSIT)

The multi-source interference task (Bush and Shin, 2006), as its name suggests, combines two sources of interference in one stimulus: identity and position interference. The stimulus used in each trial is an array of three integer numbers (drawn from 0 to 3). Two of the numbers are the same while one is different (target); the goal is to name the target number by pressing the corresponding key on the response keypad. For instance, the sequence “1 1 2” would have the number “2” as the target, since it is different from the others. However, the serial position of the numeral interferes with the stimulus-response mapping: “2” is the third in the sequence, triggering an automatic interference mapping to the third button, when in fact the second button is the correct one. Here, the stimuli can be grouped into four categories according to the patterns of interference: 1) zero interference (e.g., ‘1 0 0’); 2) identity interference only (e.g., ‘1 2 2’; the distractor ‘2’ is a member in the response options, but the position of the target number in the array is congruent with the answer key position); 3) position-interference only (e.g., ‘0 1 0’; the distractor ‘0’ is not a member in the response options, but the position of the target number is incongruent with its identity); and 4) dual interference (e.g., ‘2 1 2’ or ‘2 2 1’; presence of the distractor and the position of keypress is incongruent with the target position in the array). The presence of two types of interference provides a valuable opportunity to investigate whether there might be common or separate representations of cognitive interference in the human brain.

1.2.5 Value-based decision-making tasks

Although not typically considered in the cognitive control framework, value-based decision-making tasks have yielded insights into the validity of cognitive control as a general theoretical framework (Rangel and Hare, 2010, Hare et al., 2011). When learned values between possible options are very different, a choice can be made very easily with relatively strong confidence. However, when the learned values are close to each other, it becomes difficult to commit to a choice. Although the internal representation of choice values is stochastic and it is thus not possible that two options are of exactly the same values, the constraint that decisions need to be made within finite time probably requires a separate mechanism to hasten the choice in difficult cases. Cognitive control can be such a mechanism: when the decision difficulty is detected, cognitive control can be recruited to make an arbitration.

Many of the decision-making tasks with rewards, however, are motivated by existing theories in economics, ecology, and reinforcement learning and are designed to test specific hypotheses informed by these models. One important assumption is that decision is binary and sequential; it is made between a current and an alternative offer sequentially, instead of between many options all at the same time. Risks must be taken into consideration in the decision-making. This putative sequential nature also admits time preference as a risk factor, which is broadly considered in economic models. Building on top of this assumption is the binary meta-decision to forgo current decisions and search for alternatives, or stay with the current decisions. Given that choice value learning is indispensable for decision-making, this meta-decision can also be extended to include learning and cost components, and be framed as a decision to engage with model-free or model-based learning.

Given the success of these models in capturing the ethology of humans and animals, it is argued that specialized neural correlates of certain model assumption or parameters should exist. More specifically, neural correlates of risks (volatility), intertemporal choice, explore-exploit trade-off, and arbitration between model-free and model-based learning, among many other assumptions and parameters, should exist. Many tasks are thus developed to search for these variables. For example, motivated by optimal foraging theory in ecology, foraging tasks (see below) are designed to search for neural correlates of foraging choices. The assumption for such tasks is that the meta-decision to explore a new food patch or exploit the current patch is so important for the animal’s survival that the brain should develop dedicated neural mechanisms for the required computations. It remains a topic of debate whether or how cognitive control can provide more parsimony in explaining findings derived from these tasks.

These value-based decision-making tasks usually involve learning the values of various behavioral options, followed by a subsequent decision-making component. The choice options are arranged to fall into two types: the current default that has been chosen before with rewarding outcomes, and novel options with more risky outcomes. Subjects need to make a metacognitive decision whether to stick with their past rewarded choice options ('exploit'), or switch to new ones ('explore'). In general, the explore-exploit decision is based on two factors: devaluation in the currently rewarding option due to exploitation and risks associated with exploitation and exploration. The risks can come in the form of inherent uncertainty of a probabilistic reward source, or in the form of uncertainties in time, as the rewards only available after the decision to explore is made. In the foraging task, the subject first learns the values of a set of stimuli, and then engages in an explore-exploit choice task. In order to capture the cost incurred by the animal for leaving the default patch in a natural setting, the subject must forgo certain amount of the earned reward in order to explore. After deciding to exploit or explore, the subject choses between a pair of two options of different learned action-values that have different reward probabilities, revealed only after the engage or search choice has been made.

1.2.6 Response inhibition tasks

One frequently used task in the non-human primate (NHP) literature to study performance monitoring and cognitive control is the stop signal task (Logan and Cowan, 1984, Ito et al., 2003, Stuphorn et al., 2000). In this task, the macaque is trained to respond to the 'Go' cue by making a saccade (left or right). In a subset of trials, a 'Stop' signal appears with variable delays after the 'Go' cue, and the macaque should respond to the 'Stop' signal by withholding saccades. The difficulty of the trial can be parametrically adjusted by adjusting the 'Stop' signal delay (SSD): the longer the SSD, the more difficult it is for the macaque to cancel a prepared saccade (becoming impossible once the saccade is being made). One key benefit of using the stop signal task is that the error rate can be well controlled by adjusting the SSD, thus boosting the sample size for statistical analyses. Several aspects of cognitive control can be investigated using this task: errors are defined as the monkey's failure to stop and conflicts are defined as interference between the 'stopping' process and the 'go' process.

1.2.7 Response or rule switching tasks

This type of task usually involves the learning of multiple stimulus-response mapping rules and a switching between rules designated by the experimenter. The rule switch is communicated either explicitly via a sensory stimulus cue, or implicitly via a change in primary reinforcement.

One classic response-switching task is the pro-saccade/anti-saccade task. In the pro-saccade task, the subject is required to make a saccade to the target when it appears in either left or right side of the central fixation mark. In the anti-saccade task, the subject needs to saccade to the opposite side to the side where the target appears. In the anti-saccade task, saccades are significantly slower and there is a significant increase in error rate. This task is similar to the Simon task in the sense that it also involves a pre-potent response, which is to attend and respond to the appearance of sensory stimuli, and successful performance requires the suppression of this response.

In a task using implicit rule switching developed by Shima and colleagues (Shima and Tanji, 1998), the monkey learns to either push or turn a mechanical joystick in response to a visual trigger. After rewarding a series of pushing or turning movements, the experimenter reduces the reward, signaling the 'rule' switching to the monkey. After receiving the reduced reward, the monkey then voluntarily tries the other movement to increase possible reward. This incentive-based task switching thus shares features with explore/exploit switching in reward-based decision-making tasks more generally.

In a task using explicit rule switching developed by Isoda and colleagues (Isoda and Hikosaka, 2007), the monkey needs to make a saccade to either the right or the left depending on the color of the central cue. After a series of rewarded saccades to the same direction, the cue switches color to indicate a rule change. The error rate on the switch trial is 50%, and the reaction times are slower after successful cued-switching, a signature of interference. The Go/No-Go task is only slightly different from Isoda's task in that the cued responses are permitting and inhibiting a certain motor response, instead of making a left or right saccade.

The animal analogue of the Wisconsin Card Sorting Task ('WCST') shares some features of the rule-switching task, but with subtle differences (Mansouri et al., 2007). In the WCST, the monkey is trained to match the central target stimulus with one of the three flanking stimuli by either color or by shape. The matching rule switches between color matching and shape matching by the experimenter (conditional on the monkey's performance), unbeknownst to the monkey; the monkey has to infer which matching rule is in effect by outcome feedback. Here, there are two situations that differ in the patterns of interference: 1) high interference, when the flanking items contain two stimuli that each match with the central target in one of the dimensions (e.g. target is a blue square, one flanking stimulus is a red square and another stimulus is a blue circle); 2) low interference, when only one stimulus in the flanking items matches both of the color or shape dimension of the target stimulus. Since the matching rule is not revealed explicitly to the monkey, the two possible matching rules are in competition and the monkey has to resolve the conflicts between rules to select matching stimuli.

1.2.8 Commonalities among tasks

One common theme across the cognitive control tasks documented is situations where one needs to override learned responses in favor of a response that is appropriate given the current behavioral context. In the case of the Stroop task, one needs to override the pre-potent response of reading the word stimulus in order to name its font color. In the Simon task, the pre-potent response is pressing the button when the visual stimulus appears on the same side of the button; to achieve the goal in the incongruent trial one needs to override this tendency and press the button on the opposite side. In the Eriksen flanker task, the pre-potent response is to attend and respond to the flanking arrows by pressing the key that corresponds to the flanking arrow direction; one needs to suppress this tendency to be able to select the appropriate response (pressing the key that indicates the central arrow direction). In the MSIT task, the pre-potent responses are pressing the button that corresponds to the flanking numbers and/or one that corresponds to the position in the array where the target appears; these two pre-potent responses can map to the same or different motor output. The appropriate response requires the subject to override these pre-potent responses to report the identity of the target number. In the saccade countermanding task, the pre-potent response is to make a saccade when the ‘Go’ stimulus appear; the appropriate response is to suppress this pre-potent response so that it does not occur after the ‘Stop’ signal occurs. In tasks that involve rule or response switching, the reaction times are slower on the switch trials, similar to those in the Stroop task. The pre-potent response is specified by the previous response rule, and the appropriate response is to override the pre-potent action option in favor of the alternative response. In the WCST task, the pre-potent response is to select the stimulus according to the rule before rule switching, whereas the appropriate response is to inhibit and select a stimulus that matches with the target under the current rule.

1.3 Physiological analyses of performance monitoring and control

In this section, I will describe the basic physiological findings in human and animal studies concerning cognitive control; however, detailed interpretation and how different models were validated or invalidated by the findings listed here will be discussed in the following section (see ‘Existing models’).

1.3.1 Human studies

1.3.1.1 Error-related negativity and feedback-related negativity

A groundbreaking step taken towards understanding the human performance monitoring system was the discovery of an event-related scalp potential (ERP) named the ‘error-related negativity’ (ERN) (Gehring et al., 1993, Falkenstein et al., 1991). This signal is recorded using non-invasive EEG placed on the subject’s scalp and accompanies the subject’s erroneous action. The ERN provides the first measurable neural indices of the covert performance-monitoring process in humans. This opens up the possibility of connecting the complex behavioral signatures of performance monitoring with the dynamics of the ERN both in healthy subjects and in subjects with psychiatric disorders that are thought to feature abnormal monitoring and control (Olvet and Hajcak, 2008). The discovery of ERN led researchers to investigate whether the same signal also occurs when an error is specified externally by an explicit feedback. A signal with dynamics similar to that of ERN was discovered and named the feedback-related negativity (FRN). Such signals could be recorded in tasks where outcome is revealed after subjects make their choice, such as in a gamble that involves random pay off (Gehring and Willoughby, 2002), or estimating a time interval (Miltner et al., 1997). Notably, FRN have very similar distribution across scalp EEG electrodes to that of ERN (Potts et al., 2011). Several studies have aimed to infer the neural source of the ERN and FRN in the brain (Herrmann et al., 2004, Dehaene et al., 1994, Gehring et al., 2013, Ullsperger et al., 2014). Although the problem of inferring source locations using only scalp EEG topography and head shape is inherently an ill-posed one, this approach finds that placing two dipoles in the brain, in the dorsal anterior cingulate cortex (dACC) and the pre-supplementary motor area (pre-SMA), can explain a majority of the variance in the scalp ERN data (Dehaene et al., 1994, Herrmann et al., 2004). These two brain regions are the ones that will be the focus of my thesis.

1.3.1.2 Functional magnetic resonance imaging studies

The invention of functional magnetic resonance imaging (fMRI) in the mid-1990s has fundamentally revolutionized the field of cognitive neuroscience, and greatly advanced the understanding of the human performance monitoring system as well. fMRI monitors the level of oxygenation of hemoglobin (blood-oxygen-level dependent signal, or 'BOLD' signal) in the brain. When neurons are active, they consume energy and oxygen and recruit a hemodynamic response as a result, allowing indirect monitoring of the neural signals. Interestingly, since the BOLD-fMRI signal reflects the metabolic requirements of electrophysiological processing, the BOLD signal in fact mostly reflects synaptic activity and correlates better with field potentials than with single-unit action potentials (Logothetis and Wandell, 2004).

Using fMRI, several studies (Kerns et al., 2004, MacDonald et al., 2000, Carter et al., 1998, Botvinick et al., 1999, Ullsperger and von Cramon, 2001) have been able to confirm the involvement of dACC and pre-SMA in performance monitoring and cognitive control, but the interpretation of these BOLD signals have been a topic of intense debate. One historical difficulty has been that, unless the studies are very carefully designed, activation in particular of the dACC can be seen across very many different kinds of tasks. Equally problematic is a large bias in the publication literature that favors the dACC, among other regions cf. (Behrens et al., 2013), particularly the pre-SMA.

The basic finding is that there is BOLD activation in these areas on error trials, as well as on correct trials that have interference (Carter et al., 1998, Ullsperger and von Cramon, 2001, Kerns et al., 2004, MacDonald et al., 2000, Botvinick et al., 1999). A neural signature of the conflict adaptation effect has been identified in these pioneering studies: the BOLD activation in dACC is lower on a trial with conflict if it follows another conflict trial, than when it follows a non-conflict trial (Botvinick et al., 1999, Kerns et al., 2004). By contrast, dorsal lateral prefrontal cortex (dLPFC) has increased BOLD activation on a post-conflict conflict trial. The authors interpret this pattern of activation as dACC detecting conflict and recruiting dLPFC to implement cognitive control on the subsequent trial (see the section of 'Existing models'). One interesting study proposes that dACC monitors not conflict per se, but a likelihood of making errors. In this study, dACC was shown to be activated by a cue predicting the likelihood of error in the response inhibition task after training. Since the cue itself did not contain conflicting information, this was taken as evidence that dACC represented error likelihood and learned such association between action and outcome. Since on conflict trials, the likelihood of making an error is high as evident by the higher error rate, this study suggested that conflict signals could in fact be error likelihood signal.

One advantage of fMRI is its ability to localize the signal within certain brain regions, and when combined with simultaneous EEG recording this can provide relatively accurate temporal and anatomical information about cognitive processes. One study has used simultaneous fMRI-EEG recordings to study error processing and finds that one ICA component of the error-related EEG is correlated with the fRMI BOLD signal within dACC on a trial-by-trial basis. This is the first study that provided experimental evidence that the

ERN is tied to local activity within dACC. However, since the relation between single neuron activity and EEG and fMRI activity remains an open question, the exact meaning of this correlation remains to be investigated. One component of my thesis work below will make some direct comparisons between scalp-ERN, intracranial ERN in the dACC and pre-SMA, and single-unit responses in these brain regions to provide further insight into the standard scalp ERN.

1.3.1.3 Human intracranial studies

In recent years, technical advances together with clinical-research collaborations have made invasive electrophysiological recordings in humans possible (Fried et al., 2014, Engel et al., 2005), allowing the study of human cognitive functions at an unprecedented level of anatomical and temporal specificity. These studies are always carried out in a clinical setting. In this thesis, all work was carried out with patients with drug-intractable epilepsy. Patients are candidates for epilepsy surgery who undergo temporary surgical implantation of depth electrodes to further localize their seizures. During this time, it is possible to record, for research purposes, intracranial EEG and in some rare instances single neuron activity invasively from human brains. Dictated by its clinical nature, the electrode locations are dictated by clinical need alone. This technique doe thus not have whole-brain coverage like non-invasive techniques (fMRI, EEG, MEG). Experiments are performed in a hospital environment, which presents additional challenges in terms of experimental control. Also, due to their epilepsy, patients might have impaired cognitive function. All of these factors should be taken into account when interpreting results from intracranial studies. However, with proper control experiments in normal subjects, these studies can contribute greatly to the field of cognitive sciences and provide mechanistic understanding of human brain function not possible to achieve otherwise.

In the field of human performance monitoring, the intracranial approach so far has not been taken as frequently as in the field of memory studies, partly because of the rarity of patients that require electrode placement in the related brain regions. However, key insights have been yielded with just this handful of studies. One study (Bonini et al., 2014) investigated the neural signals of overt and covert errors using intracranial EEG recordings and EMG recordings and found that both SMA and the rostral part of dACC generated error-related ERPs (unlike the ERN, the polarity of the ERP is in fact positive, but this could be due to different electrode referencing layout and the electromagnetic properties of the tissue). This study reports that the error-related ERP occurs first in SMA and then in dACC, a finding that I confirm and further follow-up at the single-neuron level in this thesis.

Another key study (Sheth et al., 2012) reports conflict-sensitive single neurons in the human dACC. Patients who participated in this study underwent surgical ablation of dACC for the treatment of drug-intractable obsessive-compulsive disorder (OCD). These patients offered a rare opportunity to record single-neuron activity directly from dACC and monitor behavioral changes before and after the acute dACC lesion. The patients performed the MSIT task inside the operation room and exhibited typical behavioral slowing and conflict adaptation effects; both their RT and neuronal spiking patterns correlated with the three levels of interference the authors defined. This was the first evidence that single neurons within dACC carry a signature of conflict. However, the fact that the data in this study is obtained from patients with severe dACC dysfunction complicates the interpretation. Another problem with this study lies in the analysis method. Since conflict signals, according to the conflict monitoring theory, should have a stimulus-locked onset, this study rightly investigated the neural data aligned to this task event. However, there is inherent difficulty associated with this alignment: RTs differ greatly between the three interference levels. This poses a problem given that many neurons in dACC and pre-SMA had response-locked spiking patterns. For example, there are many neurons that keep spiking until an action is committed, at which point its spike rates are reduced significantly. For such neurons, a binning approach that is aligned to the stimulus onset will create spurious interference effect for the trivial reasons that spike rates differ significantly before and after an action and the RT differ between interference levels.

In summary, human intracranial studies have provided valuable insights on performance monitoring in humans. Evidence for a hierarchical error-monitoring system in MFC is strong. Conflict signals exist at the single neuron level in dACC, but failure to control for RT effects complicates the interpretation of such signals. Thus, further investigation using human single-unit recording is necessary to characterize the performance

1.3.1.3 Response inhibition literature

One large topic in the cognitive control literature is concerned with the implementation of cognitive control through response inhibition (Aron et al., 2016). The behavioral paradigm typically used is the stop signal task, where the subject is required to cancel a prepared action when an external ‘stop’ signal (usually auditory) comes on. The ‘reaction time’ to stop the action, which is the stop signal reaction time (SSRT), is inferred by a racing model with a ‘Go’ process racing against a ‘Stop’ process. Here, a cortical network has been identified that consists of right inferior frontal gyrus (rIFG) and right pre-SMA; this cortical network appears to command braking on ongoing actions, while the basal ganglia, specifically the subthalamic nucleus, implement the braking. Consistent with this framework, studies have found that BOLD activation in rIFG and STN was correlated with shorter SSRT, suggesting that subjects who were better at stopping actions had stronger activation in these stopping brain regions. Another study found that activity in the same network was correlated with the amount of conflict-induced slowing. Consistent with this braking network another group found that the white matter integrity between pre-SMA and the basal ganglia is correlated with the extent of the post-error slowing.

1.3.1.4 Lesion studies

Lesion studies provides a valuable tool to understand the causal roles of brain regions in the performance monitoring and cognitive control. Contrary to what has been suggested by the fMRI literature, studies have found that lesions in dACC do not comprise the Stroop performance (Stuss et al., 2001, Vendrell et al., 1995, Fellows and Farah, 2005). In a voxel-based lesion mapping study using a large sample of lesion patients, Stroop performance degradation is most associated with the restricted regions of the dLPFC, but not dACC (Glascher et al., 2012). Interestingly, one study even shows that patients with dACC lesions showed normal behavioral adjustments, such as the post-error slowing.

However, in the human intracranial single unit study mentioned above (Sheth et al., 2012), acute ablation of dACC impairs conflict adaption effects while leaving the conflict effects on RT intact. These studies seem to suggest that the role of dACC in performance monitoring could be more complex than just monitoring response conflicts. Some researchers, based on these lesion studies in humans and lesion/physiological studies in animal, argues that dLPFC instead of dACC monitors conflicts (Mansouri et al., 2007).

1.3.1.5 Functional manipulations

Non-invasive stimulation techniques have provided researchers with powerful tools to causally manipulate activity in the MFC and test for changes in the performance monitoring. One study uses transcranial direct-current stimulation (tDCS) to alter performance monitoring while the subjects perform a stop signal task and shows that the performance improves or deteriorates depending on the direction of current applied. Interestingly, the effect of stimulation modulates external feedback instead response-outcome processing. Several studies that use intracranial microstimulation of human pre-SMA have produced hesitation and alter manual and vocal responses in complex way, suggesting this brain region is capable of exerting inhibitory control. However, a direct demonstration of micro-stimulation induced alteration in performance monitoring has not been achieved.

1.3.2 Animal studies

Studies in nonhuman animals provide vastly improved access in electrophysiological investigations, with superior anatomical and temporal specificity that is often necessary to interrogate rapid decision-making processes. Using the saccade-countermanding task, Schall and colleagues have extensively characterized the macaque performance monitoring systems (Godlove et al., 2011, Stuphorn and Schall, 2006, Ito et al., 2003, Stuphorn et al., 2000, Emeric et al., 2008, Emeric et al., 2010). They found neurons that reported the macaque's failure to cancel a prepared saccade. These neurons increased their spike rates immediately after the stop-signal reaction time (SSRT), the inferred point in time where the 'stopping' process culminated, but before the reward feedback was revealed. These neurons were found in both dACC and supplementary eye fields (SEF), but with significance difference in latency: error neurons in dACC are activated later than those in SEF. The leading role of SEF in error monitoring holds true also at the level of local field potential (LFP). As discussed before, an important variable in these cognitive control tasks is the requirement to override a pre-potent response. The same studies also document neurons that carry a signature of interference by comparing successful 'stop' trials with 'go' trials, but this signature only existed in the SEF but not in dACC. A key feature of the error signal as well as the conflict signal identified in these studies is that they appear after the SSRT but *before* the reward feedback. Importantly, the fact that this conflict signal occurred *after* SSRT, the inferred point of successful stopping, makes it a qualitatively different signal from the conflict signal in humans, which occurred shortly after stimulus onset (Sheth et al., 2012). The average latency of the response is ~200 milliseconds, too long to result from a comparison with the efferent copy signal.

Given this, the authors conclude that these signals represent an internal monitoring mechanism of performance.

Given that conflict signals were found in humans and that animals are often faced with conflicting demands just like humans, the same signals should also be found in animals. However, the search for conflict signals in macaque dACC has not yielded convincing evidence that such signals exist. One study rigorously tested this idea by training the monkey to perform an antisaccade task, an eye-movement analogue of the Simon task (Nakamura et al., 2005). The authors noted that a signature of conflict does exist: the spike rates of certain SEF neurons were modulated by conflicts, but this modulation depended on the side where the visual stimulus appears. A pure conflict-signaling neuron should signal conflict no matter where the conflict-inducing visual stimulus appears, so the authors concluded that these neurons do not yet represent conflicts as such. This is in contrast to the aforementioned SEF conflict neuron documented by Schall and colleagues (Stuphorn et al., 2000). The authors suggest that the conflict neurons reported by Schall et al. did carry task-related information to some extent, but that the trial condition contrast used to select these neurons is confounded by the level of attention division. Importantly, on the population level, these SEF neurons that carry lateralized conflict information do summate to give elevated spike rates on the conflict trials. The authors propose that this summation of co-activated neurons could contribute to the apparent BOLD conflict signals reported in the human literature. In dACC, consistent with Schall et al. (Ito et al., 2003), the authors failed to find any signature of conflict either at the single neuron level or at the population level. One study claimed to find neurons in dACC that were sensitive to task conflict, but not action conflict.

Could the absence of conflict signals in dACC and in more dorsal parts of the medial frontal cortex (MFC) result from idiosyncrasies of the behavioral task designs? Could it be that the tasks used are not able to elicit conflict at the higher cognitive level and therefore fail to activate the MFC? One study (Ebitz and Platt, 2015) attempted to induce different kinds of conflicts by modifying the response inhibition task. Macaques were instructed to saccade to a visual target while suppressing a pre-potent response to saccade to a macaque face stimulus flashed briefly in the middle of the trial; the face stimulus could appear at a side of the central target that was either congruent or incongruent with the target saccade direction. The authors distinguished task conflict from action conflict: the former was the intrusion of a distractor (on both congruent and incongruent trials) while the latter was a competition between simultaneously activated action plans (on incongruent trials only). Referring to this definition, they found that dACC neurons reported task conflict but not action conflict, and that the pupil size was correlated with the levels of task conflict and dACC neuronal spike rates on the previous trial. However, the interpretation could be complicated by the fact that the task-conflict signal occurred after 200ms when most saccades concluded, and the significantly different saccade latencies were not controlled for across conditions. To challenge the macaque with interference at a higher level than response competition, one study (Mansouri et al., 2007) thus used the aforementioned WCST task. Here the conflict is conceptualized as the competition between two sets of matching rules as described above. The authors found that neurons in dLPFC, but not in dACC, decreased their spike rates on trials where there was competition between the matching rules. There were also neurons that increased their spike rates on the high-conflict trials, data that the authors did not include in the original paper. In addition, they also found neurons in dLPFC that tracked the level of rule conflicts on the immediately preceding trials, although the spike rate modulation appeared to be a persistent activity that started already before the onset of stimuli array. In this study, when the animal have dACC lesions, it performed the same way as the control intact animals did; only lesions in dLPFC compromised the conflict-induced behavioral adjustments. Thus, the authors made the strong claim that it was dLPFC but not dACC that detected conflict and implemented cognitive control. However, this claim seems to be at odds with a large literature documenting conflict signals in human dACC, and species difference could not be ruled out.

Studies that are traditionally concerned with reward-processing and motor planning/control can also be broadly analyzed in the framework of cognitive control. Here, the action outcomes, resulting from rules specified externally by the experimenter, are communicated to the animal via reward manipulations or explicit visual cues. The switching of a rule can signal the need to recruit cognitive control as the animal needs to explore new rules to maximize rewards. One study (Shima and Tanji, 1998) found that the dACC is causally involved in the exploration of alternative action (although in this study there are only two actions in the response set): dACC neurons encode the identity of actions that the animal was planning to switch to, triggered by a reduction of rewards, and ablation of dACC impairs the animal's ability to switch. In a similar vein, another study (Kennerley et al., 2006) finds that control animal's performance reaches asymptotic level. By contrast, when the animal's dACC is lesioned, it can still respond to reduction/omission of rewards ('error') by switching to an alternative action in the response set, suggesting that error detection is not destroyed by dACC lesion, but the animal's performance never reaches asymptotic level. Lesion animals also cannot match probability of reward by biasing their responses. The authors interpret this as evidence that dACC is crucial for integrating the reinforcement history to guide current action selection.

In summary, there is robust evidence that both dACC and more dorsal regions (supplementary eye field and pre-SMA) are sensitive to outcomes. Evidence for task-general conflict signals is weak; they mostly were manifested as a modulation on task-specific responses. By contrast, the role of dACC in reward outcome-based action selection is established. The task design difference and species difference could play a role in the discrepancies between macaque and human literature on conflict monitoring.

1.4 Theoretical accounts of dACC functions

Given the importance of dACC and its implications in a great variety of tasks, theoretical models were developed to try to provide a unifying framework that integrates all of the functions assigned to dACC. The detailed mechanism of how cognitive control is recruited by dACC and even the mechanisms of implementing cognitive control, however, have not been the main focus of these theoretical models. Many of the aforementioned tasks support (or were even designed to test) somewhat different theoretical accounts of dACC functions.

1.4.1 Conflict monitoring theory

In a pioneering connectionist model of the Stroop task, Cohen et al. put forward the idea that automaticity lies on a continuum and is malleable through learning (Cohen et al., 1990). In this model, the Stroop effect is explained as a competition between a more automatic word-reading response and a less automatic color-naming response. But the competition exists because the two pathways, one for word reading and one for color naming, map to common sets of behavioral outputs and therefore the co-activation of these two pathways (on an incongruent trial) leads to interference and slower activation for the motor units to reach their required response threshold.

The conflict monitoring theory is built upon these early models and has two main components. The first component argues that a primary function of dACC is to monitor conflicts, or 'crosstalk interference', between cognitive processes. Drawing on insights gained from information theoretic analyses of parallel processing systems, the theory aims to provide a theoretical framework for human cognitive processing, which is parallel in nature. Simply put, if two different tasks simultaneously require processing of the relevant stimuli in shared pathways, they will interfere with each other; the dACC generates a signal reflecting this processing, with magnitude commensurate with the level of interference. Of course, different tasks might share pathways at different points in the processing pipelines, but conflict is nevertheless triggered by external stimuli, and conflict timing is thus a function of stimulus onset. In the original formulation of the theory, the conflict signal is a domain-general index of the interference within neural pathways and thus does not depend on the peculiarities of the task (e.g. spatial location of the conflict-inducing stimulus). The theory proposes that dACC serves as a centralized conflict detector with the aforementioned properties.

The first key experimental evidence that supported the theory was the finding of higher BOLD activations in dACC on correct incongruent trials and on error trials. Since the ERN literature has proposed that dACC is a primary generator of the ERN, the fMRI findings suggest that there could be common mechanisms behind error and conflict monitoring. A connectionist modeling study aimed to integrate findings on ERN and the conflict signal reported in fMRI studies (Botvinick et al., 2001). In this study, conflict was defined as the product of activation values from two simultaneously active units. A unit is thought to signal conflict if its activation is a function of this product. Error was modeled as conflict between ongoing residual stimulus processing and the committed error response. With this assumption, the dynamics of the ERN can be produced in the activation profile of conflict-sensitive units in this network(Yeung et al., 2004).

1.4.2 Expected value of control (EVC) theory

This theory is an extension and improvement of the conflict monitoring theory, as it extents the range of scenarios where cognitive control is recruited: conflict is just one of the signal that dACC monitors, among many other, such as errors, negative feedback, difficulty, pain, etc. The EVC theory proposes that the primary function of dACC is to decide which and how much control is to be exerted (Shenhav et al., 2013). As in conflict monitoring theory, the key feature is a separation between the decision to engage control and the actual implementation of control. dACC is thought to be responsible for the former while dLFPC is responsible for the latter. In this theory, the dACC learns the EVC using both positive and negative outcome, the time-discounted EVC from the past, and the cost of control (which is a function of the control signal itself). It then specifies the identity as well as the intensity of control (e.g. which task to perform, or which choice to make) based on the current EVC. This theory anchors well to several key experimental findings. First, the chosen and alternative value signals in dACC is taken to be evidence that support the role of dACC in maintaining EVC for each possible control signals and determine which of the choices these values correspond to can lead to maximal expected rewards. Second, the prediction error signals prominent in dACC are taken as evidence for the EVC updating process (similar to the temporal difference algorithm in reinforcement learning). Third, the fact that dACC maintains information about past outcomes and intertemporal choice is captured by recursive incorporation of time-discounted EVC. Fourth, the incorporation of cost in the EVC speaks to the findings where subjects learns the cognitive efforts needed and tends to avoids tasks with higher learned efforts if given the choice. This powerful framework is able to explain a wide variety of experimental findings and provide specific testable predictions.

1.4.3 Action-outcome predictor

An elegant model (Alexander and Brown, 2011) argues that the major function of dACC can be as simple as learning action – outcome associations and making predictions of action outcomes. The theory was built upon a previous fMRI findings that dACC responded to cues indicating error likelihood (Brown and Braver, 2005). The model generalizes the standard reinforcement learning algorithms and uses temporal difference learning law, but differs from reinforcement learning in several important ways. The model does not learn stimulus-response mapping, but instead learns the response-outcome mappings, and keeps a separate prediction error for each possible outcomes (instead of one scalar prediction error as in a standard framework). The prediction error here refers specifically to a mismatch between the predicted outcome and the actual outcome, instead of a mismatch between the predicted value and actual value of a stimulus. Models implemented according to this theory have successfully reproduced a wide range of experimental findings, such as activity profile of ERN and error-signaling neurons, and the sensitivity to volatility of reinforcement contingencies.

1.5 Motivation and open questions

Over the past twenty years, consensus has been reached that cognitive control is a useful and powerful framework to conceptualize the mechanisms required to enable deliberate goal pursuit, or even more broadly, volitional behavior. There is converging evidence that frontal regions (MFC and dLPFC), parietal regions, and the basal ganglia are involved in cognitive control. There is a wealth of behavioral and neuroscience data in both humans and animals supporting these broad conclusions. Several elegant theoretical frameworks attempt to tie together these findings and to provide normative models. However, key open questions remain, in large part because the precision and resolution of measures available, especially in humans, has so far been insufficient to provide detailed causal mechanisms. Here I sketch the open questions that motivate the studies I carried out. More detailed motivation and background related to these open questions will be provided in the relevant chapters that follow.

What is the relation between error signals at different scales? In humans, error signals reported so far are all at the macroscopic level: it involves the bulk activity of many neurons. ERN is an event-related potential and reflects the summation of synaptic activity from a large number of pyramidal neurons in the cortex. Error-related BOLD activation also reflects the activity of a large number of neurons. However, the relationship between these macroscopic signals and spiking activity of neurons at the microscopic level is unclear. This open question was the motivation for Chapter 2 (c.f. Figure 2.6a).

On a single neuron level, can performance monitoring or cognitive control signals be represented by a temporal code? It is well established that neurons not only carry information by modulating their spike rates (rate code), but also by modulating the precise timings of each spike. Given that oscillatory activity are prominent and ubiquitous in the brain, it can serve as a reference frame for spike timings. The phase relationship between spikes and brain oscillations might code information about cognitive processes and might be relevant for coding information about action monitoring and control. This open question was the motivation for Chapter 3.

Could performance monitoring or cognitive control signals be represented by a population code? Neurons does not function alone; its activity is related and constrained by its neighboring neurons. The joint activity pattern between neurons could carry information important for cognitive control. As mentioned above, MFC, especially dACC, subserves a variety of functions important for controlling behavior in the service of a goal. This complex nature of goal-directed behavior requires that the neuronal population in these brain regions needs to code multiple pieces of information at the same time. This open question was the motivation for Chapter 4.

How is cognitive control implemented? In addition, what would be possible neural network architecture that enable cognitive control? Existing studies have provided with several possibilities: 1) adjustments of the parameters of the sensory evidence accumulation (e.g. the decision threshold); 2) increase of attention; 3) inhibition of specific motor responses; 4) broad and unspecific adjustment of motor cortex excitability. These accounts can all be broadly categorized as diffusion-to-bound gating mechanisms. However, the difficulty with these accounts is that cognitive control seem to be task-specific (does not generalize easily to other tasks) but at the same time general-purpose (it can be applied to many different scenario), yet these accounts lack this flexibility. After all, we do not miraculously become better at driving when we practice Stroop task more. Many of the theories mentioned above also point to the dLFPC as the primary locus that implements cognitive control. It is thus implied that a centralized ‘controller’, situated in dLPFC, can exert control through all the means mentioned above. However, there is convincing evidence that it is rIFG-pre-SMA-basal ganglia network that is shown to be associated with response inhibition, but not dLPFC. What other kinds of cognitive control does dLFPC implement?

Given the apparent specificity of cognitive control, how is any centralized controller able to specify the parameters for a certain task-relevant neural pathways out of so many other pathways? The specificity of cognitive control implies that cortex might be organized in such a way that decentralized control is readily installed in the circuit motifs. The central controller then needs only to broadcast a general alert and the decentralized control units can be evoked in a pathway-specific manner. However, the ultimate question remains of who controls the centralized controller, or the dLPFC? The study of cognitive control might shed light on general principles of controlling complex networked dynamical systems.

What is the relation between learning and cognitive control? From experimental findings, cognitive control seems to have a broad range of time constants. Within trial, cognitive control can rapidly suppress goal-incompatible actions, leading to covert errors (no overt execution, but partial activation of incorrect response); in a stop task, it can be engaged to inhibit a response rapidly. However, cognitive control can also be engaged across trials: the post-error slowing occur seconds after an error is made (on the next trial) and many of the trial-by-trial RT adaptation effect also occurs on the scale of seconds. It could be that at different time scales there are completely different mechanisms for implementing control. For example, for the longer across-trial control, a finite state machine implemented by recurrent network can be used, whereas for short timescale control, possible mechanisms could be adjustment of evidence accumulation process via a change in local excitatory and inhibitory balance. However, could cognitive control involve structural changes such as synaptic plasticity just as proposed in the existing theory of learning?

All of these questions require data with resolution that matches with the timescale of a single spike to answer. I will thus try to tackle some of these questions with human intracranial single-neuron recordings, taking advantage the superior temporal and spatial specificity required by the nature of these problems.

References

Alexander, W. H. & Brown, J. W. 2011. Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience,* 14**,** 1338-U163.

Anderson, S. W., Damasio, H., Jones, R. D. & Tranel, D. 1991. Wisconsin Card Sorting Test-Performance as a Measure of Frontal-Lobe Damage. *Journal of Clinical and Experimental Neuropsychology,* 13**,** 909-922.

Aron, A. R., Herz, D. M., Brown, P., Forstmann, B. U. & Zaghloul, K. 2016. Frontosubthalamic Circuits for Control of Action and Cognition. *J Neurosci,* 36**,** 11489-11495.

Behrens, T. E. J., Fox, P., Laird, A. & Smith, S. M. 2013. What is the most interesting part of the brain? *Trends in Cognitive Sciences,* 17**,** 2-4.

Bonini, F., Burle, B., Liegeois-Chauvel, C., Regis, J., Chauvel, P. & Vidal, F. 2014. Action monitoring and medial frontal cortex: leading role of supplementary motor area. *Science,* 343**,** 888-91.

Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S. & Cohen, J. D. 1999. Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature,* 402**,** 179-81.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S. & Cohen, J. D. 2001. Conflict monitoring and cognitive control. *Psychol Rev,* 108**,** 624-52.

Brown, J. W. & Braver, T. S. 2005. Learned predictions of error likelihood in the anterior cingulate cortex. *Science,* 307**,** 1118-1121.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D. & Cohen, J. D. 1998. Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science,* 280**,** 747-9.

Cohen, J. D., Dunbar, K. & McClelland, J. L. 1990. On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychol Rev,* 97**,** 332-61.

Dehaene, S., Posner, M. I. & Tucker, D. M. 1994. Localization of a Neural System for Error-Detection and Compensation. *Psychological Science,* 5**,** 303-305.

Ebitz, R. B. & Platt, M. L. 2015. Neuronal activity in primate dorsal anterior cingulate cortex signals task conflict and predicts adjustments in pupil-linked arousal. *Neuron,* 85**,** 628-40.

Emeric, E. E., Brown, J. W., Leslie, M., Pouget, P., Stuphorn, V. & Schall, J. D. 2008. Performance monitoring local field potentials in the medial frontal cortex of primates: anterior cingulate cortex. *J Neurophysiol,* 99**,** 759-72.

Emeric, E. E., Leslie, M., Pouget, P. & Schall, J. D. 2010. Performance monitoring local field potentials in the medial frontal cortex of primates: supplementary eye field. *J Neurophysiol,* 104**,** 1523-37.

Engel, A. K., Moll, C. K., Fried, I. & Ojemann, G. A. 2005. Invasive recordings from the human brain: clinical insights and beyond. *Nat Rev Neurosci,* 6**,** 35-47.

Eriksen, B. A. & Eriksen, C. W. 1974. Effects of Noise Letters Upon Identification of a Target Letter in a Nonsearch Task. *Perception & Psychophysics,* 16**,** 143-149.

Fellows, L. K. & Farah, M. J. 2005. Is anterior cingulate cortex necessary for cognitive control? *Brain,* 128**,** 788-796.

Fried, I., Rutishauser, U., Cerf, M. & Kreiman, G. 2014. *Single neuron studies of the human brain : probing cognition,* Cambridge, Massachusetts, The MIT Press.

Friederici, A. D. 2017. *Language in our brain : the origins of a uniquely human capacity,* Cambridge, Massachusetts, The MIT Press.

Fuster, J. M. 2015. *The prefrontal cortex,* Amsterdam ; Boston, Elsevier/AP, Academic Press is an imprint of Elsevier.

Gehring, W. J., Liu, Y., Orr, J. M. & Carp, J. 2013. The error-related negativity (ERN/Ne). *In:* Luck, S. J. & Kappenman, E. S. (eds.) *The Oxford handbook of event-related potential components.* Oxford: Oxford University Press.

Gehring, W. J. & Willoughby, A. R. 2002. The medial frontal cortex and the rapid processing of monetary gains and losses. *Science,* 295**,** 2279-82.

Glascher, J., Adolphs, R., Damasio, H., Bechara, A., Rudrauf, D., Calamia, M., Paul, L. K. & Tranel, D. 2012. Lesion mapping of cognitive control and value-based decision making in the prefrontal cortex. *Proc Natl Acad Sci U S A,* 109**,** 14681-6.

Godlove, D. C., Emeric, E. E., Segovis, C. M., Young, M. S., Schall, J. D. & Woodman, G. F. 2011. Event-related potentials elicited by errors during the stop-signal task. I. Macaque monkeys. *J Neurosci,* 31**,** 15640-9.

Gold, J. M., Carpenter, C., Randolph, C., Goldberg, T. E. & Weinberger, D. R. 1997. Auditory working memory and Wisconsin Card Sorting Test performance in schizophrenia. *Archives of General Psychiatry,* 54**,** 159-165.

Gratton, G., Coles, M. G. H. & Donchin, E. 1992. Optimizing the Use of Information - Strategic Control of Activation of Responses. *Journal of Experimental Psychology-General,* 121**,** 480-506.

Hare, T. A., Schultz, W., Camerer, C. F., O'Doherty, J. P. & Rangel, A. 2011. Transformation of stimulus value signals into motor commands during simple choice. *Proceedings of the National Academy of Sciences of the United States of America,* 108**,** 18120-18125.

Herrmann, M. J., Rommler, J., Ehlis, A. C., Heidrich, A. & Fallgatter, A. J. 2004. Source localization (LORETA) of the error-related-negativity (ERN/Ne) and positivity (Pe). *Brain Res Cogn Brain Res,* 20**,** 294-9.

Ito, S., Stuphorn, V., Brown, J. W. & Schall, J. D. 2003. Performance monitoring by the anterior cingulate cortex during saccade countermanding. *Science,* 302**,** 120-2.

Kennerley, S. W., Walton, M. E., Behrens, T. E., Buckley, M. J. & Rushworth, M. F. 2006. Optimal decision making and the anterior cingulate cortex. *Nat Neurosci,* 9**,** 940-7.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., 3rd, Cho, R. Y., Stenger, V. A. & Carter, C. S. 2004. Anterior cingulate conflict monitoring and adjustments in control. *Science,* 303**,** 1023-6.

Logothetis, N. K. & Wandell, B. A. 2004. Interpreting the BOLD signal. *Annu Rev Physiol,* 66**,** 735-69.

MacDonald, A. W., 3rd, Cohen, J. D., Stenger, V. A. & Carter, C. S. 2000. Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science,* 288**,** 1835-8.

MacLean, E. L., Hare, B., Nunn, C. L., Addessi, E., Amici, F., Anderson, R. C., Aureli, F., Baker, J. M., Bania, A. E., Barnard, A. M., Boogert, N. J., Brannon, E. M., Bray, E. E., Bray, J., Brent, L. J. N., Burkart, J. M., Call, J., Cantlon, J. F., Cheke, L. G., Clayton, N. S., Delgado, M. M., DiVincenti, L. J., Fujita, K., Herrmann, E., Hiramatsu, C., Jacobs, L. F., Jordan, K. E., Laude, J. R., Leimgruber, K. L., Messer, E. J. E., Moura, A. C. D., Ostojic, L., Picard, A., Platt, M. L., Plotnik, J. M., Range, F., Reader, S. M., Reddy, R. B., Sandel, A. A., Santos, L. R., Schumann, K., Seed, A. M., Sewall, K. B., Shaw, R. C., Slocombe, K. E., Su, Y. J., Takimoto, A., Tan, J. Z., Tao, R., van Schaik, C. P., Viranyi, Z., Visalberghi, E., Wade, J. C., Watanabe, A., Widness, J., Young, J. K., Zentall, T. R. & Zhao, Y. N. 2014. The evolution of self-control. *Proceedings of the National Academy of Sciences of the United States of America,* 111**,** E2140-E2148.

MacLeod, C. M. 1991. Half a century of research on the Stroop effect: an integrative review. *Psychol Bull,* 109**,** 163-203.

Mansouri, F. A., Buckley, M. J. & Tanaka, K. 2007. Mnemonic function of the dorsolateral prefrontal cortex in conflict-induced behavioral adjustment. *Science,* 318**,** 987-90.

Metcalfe, J. 2008. Evolution of metacognition. *Handbook of metamemory and memory.* New York, NY, US: Psychology Press.

Miltner, W. H., Braun, C. H. & Coles, M. G. 1997. Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a "generic" neural system for error detection. *J Cogn Neurosci,* 9**,** 788-98.

Nakamura, K., Roesch, M. R. & Olson, C. R. 2005. Neuronal activity in macaque SEF and ACC during performance of tasks involving conflict. *J Neurophysiol,* 93**,** 884-908.

Notebaert, W., Houtman, F., Opstal, F. V., Gevers, W., Fias, W. & Verguts, T. 2009. Post-error slowing: an orienting account. *Cognition,* 111**,** 275-9.

Passingham, R. E. & Wise, S. P. 2012. *The neurobiology of the prefrontal cortex : anatomy, evolution, and the origin of insight,* Oxford, United Kingdom, Oxford University Press.

Potts, G. F., Martin, L. E., Kamp, S. M. & Donchin, E. 2011. Neural response to action and reward prediction errors: Comparing the error-related negativity to behavioral errors and the feedback-related negativity to reward prediction violations. *Psychophysiology,* 48**,** 218-228.

Purcell, B. A. & Kiani, R. 2016. Neural Mechanisms of Post-error Adjustments of Decision Policy in Parietal Cortex. *Neuron,* 89**,** 658-71.

Rangel, A. & Hare, T. 2010. Neural computations associated with goal-directed choice. *Curr Opin Neurobiol,* 20**,** 262-70.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A. & Nieuwenhuis, S. 2004. The role of the medial frontal cortex in cognitive control. *Science,* 306**,** 443-7.

Schneider, W. & Shiffrin, R. M. 1977. Controlled and Automatic Human Information-Processing .1. Detection, Search, and Attention. *Psychological Review,* 84**,** 1-66.

Shenhav, A., Botvinick, M. M. & Cohen, J. D. 2013. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron,* 79**,** 217-40.

Sheth, S. A., Mian, M. K., Patel, S. R., Asaad, W. F., Williams, Z. M., Dougherty, D. D., Bush, G. & Eskandar, E. N. 2012. Human dorsal anterior cingulate cortex neurons mediate ongoing behavioural adaptation. *Nature,* 488**,** 218-21.

Shiffrin, R. M. & Schneider, W. 1977. Controlled and Automatic Human Information-Processing .2. Perceptual Learning, Automatic Attending, and a General Theory. *Psychological Review,* 84**,** 127-190.

Shima, K. & Tanji, J. 1998. Role for cingulate motor area cells in voluntary movement selection based on reward. *Science,* 282**,** 1335-8.

Smith, J. D., Shields, W. E. & Washburn, D. A. 2003. The comparative psychology of uncertainty monitoring and metacognition. *Behav Brain Sci,* 26**,** 317-39; discussion 340-73.

Stroop, J. R. 1935. *Studies of interference in serial verbal reactions.* Ph D, George Peabody College for Teachers.

Stuphorn, V. & Schall, J. D. 2006. Executive control of countermanding saccades by the supplementary eye field. *Nat Neurosci,* 9**,** 925-31.

Stuphorn, V., Taylor, T. L. & Schall, J. D. 2000. Performance monitoring by the supplementary eye field. *Nature,* 408**,** 857-60.

Stuss, D. T., Floden, D., Alexander, M. P., Levine, B. & Katz, D. 2001. Stroop performance in focal lesion patients: dissociation of processes and frontal lobe lesion location. *Neuropsychologia,* 39**,** 771-86.

Ullsperger, M., Danielmeier, C. & Jocham, G. 2014. Neurophysiology of performance monitoring and adaptive behavior. *Physiol Rev,* 94**,** 35-79.

Ullsperger, M. & von Cramon, D. Y. 2001. Subprocesses of performance monitoring: a dissociation of error processing and response competition revealed by event-related fMRI and ERPs. *Neuroimage,* 14**,** 1387-401.

Vendrell, P., Junque, C., Pujol, J., Jurado, M. A., Molet, J. & Grafman, J. 1995. The role of prefrontal regions in the Stroop task. *Neuropsychologia,* 33**,** 341-52.

Yeung, N., Botvinick, M. M. & Cohen, J. D. 2004. The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychol Rev,* 111**,** 931-59.

Yeung, N. & Summerfield, C. 2012. Metacognition in human decision-making: confidence and error monitoring. *Philos Trans R Soc Lond B Biol Sci,* 367**,** 1310-21.

Chapter 2. Single-neuron Correlates of Error Monitoring and Post-error Adjustments in Human Medial Frontal Cortex

The following chapter is adapted from Fu et al., 2018 and modified according to the format of Caltech Thesis.
Fu et al., Single-neuron Correlates of Error Monitoring and Post-error Adjustments in Human Medial Frontal Cortex, Neuron (2018),
https://doi.org/10.1016/j.neuron.2018.11.016

2.1 Abstract

Humans can self-monitor errors without explicit feedback, resulting in behavioral adjustments on subsequent trials such as post-error slowing (PES). The error-related negativity (ERN) is a well-established macroscopic scalp EEG correlate of error self-monitoring, but its neural origins and relationship to PES remain unknown. We recorded in the frontal cortex of patients performing a Stroop task and found neurons that track self-monitored errors and error history in dorsal anterior cingulate cortex (dACC) and pre-supplementary motor area (pre-SMA). Both the intracranial ERN (iERN) and error neuron responses appeared first in pre-SMA, and ~50ms later in dACC. Error neuron responses were correlated with iERN amplitude on individual trials. In dACC, such error neuron-iERN synchrony and responses of error-history neurons predicted the magnitude of PES. These data reveal a human single-neuron correlate of the ERN and suggest that dACC synthesizes error information to recruit behavioral control through coordinated neural activity.

2.2 Introduction

A fundamental feature of behavior is the ability to optimize performance based on outcomes (Ullsperger et al., 2014). In humans, performance failure can be monitored not only by explicit external feedback, but also through self-monitoring in the absence of such feedback. Successful detection of errors then initiates behavioral adjustments on various timescales. These include within-trial adjustment such as on-line error avoidance (leading to 'covert errors') (Bonini et al., 2014) and immediate correction of the response (Rabbitt, 1966), next-trial adjustment that requires cognitive control such as delaying an impending action (Laming, 1979, Ridderinkhof et al., 2004, Ullsperger et al., 2014), as well as more deliberate adjustments that span several trials to maximize potential rewards (Frank et al., 2005, Quilodran et al., 2008, Shima and Tanji, 1998).

Previous work on identifying the neural substrates for the different components of this behavioral feedback-control loop has revealed that the medial frontal cortex (MFC), which includes the dorsal anterior cingulate cortex (dACC, sometimes also referred to as anterior mid-cingulate cortex (Vogt et al., 2003)) and the pre-supplementary motor area (pre-SMA), serves a critical role for both self-monitoring and control of actions (Ullsperger et al., 2014). While self-monitored errors are robustly signaled by the error-related negativity (ERN) (Gehring et al., 1993, Burle et al., 2008, Godlove et al., 2011, Falkenstein et al., 1991), no single-neuron correlates of this process have yet been reported in humans.

A second large topic concerns the changes in cognitive control that ensue either as a consequence of ongoing prediction of action outcomes, or subsequent to having detected an outcome such as an error. The MFC is also crucially involved in these processes (Kolling et al., 2016, Rushworth and Behrens, 2008, Kerns et al., 2004, Behrens et al., 2007, Brown and Braver, 2005, Shenhav et al., 2013, Sheth et al., 2012, Alexander and Brown, 2011). Such control mechanisms can either trigger switching to a different action based on estimated action values, or influence the production of an action, such as delaying an action or adjusting the force with which an action is executed (Gehring et al., 1993, Ullsperger et al., 2014). As an example for the former, MFC neurons encode plans to switch to the alternative action triggered by a reduction of reward (Shima and Tanji, 1998, Williams et al., 2004, Kennerley et al., 2006). Similarly, MFC neurons signal the need to switch saccade directions in response to an externally cued rule change (Isoda and Hikosaka, 2007). Lesioning or pharmacological manipulation of the MFC disrupt such reward history dependent action selection (Shima and Tanji, 1998, Kennerley et al., 2006), illustrating a critical role for the MFC in explore-exploit decisions.

Less is known about the MFC's involvement in control of action production triggered by monitored outcomes (mentioned above as the second type of behavioral adjustments). In the case of externally cued response inhibition, electrical stimulation of the supplementary eye field or pre-SMA has been shown to delay saccades in service of avoiding errors (Stuphorn and Schall, 2006, Isoda and Hikosaka, 2007). These studies provide crucial causal evidence that MFC can influence action production, but the neuronal mechanisms that bridge monitoring to such control and the possible roles of other brain regions in this process remain unclear. Self-monitored errors, on the other hand, have a typical behavioral consequence: they can delay successive actions, a phenomenon known as the post-error slowing ('PES') (Ullsperger et al., 2014). Functional imaging studies have revealed the complex neural mechanism that may underlie PES with MFC being the central node of this control network. In this framework, the need for PES is signaled by MFC after detection of an error. PES involves inhibitory activity in the cortico-subthalamic pathways (Danielmeier et al., 2011, Aron and Poldrack, 2006, Aron et al., 2007), as well as adaptations in motor cortex (Danielmeier et al., 2011) and sensory processing and integration regions (Purcell and Kiani, 2016, Ullsperger and Danielmeier, 2016, King et al., 2010). This argument is principally supported by the finding that BOLD activation in dACC is correlated with the magnitude of PES (Kerns et al., 2004). In addition, in rodents, pharmacological inactivation of MFC abolishes PES (Narayanan et al., 2013).

A natural hypothesis thus links the detection of self-generated errors, as reflected in the ERN, with changes in cognitive control, as exhibited behaviorally in PES, predicting that the two measures should be correlated. However, several EEG studies have failed to find a significant relationship between PES and ERN (Gehring and Fencsik, 2001, Nieuwenhuis et al., 2001, Hajcak et al., 2003). Curiously, while BOLD activity in MFC predicts PES, the ERN does not. Based on these discrepancies in the literature, we tested a more detailed mechanistic hypothesis that might reconcile them. The ERN is thought to be produced by the summation of postsynaptic potentials within MFC and may thus, in part, reflect inputs to this region (Holroyd and Coles, 2002, Luck, 2014). One possibility explaining the aforementioned discrepancies is that the inputs to the MFC that produce the ERN only carry information about error monitoring, but not about the engagement of control. The computations within MFC that underlie cognitive control, while not reflected in the ERN, might instead be evident in oscillatory components in the local field potential (LFP) (Siegel et al., 2012, Pesaran et al., 2018) or in correlations between spike rates of neurons and the LFP (Nir et al., 2007). Such correlated neuronal activity could also explain why BOLD signals are associated with PES (Niessing et al., 2005). This model predicts that spike rates and iEEG power within MFC would be correlated with the strength of PES, even though the ERN is not.

2.3 Results

2.3.1 Task and behavior

Subjects performed a color-naming Stroop task, which required subjects to name the color of words while ignoring their semantic meaning (Fig. 2.1a). RTs were longer on word-color incongruent trials than word-color congruent trials (the "Stroop effect"; 224.9 ± 19.2 ms difference, mean ± s.e.m. across sessions, F (1, 84) = 116.6, p < 0.001, mixed-effects one-way ANOVA). Subjects responded incorrectly ('error trials') in 7.2 ± 0.5 % (± s.e.m) of all trials. On correct trials that follow an error ('EC' trials), responses were significantly slower than on correct trials that follow another correct trial ('CC' trials) (Fig. 2.1b, amount of PES: 64.3 ± 11.0 ms, mean ± s.e.m. across sessions, mixed-effect one-way ANOVA, F (1,184) = 23.4, p < 0.001). To quantify PES for individual trials in the analysis below, we used sequences of 'CCEC' trials ('C' represents correct trials, 'E' represents error trials, see Methods; median single-trial PES = 33ms, p = 0.0016, z = 3.154; signed rank test).

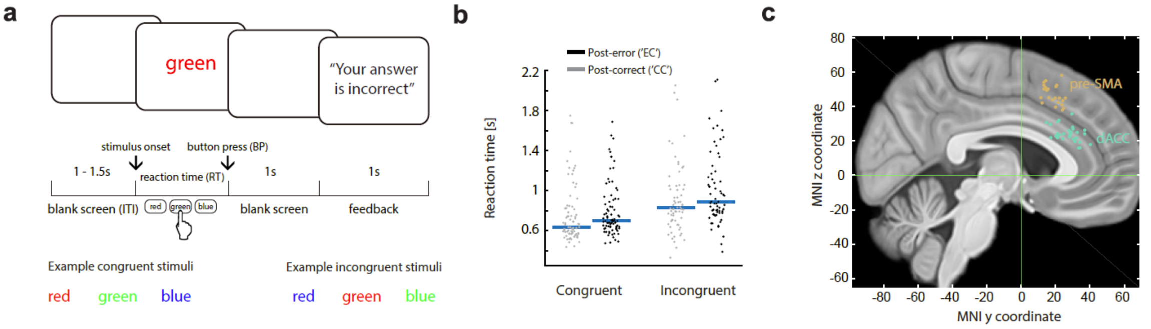

Figure 2.1 Task, behavior, and electrode localization
(a) Task structure.
(b) Behavior. Each dot represents the mean RT of 'EC' or 'CC' trials of a session.
(c) Recording locations, projected onto the x=5 mm slice. Each dot represents the location of a micro-wire bundle in a patient.
See also Figure S1.

2.3.2 Single-neuron correlates of error self-monitoring

We isolated 1171 single units from dACC (n = 618) and pre-SMA (n = 553) across 29 patients (Fig. 2.11c, Table S1; see also Fig. S1a-c and Fig. S1d-i). Some neurons were in sessions with fewer than seven error trials and thus were excluded from the analyses that involve errors (number of neurons included in dACC is n=399; in pre-SMA is n=431). Error neurons were identified using a Poisson regression model. Spike rates in a one-second epoch starting immediately after the action (button press), but before feedback, were regressed against trial labels (‘error’ or ‘correct’) and RTs. 34% (n = 134) of dACC and 46% (n = 198) of pre-SMA neurons signaled errors (see Fig. 2.2a-d, Fig. 2.3a, Fig. 2.3b-c, Fig. S2c-d and Table S2). We classified error neurons based on whether they had higher (“Type I”, error > correct, n = 99 and 118 in dACC and pre-SMA, respectively; see Fig. 2.2a,c and Fig. 2.3b,c, left) or lower (“Type II”, error < correct, n = 35 and 80 in dACC and pre-SMA, respectively, Fig. 2.2b,d and Fig. 2.3b,c, right) spike rates for error than correct trials. The responses of error neurons on individual trials differed reliably between error and correct trials as evaluated using receiver operating characteristic (ROC) analysis (see methods and Fig. 2.3f): Area-under-the-curve (AUC) values were, on average, 0.61 and 0.60 for dACC and pre-SMA, respectively (significantly greater than 0.5 with $p < 10^{-10}$, $t(133) = 12.86$ and $p < 10^{-10}$, $t(197) = 18.5$, respectively; t-test). AUC values of error neurons did not differ significantly between dACC and pre-SMA (Fig. 2.3f; $p = 0.52$, $t(330) = 0.64$, t-test).

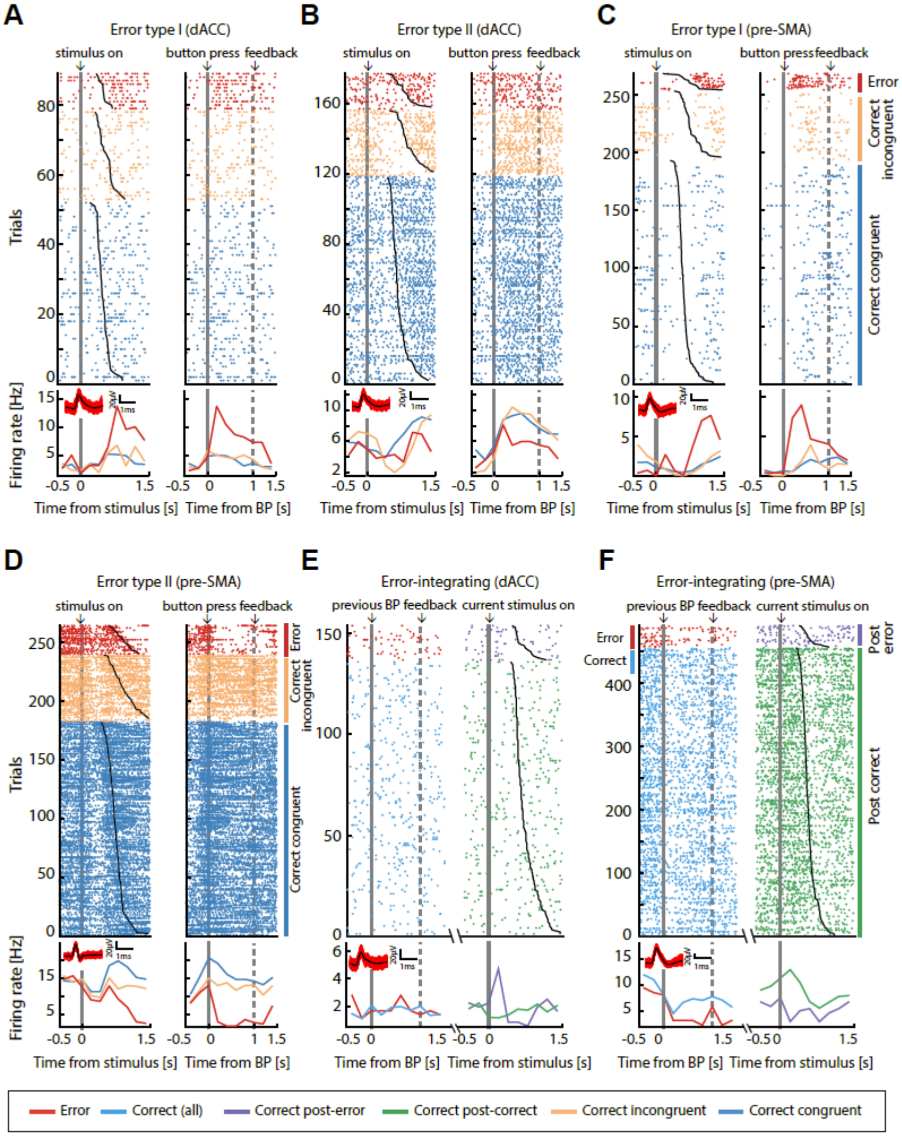

Figure 2.2 Examples of error and error-integrating neurons
(a-d) Error neurons (e-f) Error-integrating neurons. (a-f) Raster (top) and mean spike rates (bottom) aligned at stimulus onset (left) and button press (right; 'BP') for (a-d); aligned to previous-trial button press (left) and to current-trial stimulus onset (right) for (e-f). Trials are sorted by RT (black line overlaying raster plots) and trial types (color; from top to bottom, error, correct incongruent, correct congruent for (a-d); 'EC' and 'CC' trials for (e-f)). Solid gray bars, time points for alignments. Broken gray bars, onset of feedback. Insets show the waveforms associated with each neuron and the corresponding scale bars.

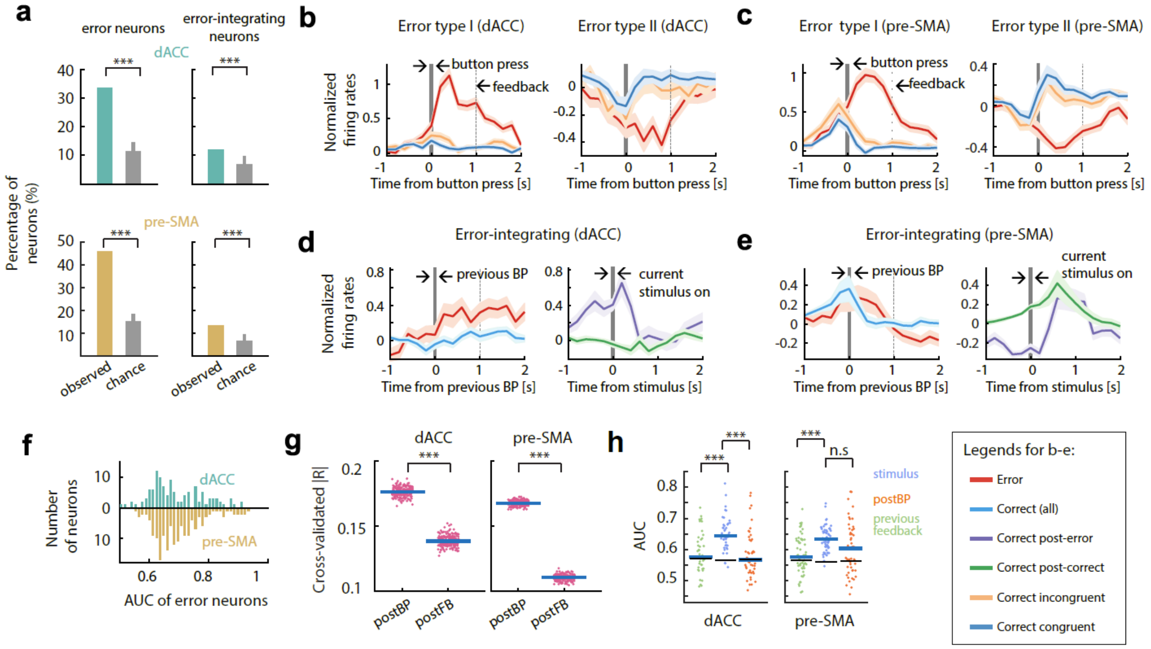

Figure 2.3 Temporal profile of error and error-integrating neurons
(a) Percentage of significant error and error-integrating neurons in dACC and pre-SMA. Gray bar is null distribution (mean and 95% confidence interval).
(b) Average standardized spike rates for all dACC error neurons, aligned at button press (t=0, gray bar). Broken bars, 1s after button press. Shading is ± s.e.m. across neurons.
(c) Same as (b), but for pre-SMA error neurons.
(d) Average standardized spike rates as a function of time for dACC error-integrating neurons, aligned at preceding-trial button press (left) or current-trial stimulus onset (right).
(e) Same as (d) but for the pre-SMA.
(f) ROC analysis. Error signal can be reliably decoded at the single-trial level (Type I and Type II error neurons pooled).
(g) Statistics for (b-c). Error neurons distinguished between error and correct trials more strongly after button press compared to after onset of feedback. Shown are cross-validated partial correlation coefficients across all error neurons (Type I and II pooled). Each data point represents the mean effect size across all error neurons in one cross-validation run.
(h) Statistics for (d-e). ROC analysis of the response of error-integrating neurons in three different time windows. The spike rates of error-integrating neurons differentiated between 'EC' and 'CC' trials in the current-trial peri-stimulus time window (blue; [-500 500]ms relative to stimulus onset) significantly better than those in the preceding-trial post-feedback period in differentiating between error and correct trials. Error bars, ± s.e.m. across neurons. Broken horizontal lines, the 97.5th percentile of the null distribution. '*', '**', and '***' mark statistical comparisons with p value <0.05, ≤0.01, or ≤0.001, respectively. 'n.s' marks not significant (p>0.05). BP=button press.

The majority of errors (67%) occurred on incongruent trials. Spike rates of error neurons on the error trials (within the post-action epoch; Fig. S2a) did not correlate with RT (Fig. S3a-b; for Type I error neurons, $p > 0.4$, $t(98) = 0.86$ in dACC, $p > 0.5$, $t(117) = -0.41$ in pre-SMA; for Type II error neurons, $p > 0.5$, $t(34) = -0.54$ in dACC and $p > 0.5$, $t(79) = -0.63$ in pre-SMA; t-test) and did not distinguish significantly between congruent and incongruent errors (Fig. S4a,c, see Fig. S4b,d for statistics). We thus pooled congruent and incongruent error trials in all subsequent analyses. Unlike the responses of error neurons, RTs were significantly longer on incongruent compared to congruent error trials (Fig. S4g; $p < 0.001$, $t(57) = 4.03$, paired t-test), arguing that errors were not due to lapses in stimulus processing.

While the neuronal error signal persisted into the post-feedback epoch (which appeared 1 sec after button press; Fig. 2.1a and Fig. S2a), the maximal spike rate modulation for both types of error neurons occurred before onset of feedback (Fig. 2.3b,c). An out-of-sample analysis of effect sizes (see Methods) confirmed this impression: spike rates of error neurons in the epoch between action and feedback onset carried significantly more information about the occurrence of an error than those in the post-feedback epoch (Fig. 2.3g; $p < 10^{-10}$, $t(199) = 98.3$ in dACC, $p < 10^{-10}$, $t(199) = 288.2$ in pre-SMA, paired t test). Thus, feedback onset did not reactivate error or terminate their ongoing response on error trials (Fig. 2.3b,c). In summary, error neurons were action-triggered and encoded the detection of a mismatch between the intended action and the actual action performed.

2.3.3 Error-integrating neurons

We hypothesized that MFC neurons signal information about the history of self-monitored outcomes (Shima and Tanji, 1998, Kennerley et al., 2006). We identified a significant proportion of MFC neurons (see Fig. 2.3a; $n = 46$, 11.5% in the dACC; $n = 58$, 13.5% in pre-SMA, $p < 0.001$ for both areas, permutation test; also see Table S2) whose spike rates signaled whether the response in the preceding trial was an error or not (Fig. 2.2e-f, Fig. 2.3a, Fig. S2c-d). Response patterns of these ‘error-integrating’ neurons differed between dACC and pre-SMA: whereas dACC neurons (Fig. 2.3d) showed a peri-stimulus onset spike rate increase on trials that followed an error, responses in pre-SMA were characterized by an extended decrease starting in the pre-stimulus baseline period (Fig. 2.3e).

We next tested whether this response pattern was the result of error signals persisting from the preceding error trial, in which case the error-integrating neurons would also be classified as error neurons. While there was some overlap between the two categories (overlap: n= 12 and 20 for dACC and pre-SMA), many error-integrating neurons were not also error neurons (non-overlap: n= 34 and n=38 for dACC and pre-SMA, respectively). The time course of the population activity of all error-integrating neurons confirmed this: while these neurons did signal errors to some degree during the post-action epoch (definition see Fig. S2) on the preceding trial (Fig. 2.3h, orange; mean AUC for dACC 0.59±0.01, for pre-SMA 0.63±0.01; $p < 0.05$ versus chance for both areas, permutation tests), this error signal was attenuated after feedback (Fig. 2.3h, green; mean AUC for dACC 0.59±0.01, for pre-SMA 0.57±0.01), reinforced before stimulus onset, then continued on to after the stimulus onset on the next correct trial (Fig. 2.3h, blue; mean AUC for dACC 0.65±0.01, for pre-SMA 0.62±0.01; blue vs green, $p < 0.001$, $z = 4.74$ in dACC and $p < 0.001$, $z = 4.72$ in pre-SMA, rank sum test). In summary, we found error-integrating neurons carried a sustained error signal that was reinforced *around* stimulus onset on the subsequent trial, consistent with a putative role in post-error behavioral control.

2.3.4 Relationship between error and conflict neurons, and a signature of control

Response conflict is thought to be the stimulus-evoked competition between a prepotent but task-irrelevant response (reading the word) and a task-relevant response (the ink color) (Botvinick et al., 2001, Shenhav et al., 2013). In this framework, error signals are generated by conflict between the committed erroneous response and continuing development of the correct response. This implies that error neurons should not only signal errors, but also signal conflict as soon as it arises following stimulus onset. Here, we tested this hypothesis. We found that, as a group, the spike rates of error neurons within the post-stimulus epoch ([0 500]ms relative to stimulus onset; Fig. S2a) did not distinguish significantly between incongruent and congruent stimuli (Fig. S3c-d; see legend for statistics). For the second analysis, we first identified conflict neurons in both dACC (Fig. S3e; $p = 0.03$, $n = 41$, 6.7% of recorded neurons for Type I and $p < 0.001$, $n = 43$, 7% of recorded neurons for Type II; permutation tests) and pre-SMA ($p < 0.001$, $n = 54$, 10%, Type I only; permutation test), confirming earlier work (Sheth et al., 2012, Ebitz and Platt, 2015). These neurons changed their spike rates to signal conflict, with the signal culminating ~500ms after stimulus onset (Fig. S3f). The majority of error neurons were not conflict neurons (81% of error neurons in dACC and 87% of error neurons in pre-SMA were not conflict neurons) and vice-versa (Table S3). The number of neurons qualified as both error and conflict neurons was not significantly greater than what was expected if these two categories were independent (Fisher's exact test for association, see Table S3). Also, error neurons are significantly more common in MFC relative to conflict neurons (28% vs 12%, $p< 0.001$, $\chi^2(1) = 93.64$, Chi-squared test). Thus, the substrates for error monitoring and conflict detection are largely separate at the neuronal level.

According to the model mentioned above, on an incongruent and correct trial, conflict arises accompanying stimulus onset and recruits cognitive control, which in turns resolves the conflict and results in a correct response. Neural activity reflecting conflict detection and the state of cognitive control are thus intermingled. To separate them, we compared spike rates within the post-stimulus epoch between error incongruent and correct incongruent trials for the previously identified groups of neurons. We found that, at the group level, only Type II error neurons in dACC (Fig. S3g) as well as conflict neurons in both dACC (Fig. S3j,k) and pre-SMA (Fig. S3l) carry a signature of control state according to this metric (See legend for statistics). We also confirmed these results by a multi-level Poisson regression model where the RT effect is controlled, with qualitatively similar results (data not shown). None of the other types of neurons changed their spike rates significantly to reflect the control state ($p = 0.41$, $z = 0.82$ for Type I error neuron in dACC; $p = 0.87$, z value $= 0.16$ for Type I error neuron and $p = 0.26$, $z = -1.12$ for Type II error neuron in pre-SMA; $p = 0.17$, $z = -1.37$ for error-integrating neurons in dACC, $p = 0.24$, $z = -1.16$ for error integrating neurons in pre-SMA; signed rank test). Notably, the Type I error neurons and error-integrating neurons in both dACC and pre-SMA did not carry this signature of control state, consistent with a more specialized role in monitoring and control, respectively.

2.3.5 Waveforms of error neurons and error-integrating neurons

We quantified the duration of the extracellular waveforms of neurons (‘trough-to-peak time’) to differentiate between putative cell types (Bartho et al., 2004, Mitchell et al., 2007, Rutishauser et al., 2015). The distribution of spike duration is significantly bimodal in both dACC and pre-SMA (Fig. S5a,e; $p < 0.001$ for both areas, Hartigan’s dip test). 80% of neurons had broad waveforms (trough-to-peak time greater than 0.5ms), a feature indicative of putative pyramidal cells (Mitchell et al., 2007). Comparing the proportion of putative pyramidal and inhibitory neurons within each category with the overall population revealed that most error and error-integrating neurons are putatively excitatory (Fig. S5 legend for statistics).

2.3.6 Error neurons signal errors earlier in pre-SMA than in dACC

We next sought the point in time when error information first became available in each brain region. We first estimated the differential onset latency (the first point in time when the spike rates significantly differentiated between two conditions, see Methods), which showed that the error signal in pre-SMA occurred significantly earlier than in dACC by 55ms (Fig. 2.4a,b; median dACC latency, 165ms; median pre-SMA latency, 110ms; $p = 0.002$ and $z = 3.05$, rank sum test). A putative downstream readout (here a decoder), however, only has access to the response of an error neuron on a single trial. We used a Poisson-based method to detect, for each trial, the point of time the spike rate of a given error neuron departed significantly from the baseline (Type I only; see Methods for details). This analysis revealed that the error signal appeared first in pre-SMA 52ms after button press (Fig. 2.4c; $p = 0.0025$, $z = 3.02$, rank sum test), followed by the response in the dACC 60ms later (median difference). Repeating this analysis restricting to simultaneously recorded error neurons revealed quantitatively similar results ($p = 0.002$ and $z = 2.89$; one-tailed rank sum tests).

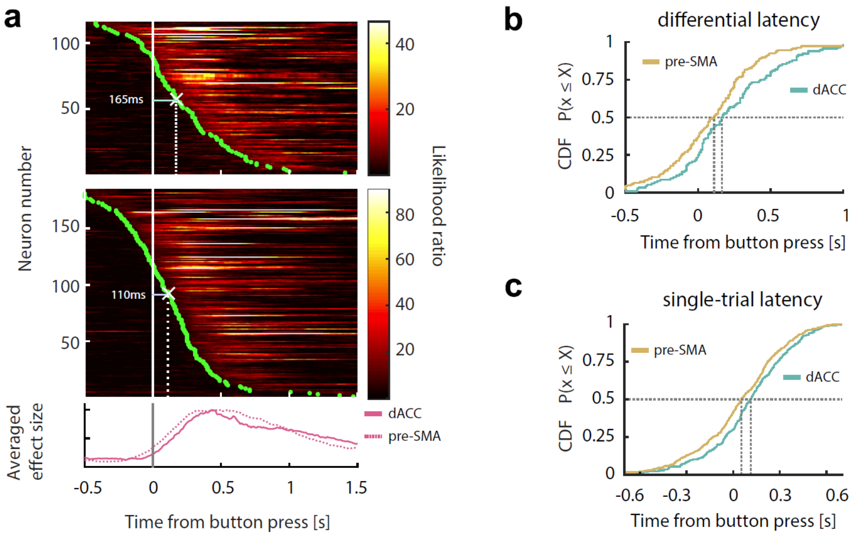

Figure 2.4 Error neurons in pre-SMA respond earlier than error neurons in dACC

(a) Temporal profile of error information carried by the error neuronal population (Type I and II pooled), aligned at button press (gray bar) and sorted by the onset latencies of error information (green dots). Each row represents one error neuron in dACC (upper) or pre-SMA (middle). White crosses mark the medians of onset latencies. Bottom plot shows the average likelihood ratio normalized by the peak value (solid line, dACC; broken lines, pre-SMA).
(b) CDF of differential latencies (see Methods for details) are shown for error neurons.
(c) CDF of single-trial onset latencies for error neurons.
CDF=cumulative distribution function.

2.3.7 Error-related negativity

Simultaneously with single neurons, we recorded the intracranial EEG (iEEG) using low-impedance macro contacts in both dACC and pre-SMA (see Table S1 and Fig. S1a). Following an erroneous button press, the iEEG revealed a prominent intracranial error-related negativity (iERN) visible on single trials in both dACC and pre-SMA (Fig. 2.5a-c, Fig. S6a-b). We also repeated the same task with scalp EEG in control subjects (see Methods) and found that the scalp ERN (Fig. S6c,d) had waveforms similar to the iERN, but with 5-10 times samller amplitude (Compare Fig. 5c and Fig. S6c). The extracted iERN amplitude values significantly distinguished error from correct trials (see Methods for details; Fig. 2.5d; median AUC for dACC electrodes is 0.59, $p<10^{-10}$, z=7.72; median AUC for pre-SMA electrodes is 0.67, $p<10^{-10}$, z=7.78; signed rank test).

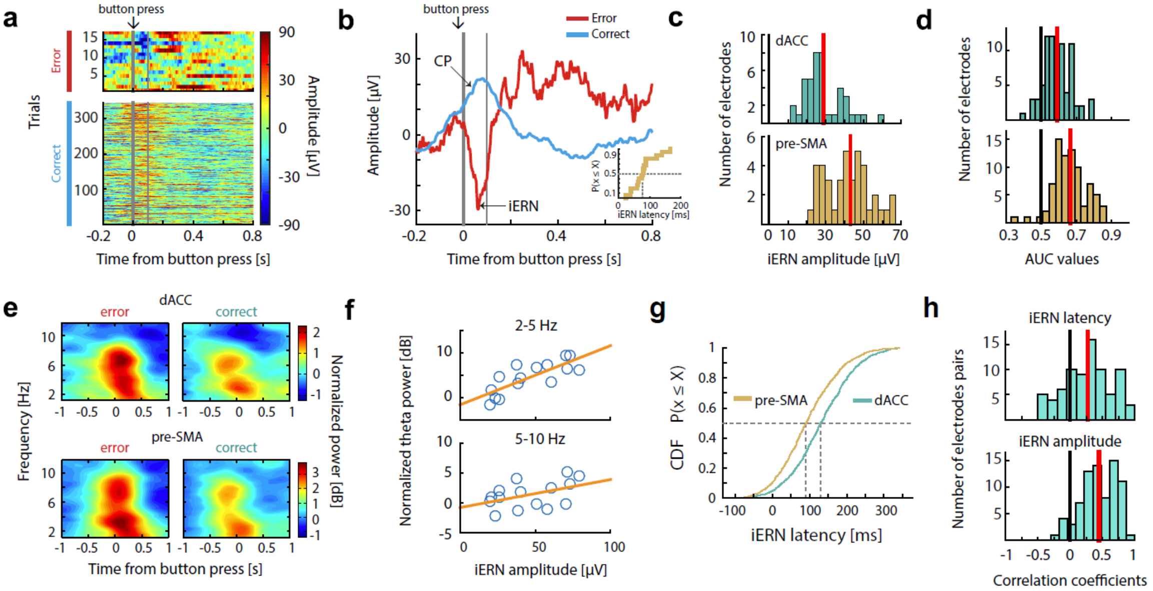

Figure 2.5 Intracranial error-related negativity (iERN)
(a) Example single-trial event-related potentials recorded from dACC, sorted by RT (RT increases from top to bottom rows) and trial types. t=0 is button press. Thin vertical bar marks 100ms after button press.
(b) Average of data shown in (a) grouped by trial types (colors; red for error, blue for correct), aligned at button press (t = 0, thick vertical gray bar). Inset, distribution of iERN latencies for the same data. Thin vertical bar marks 100ms after button press.
(c) Mean iERN amplitudes over all electrodes placed in dACC (green) and pre-SMA (brown). Red vertical bars show the median values.

(d) iERN amplitudes differ significantly between correct and error trials, evaluated using ROC analysis (see main text for details). Red vertical bars show the mean values.
(e) Spectral signature of the error signal. Power spectrum is aligned at button press (t = 0; averaged across n = 42 sessions). The region of power increase visibly splits into two frequency bands (2-5Hz and 5-10Hz). See Fig. S6e-f for statistics.
(f) Trial-by-trial correlation between iERN amplitude and slow-theta (2-5Hz; top) and (5-10Hz; bottom) power for the example session shown in (a,b).
(g) Comparison of iERN latency across all sessions. The iERN peak occurred significantly earlier in the pre-SMA compared to the dACC.
(h) Trial-by-trial correlation of iERN latency (upper) and iERN amplitude (lower) between pairs of iERNs recorded simultaneously in dACC and pre-SMA. For both, the correlation coefficients have a mean significantly greater than zero. Red vertical bars show the mean values.

Time-frequency analyses revealed that iEEG power increased following button press in two frequency bands: 2-5Hz (‘slow theta’) and 5-10Hz (‘theta’) on both error and correct trials (Fig. 2.5e), with a significantly stronger increase on error trials (Fig. S6e-f; see legend for statistics). Previous studies have demonstrated that volume conduction from the hippocampus can account for theta oscillations in neocortex (Sirota et al., 2008, Gerbrandt et al., 1978). For this reason, we next repeated the same analysis for simultaneously recorded hippocampal iEEG. This revealed that although there were significant differences between error and correct trials, the differences were of opposite sign (Fig. S6e-f; see legend for statistics), suggesting that the signals we reported in MFC are not volume conducted from the hippocampus.

Power increase in both bands (averaged within [-0.5, +0.5]s around button press) was correlated with the iERN peak amplitude on the same trial (Fig. 2.5f shows this relationship for the data in Fig. 2.5a,b; Fig. S6g,h shows population summary; for theta-iERN correlation, mean correlation = 0.33, $p < 10^{-10}$, t(78) = 12.15 in dACC and mean correlation = 0.41, $p < 10^{-10}$, t(79) = 16.52 in pre-SMA; for slow theta-iERN correlation, mean correlation= 0.44, $p < 10^{-10}$, t(78) = 19.2; mean correlation = 0.48, $p < 10^{-10}$, t(79) = 19.4 in pre-SMA; mean-versus-zero comparisons, t-test). The ERN is thought to contain a combination of phase-locked theta-frequency band activity and non-phase-locked theta-frequency band power increases (Yeung et al., 2007, Trujillo and Allen, 2007, Wang et al., 2005, Luu et al., 2004). Induced theta power (Fig. S6i) alone in the same time-frequency region-of-interest was also significantly correlated with iERN amplitude (Fig. S6j-k; see legend for statistics).

Consistent with the spiking activity of error neurons reported above, the iERN amplitude, theta and slow theta power also did not differ significantly between congruent and incongruent errors (Fig. S4e,f; see legend for statistics). Although the iERN in dACC and pre-SMA had similar waveforms, their peak latency differed: the iERN occurred on average 40ms earlier in pre-SMA than in dACC (Fig. 2.5g; For a comparison with spike latency, see Fig. S6n; median dACC latency is 140ms, median pre-SMA latency is 100ms; $p <10^{-10}$, z = 13.04, rank sum test; this effect held even after equalizing amplitudes across areas, $p <10^{-10}$, z = 10.5, rank sum test). We also investigated the difference as well as correlation in latency and amplitude between pairs of simultaneously recorded iERNs. The distribution of these latency difference values between the iERN pairs have a significantly non-zero median (Fig. S6l; median = 18ms; $p < 0.001$, z =19.27, rank sum test), further confirming the leading role of pre-SMA. This latency difference also provides evidence against the hypothesis that the iERN is volume conducted because this would result in simultaneous onset (Logothetis et al., 2007). Similarly, the amplitude difference between iERN pairs was significantly positive (Fig. S6m; median = 11 μV; $p < 0.001$, $z = 20.14$, rank sum test). In addition, both the latency and amplitudes of pairs are significantly correlated (Fig. 2.5h; mean correlation coefficient for latency correlation is 0.27 and for amplitude correlation is 0.44; $p < 0.001$, $t (77) = 6.81$ for latency correlation and $p < 0.001$, $t (77) = 0.29$ for amplitude correlation, t test). Together, this data shows that the iERN is accompanied by theta and slow theta activity in MFC, and that the iERNs appeared earlier and with larger amplitude in pre-SMA.

2.3.8 Linking spikes, iERN, and behavior

To gain insights into the processes that contribute to the iERN, we began by correlating its amplitude with the spike rates of error neurons. We used a multi-level linear model in which iERN amplitude was the dependent variable, and RT and spike rates were fixed effects. We then tested whether this model explained the data significantly better than a null model (see Methods). Here, the null model has the iERN amplitude as the dependent variable, and only RT as the fixed effect (and all the random effects remained the same as before). Note that only error trials were included in this analysis. The spike rates of Type I error neurons significantly co-varied with the iERN amplitude recorded in the same brain region in a trial-by-trial fashion (Fig. 2.6a, $p = 0.01$ for dACC error neurons, $p < 0.001$ for pre-SMA error neurons; cluster-based permutation test for the time course, details see Methods). This effect was evident at the single-cell level: each error neuron's mean spike rate was greatest on trials with the largest iERN amplitude (Fig. 2.6b). This correlation began around action onset (button press), peaked ~400ms after erroneous actions with a maximal likelihood ratio of 7.9 for dACC and 15.4 for pre-SMA, and occurred earlier in pre-SMA compared to dACC (Fig. 2.6a). This is consistent with the shorter iERN latencies in pre-SMA reported above (Fig. 2.5g). This effect held when we used spike counts within the post-action epoch ([0 1]s after button press; Fig. S7a; $p = 0.008$, $\chi^2(1) = 6.56$ in dACC and $p = 0.012$, $\chi^2(1) = 5.81$ in pre-SMA). We found no significant correlation between iERN amplitude and spike rates of Type II error neurons (Fig. S7b; spike counts within [0 1]s after button press were used in the GLM; $p = 0.19$, $\chi^2(1) = 1.64$ in dACC, $p = 0.07$, $\chi^2(1) = 3.36$ in pre-SMA, likelihood ratio test) or non-error neurons ($p > 0.1$, cluster-based permutation test).

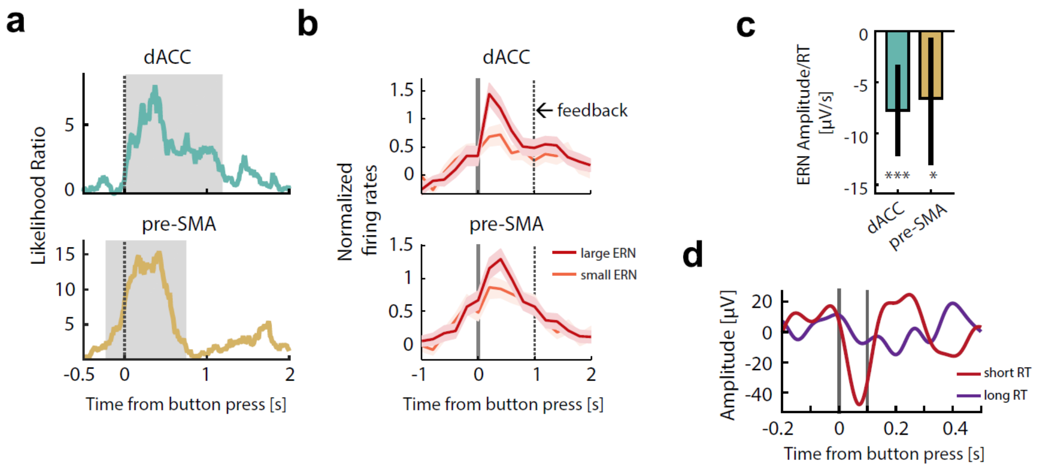

Figure 2.6 The iERN amplitude is correlated with error neuron spike rate and RT
(a) iERN amplitude correlated significantly with the spike rates of error neurons (Type I). The likelihood ratio peaked around ~400ms after button press. t=0 is button press. Grey shading delineates the extent of the significant cluster as determined by a cluster-based permutation test. Note that the significant cluster started earlier in pre-SMA.
(b) Illustration of the relationship between iERN amplitude and spike rates of the error neurons (Type I). Color code: red for error trials with largest ERN (iERN larger than the 80th percentile), orange for error trials with smallest ERN (iERN smaller than the 20th percentile). t=0 marks button press. Solid bar marks button press; dotted bar marks feedback onset.
(c) iERN amplitude correlated significantly with RT. Bar plots represent values of regression coefficient for the fixed effect of RT in a mixed effect model. Error bars represent 95% confidence intervals (see Methods).
(d) Illustration of the relationship between RT and iERN amplitude (data from one session). iERN amplitudes were larger when the corresponding RTs were short (red; RTs shorter than the median) than when RTs were long (purple; RTs longer than the median). Thick vertical bar marks button press; thin vertical bar marks 100ms after button press. See panel c for statistics.
'*', '**', and '***' mark statistical comparisons with p value ≤ 0.05, ≤ 0.01, or ≤ 0.001, respectively.

Does the same relationship hold on correct trials? To answer this question, we first extracted the positive peaks on the correct trials as informed by the average event-related potential (ERP) shape (Fig. 2.5b, see Methods). We then constructed a similar multi-level model but with evoked potentials on the correct trial ('CP') as the response variable, and spike rates of error neurons and RT on the same trial as fixed effects. We found no significant correlation between the evoked potential amplitude and spike rates of error neurons on correct trials (Fig. S7c; p = 0.34. $\chi^2(1) = 0.92$ for Type I error neurons and p = 0.74, $\chi^2(1) = 0.11$ for Type II error neurons in dACC; p = 0.88, $\chi^2(1) = 0.023$ for Type I error neurons and p = 74, $\chi^2(1) = 0.11$ for Type II error neurons in pre-SMA). The relationship between spiking activity and amplitude of evoked potential is thus specific to error neurons.

Each trial was characterized not only by whether an error occurred (indexed by error neurons) but also by its RT, which likely index the degree of cognitive control recruited as well as prediction of outcomes. Notably, RT and error neuron spike rates are internal variables indicative of different processes, as they were uncorrelated (Fig. S3a,b). We thus next investigated whether iERN amplitude might be correlated with RT using the same multi-level linear model approach (Fig. 2.6c). We found that larger iERN amplitudes were associated with shorter RTs in both dACC and pre-SMA (Fig. 2.6d shows this effect of RT on the iERN amplitude; Fig. 2.6c provides statistics; The significance of this RT effect was evaluated by a likelihood ratio test: For dACC, $\chi^2(1) = 14.61$, p = 0.0001; For pre-SMA, $\chi^2(1) = 5.325$, p = 0.021). This negative correlation was significant after controlling for stimulus congruence, which by itself would have resulted in RT differences (See Fig. S4g for RT comparisons for error trials; for dACC, $\chi^2(1) = 9.54$, p = 0.002; for pre-SMA, $\chi^2(1) = 4.83$, p = 0.028). Thus, the faster an error was made, the larger the iERN amplitude was on that trial. Together, these data revealed two distinct components of the iERN: one that is positively correlated with error neuron spike rate (action outcome information) and one that is negatively correlated with RT, putatively action-outcome prediction error (Alexander and Brown, 2011).

2.3.9 Neural signatures of PES in dACC

We next sought to determine which aspects of the performance monitoring circuitry interface with the control processes that result in PES. Note that previous efforts to correlate the magnitude of error monitoring signals measured using scalp EEG to PES have yielded contradictory results (Gehring and Fencsik, 2001, Debener et al., 2005, Nieuwenhuis et al., 2001, Hajcak et al., 2003). The evoked potential likely reflects synaptic inputs to a brain region. If so, this synaptic input would then subsequently cause the local responses we measure as spiking activity of neurons in the same region. Given this, we investigate the hypothesis that the ERN itself does not predict PES, but that the ensuing relationship between the ERN and the activity of error neurons does.

We first tested whether the amplitude of the iERN is indicative of PES. Error trials were separated into two groups (for each session): one that leads to PES larger than the median value, and the other that leads to PES smaller than the median value. We then assessed whether the iERN amplitude differed between these two groups (quantified by the 'large-small PES' index, zero equals no difference, see Methods). Consistent with some previous EEG studies (Gehring and Fencsik, 2001, Nieuwenhuis et al., 2001, Hajcak et al., 2003), we did not find a significant relationship between iERN amplitude and PES (Fig. 2.7a).

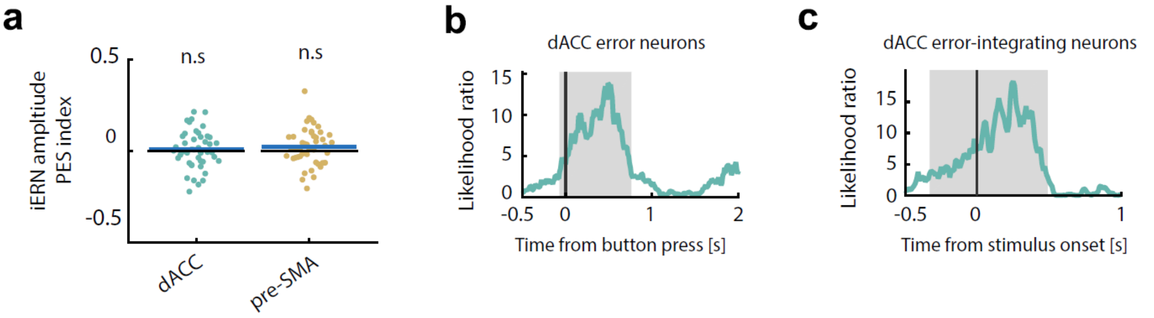

Figure 2.7 Error neuron-iERN synchrony during errors predicts engagement of control
(a) iERN amplitude did not predict PES significantly. Mean values of the PES index (see Methods) for iERN amplitudes were not significantly different from zero. Blue bars denote mean values; black bars denote zero.
(b) The correlation between iERN amplitude and error neuron spike rates (as a function of time; quantified as the likelihood ratio in model comparison; see Methods) predicted the extent of post-error slowing (PES) in the dACC. t=0 is button press. Grey shading delineates the extent of the significant cluster as determined by a cluster-based permutation test ($p < 0.05$). The same analysis in the pre-SMA did not yield a statistically significant relationship.
(c) The spike rates of error-integrating neurons in dACC around the time of stimulus onset predicted PES. '*', '**', and '***' mark statistical comparisons with p value ≤ 0.05, ≤ 0.01, or ≤ 0.001, respectively. Error bars represent ± s.e.m across cells.

We next investigated whether neural synchrony would predict PES. Here we assessed neural synchrony by the extent to which spike rates of an error neuron co-vary with the amplitude of the iERN (Nir et al., 2007). This correlation measure could also indicate the efficacy of iERN inputs in driving the local neuronal error signal that is important for control recruitment. We used a multilevel model (see Methods) to assess whether there was a significant interaction between spike rate of error neurons and the large-small PES categorical variable in predicting iERN amplitude trial-by-trial. This revealed that in dACC, the stronger the iERN- spike rate correlation around the time an error was committed, the larger was the subsequent PES (Fig. 2.7b; the maximal likelihood ratio is 13.9; $p = 0.015$ obtained by cluster-based permutation test. See Methods for details; the same analysis with Type II error neurons in dACC and both types of error neurons in pre-SMA did not yield a statistically significant relationship, see Fig. S7d,e). Note that while the strength of the correlation between the iERN and error neuron firing rate (in dACC) was thus predictive of PES, both underlying variables themselves were not ('large-small PES' index. $p > 0.5$, $z = 0.46$ for iERN and $p > 0.5$, $z = -0.17$ for spike rate within [0 1]s post button-press, signed rank test; See also Fig. 2.7a).

Error-integrating neurons in dACC signaled whether an error was committed in the previous trial by increasing their spike rates around stimulus onset. This pattern suggests that these neurons could be involved in implementing PES. To investigate this, we tested the relationship between spike rates of error-integrating neurons and PES (see Methods). Spike rates of dACC error-integrating neurons around the time of stimulus onset in post-error trials were significantly predictive of the size of PES (Fig. 2.7c; maximal likelihood ratio is 18.3; $p < 0.001$, cluster-based permutation test; as shown in Fig. 2.3d). This effect also holds if we used the spike counts within the peri-stimulus epoch ([-500 500]ms relative to the stimulus onset; Fig. S7f; $p < 0.001$, $\chi^2(1) = 15.76$, likelihood ratio test). We found no significant relationship between their spike rates and the levels of PES for pre-SMA error integrating neurons (Fig. S7f; $p = 0.07$, $\chi^2(1) = 3.31$, likelihood ratio test). We thus found two aspects of error monitoring that were predictive of the extent to which control was engaged (all in dACC only): iERN-error neuron spike rate coupling, and spike rates of error-integrating neurons. These two signals occurred at different points in time, suggesting that they are involved in bridging monitoring and corrective control.

2.4 Discussion

Here we provide direct recordings of single neurons in the human MFC that signal errors that are detected endogenously, before external feedback was presented and without the presence of an additional sensory signal to indicate task set (such as a stop signal). Error neurons were largely distinct from neurons signaling conflict shortly following stimulus onset, arguing that the representation of conflict detection and error monitoring in MFC are largely distinct. Conflict neurons were also modulated by the state of control: their activity differed between error incongruent and correct incongruent trials. This was not the case for Type I error neurons nor for error-integrating neurons, highlighting their putative roles in monitoring and actively mediating control, respectively. It remains an open question whether the error neurons that signal self-monitoring are functionally distinct from neurons that monitor external feedback, reward manipulations, or prediction errors that have been described in detail in macaques (Ito et al., 2003, Stuphorn et al., 2000, Scangos et al., 2013, Matsumoto et al., 2007, Matsumoto et al., 2003, Amiez et al., 2006, Ebitz and Platt, 2015, Hayden et al., 2011).

Despite evidence that the ERN (Gehring et al., 1993, Bonini et al., 2014, Brazdil et al., 2005, Godlove et al., 2011, Emeric et al., 2008, Falkenstein et al., 1991) originates from within dACC and/or pre-SMA (Dehaene et al., 1994, Debener et al., 2005), its relationship with neuronal spiking activity has not been clear. Our report shows that error neuron responses predict the amplitude of the iERN in both of these areas. Further, we showed that the iERNs recorded in pre-SMA 1) occurred earlier, 2) had larger amplitude, 3) were correlated in both amplitude and latency on a trial-by-trial basis with iERNs recorded simultaneously in dACC. These results are consistent with earlier studies (Bonini et al., 2014, Emeric et al., 2010). Our findings argue that both dACC and pre-SMA contribute to the ERN, but at different points in time.

This pattern of findings is consistent with two interpretations. One interpretation is that pre-SMA and dACC both receive inputs carrying error information in parallel, but pre-SMA receives the information earlier than dACC. This scenario is consistent with an influential computational account where synchronized disinhibition of dACC pyramidal cells by dopaminergic projections generates the iERN in dACC (Holroyd and Coles, 2002), and suggest that in pre-SMA similar disinhibition can also occur, but at earlier points of time. But a second possible interpretation is that pre-SMA provides error-related signals as an input to dACC, an interpretation which is consistent with a previous report where error-related evoked potentials in pre-SMA andSMA strictly precede those in the rostral cingulate zone (Bonini et al., 2014). Such a feedforward architecture could interpose additional relays as error signals are communicated indirectly from pre-SMA to dACC, for instance through the basal ganglia (Nachev et al., 2008, Jahanshahi et al., 2015). Future experiments utilizing causal manipulations will be necessary to probe the role of this putative feedforward connection in error processing.

Strong coupling between LFP components (here measured by the iERN) and spike rates is well documented in sensory cortices, where the coupling is often driven in part by common sensory inputs [but see (Kayser et al., 2004)]. However, in brain areas removed from direct sensory inputs, such as the hippocampus and inferior temporal cortex, these two measures of neural activity diverge and encode information independently (Kreiman et al., 2006, Ekstrom et al., 2007, Ekstrom, 2010). The strong and transient ERN-spike rate coupling in MFC reported we found is thus notable, because it shows that such phenomenon can occur in brain areas whose primary functions are not sensory information processing. Evoked potentials such as the ERN are thought to reflect spatial summation of large numbers of postsynaptic potentials that synchronize to a substantial degree. Previous work has demonstrated that variation in LFP – spike rate coupling strength is commensurate with the level of synchronization between two neurons within a local population (Nir et al., 2007) and that the LFP can serve as an index of local information content carried by neurons (Kreiman et al., 2006). The correlation between iERN amplitude and spike rates of error neurons we find here is likely a reflection of the neuronal synchronization that underlies the detection and representation of self-generated errors, and/or more effective transmission of error information from other brain structures to the MFC. Notably, this relationship was specific to error trials: we found no significant correlation between similar deflections in the intracranial LFP during correct trials. It is thus likely the case that a separate group of neurons (which we did not describe here) receives the synaptic inputs that are synchronized during correct trials.

PES is one of the most studied consequences of error detection. PES is thought to be jointly produced by two types of cognitive control processes. One type is concerned with sensory information processing, reflected in the up- and down-regulation of task-relevant and task-irrelevant sensory areas (Danielmeier et al., 2011, King et al., 2010), as well as adjustments to the parameters of parietal sensory integration processes (Purcell and Kiani, 2016). The second type is concerned with engagement of response inhibition by error monitoring by MFC. BOLD activity within MFC is correlated with activity in task-related visual and motor areas, as well as the size of PES (Danielmeier et al., 2011, Kerns et al., 2004). Inactivation and lesioning of MFC abolishes PES (Narayanan et al., 2013, Kennerley et al., 2006), and individual differences in the white matter integrity of inhibitory networks that include pre-SMA (Aron and Poldrack, 2006, Aron et al., 2007, Jahanshahi et al., 2015) are correlated with the size of PES (Danielmeier et al., 2011). Although these studies unequivocally demonstrate the involvement of MFC in PES, they do not provide insight into how MFC neurons communicate error signals to the control processes that mediate PES. Here, we show that neuronal synchronization may provide a basis for recruiting control by MFC. We find that the strength of the correlation between iERN amplitude and the spike rates of error neurons is predictive of PES in dACC (but not pre-SMA). This suggests that the more synchronized the dACC error neurons are with neighboring neuronal population during errors, the larger the ensuing PES is. Given that neuronal synchronization can potentially represent information with high fidelity (Rutishauser et al., 2010, Wong et al., 2016) and thus have stronger impact on downstream targets (Siegel et al., 2012), our finding suggests that neuronal synchronization may underlie dACC-mediated PES.

Our results suggest that coordinated neural activity can serve as a substrate for information routing that enables the performance-monitoring system to communicate the need for behavioral control to other brain regions, including those that maintain flexible goal information, such as the lateral prefrontal cortex and the frontal polar cortex (Koechlin and Hyafil, 2007, Tsujimoto et al., 2010, Mansouri et al., 2017, Voytek et al., 2015). The present study offers new insights into the mechanisms of ERN generation and provides potential neural targets for validating the use of the ERN as an endophenotype for psychiatric illness (Olvet and Hajcak, 2008).

2.5 Supplementary materials

2.5.1 Supplementary figures

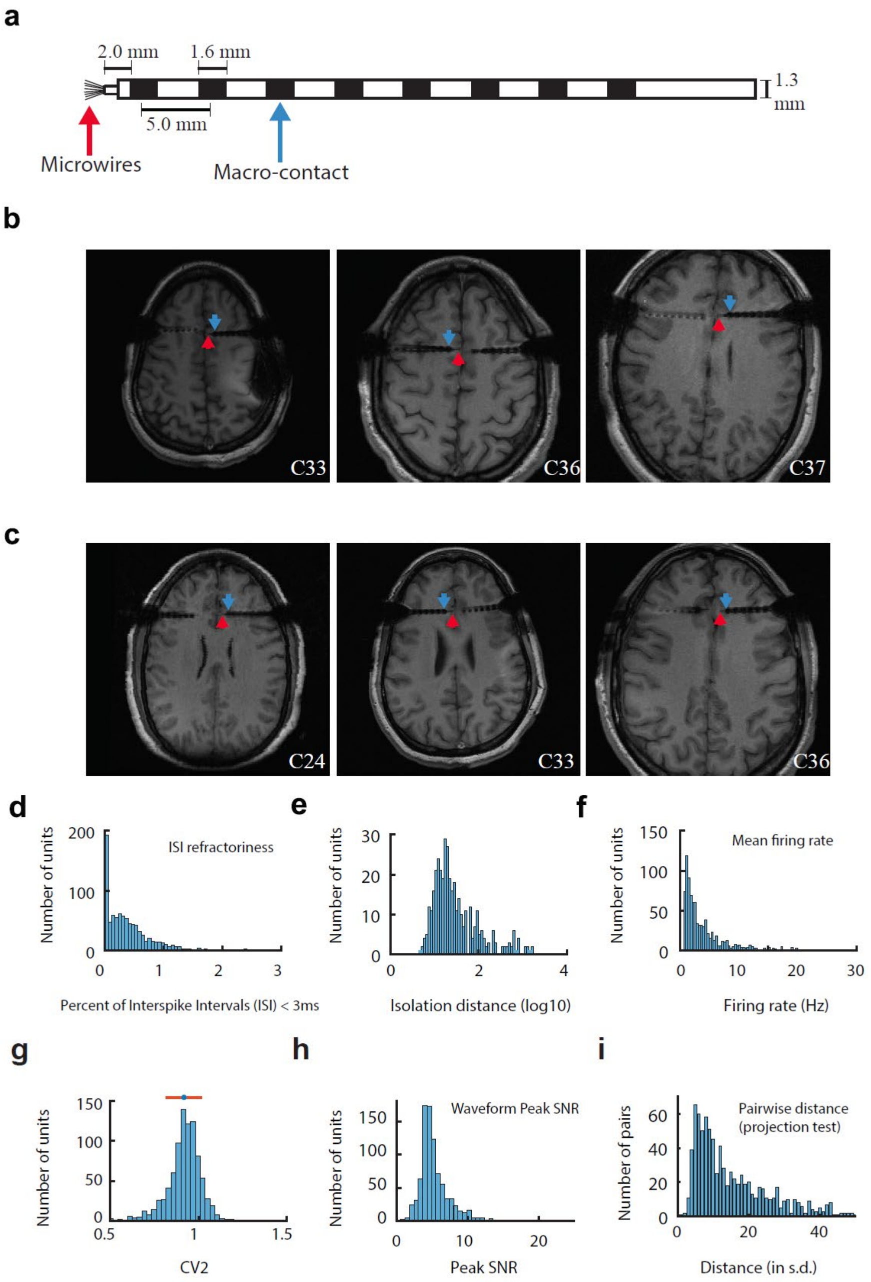

Figure S1 Recording electrode, example post-operative structural MRIs and spike sorting quality. Related to Figure 1-2.

(a) The hybrid macro-micro electrode used. Individual neurons are recorded using high-impedance microwires (red arrow; diameter 40µ, impedance 400-600kΩ). Field potentials are recorded from the low-impedance (<2kΩ) macro-contact most adjacent to the micro-wires (blue).

(b,c) Example axial T1 MRI scans of recording locations used in pre-SMA (b) and dACC (c). Red and blue arrowheads indicate tips of microwires and the macro contacts used, respectively.

(d-i) Spike sorting quality. Metrics quantifying the individual clusters that we used as putative single-units. (d) Histogram of proportion of inter-spike intervals (ISIs) shorter than 3ms. Most of our recorded clusters had less than 0.5% of their ISIs smaller than 3ms. (e) Isolation distance of all units for which this metric was defined (median 21.5). (f) Histogram of mean spike rates. (g) Histogram of coefficient-of-variation (CV2) values of all units. (h) Histogram of the signal to noise ratio (SNR) of the mean waveform peak computed for each unit. (i) Pairwise distance between all possible pairs, calculated using the projection test (see methods), of units on all wires with at least one cluster isolated. Distance is in unit of standard deviation after normalizing the data such that the distribution of waveforms around their mean is equal to one.

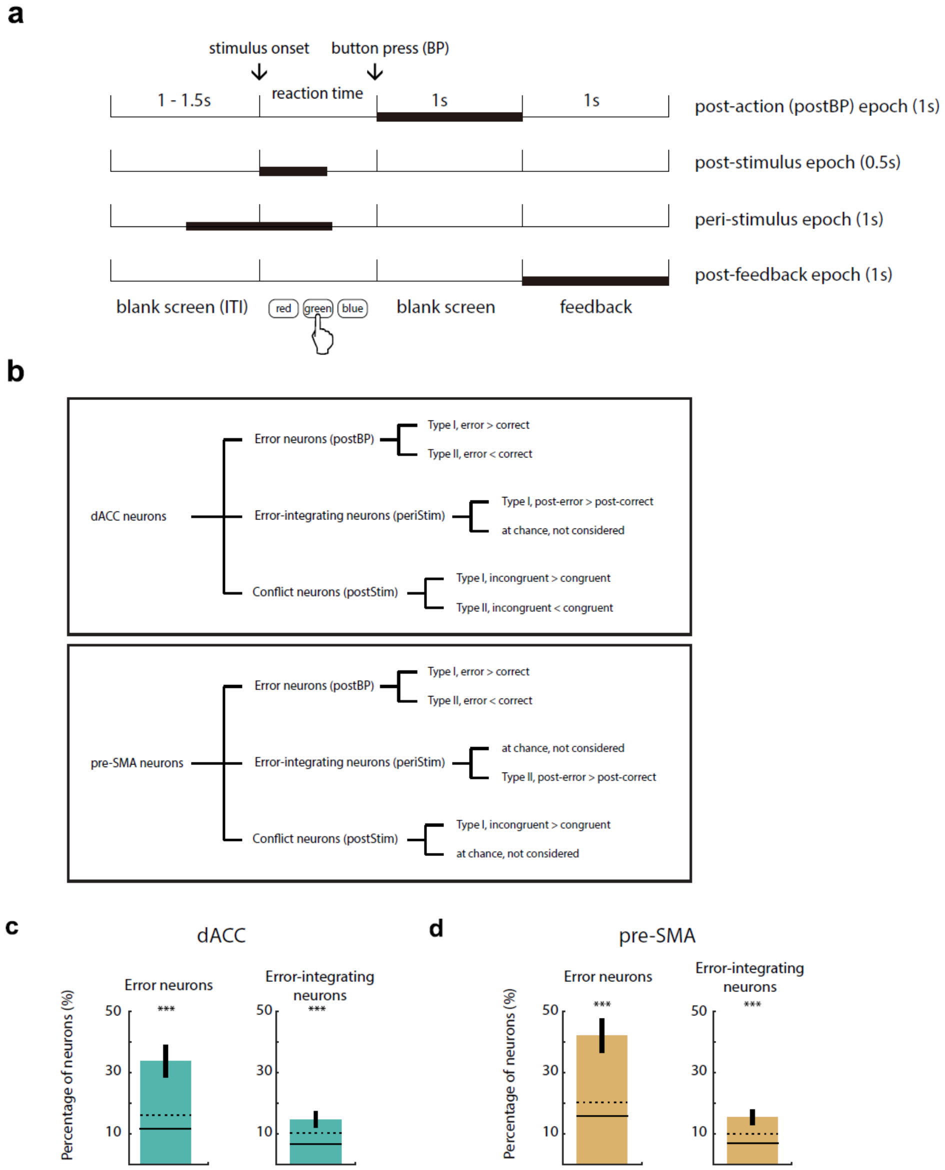
a
stimulus onset
button press (BP)
1 - 1.5s
reaction time
1s
1s
post-action (postBP) epoch (1s)
post-stimulus epoch (0.5s)
peri-stimulus epoch (1s)
post-feedback epoch (1s)
blank screen (ITI)
red
green
blue
blank screen
feedback
b
dACC neurons
Error neurons (postBP)
Type I, error > correct
Type II, error < correct
Error-integrating neurons (periStim)
Type I, post-error > post-correct
at chance, not considered
Conflict neurons (postStim)
Type I, incongruent > congruent
Type II, incongruent < congruent
pre-SMA neurons
Error neurons (postBP)
Type I, error > correct
Type II, error < correct
Error-integrating neurons (periStim)
at chance, not considered
Type II, post-error > post-correct
Conflict neurons (postStim)
Type I, incongruent > congruent
at chance, not considered
c
dACC
Error neurons
Error-integrating neurons
Percentage of neurons (%)
50
30
10
***
***
d
pre-SMA
Error neurons
Error-integrating neurons
Percentage of neurons (%)
50
30
10
***
***

Figure S2. Summary of epochs of interest, neuronal categories and distribution of neuronal categories separately for each session. Related to Figures 2-3.
(a) Epochs used to analyze spiking activity. Thick lines indicate the extent of the time windows. Length of each analysis epoch is indicated in brackets on the right.
(b) Summary of neuronal categories identified in dACC and pre-SMA. The second level lists all neuronal types and the time window (brackets) in which we identified more neurons than expected by chance. The

third level lists all sub-types (Type I or II) which were identified at levels higher than expected by chance. The contrasts listed refer to the spike rates during the trial types mentioned (e.g. ‘error > correct’ means spike rates in the error trials were larger than those in the correct trials for this particular type of neurons). (c-d) Percentage of significant neurons identified in dACC (c) and pre-SMA (d) that qualified as error or error-integrating neurons across recording sessions. Error bars represent ± s.e.m across sessions, solid and broken horizontal lines, the mean and the 97.5th percentile of the null distribution of the number of neurons expected by chance as estimated using permutation tests.
‘***’ marks groups of neurons which were observed more than expected by chance with p values ≤ 0.001.

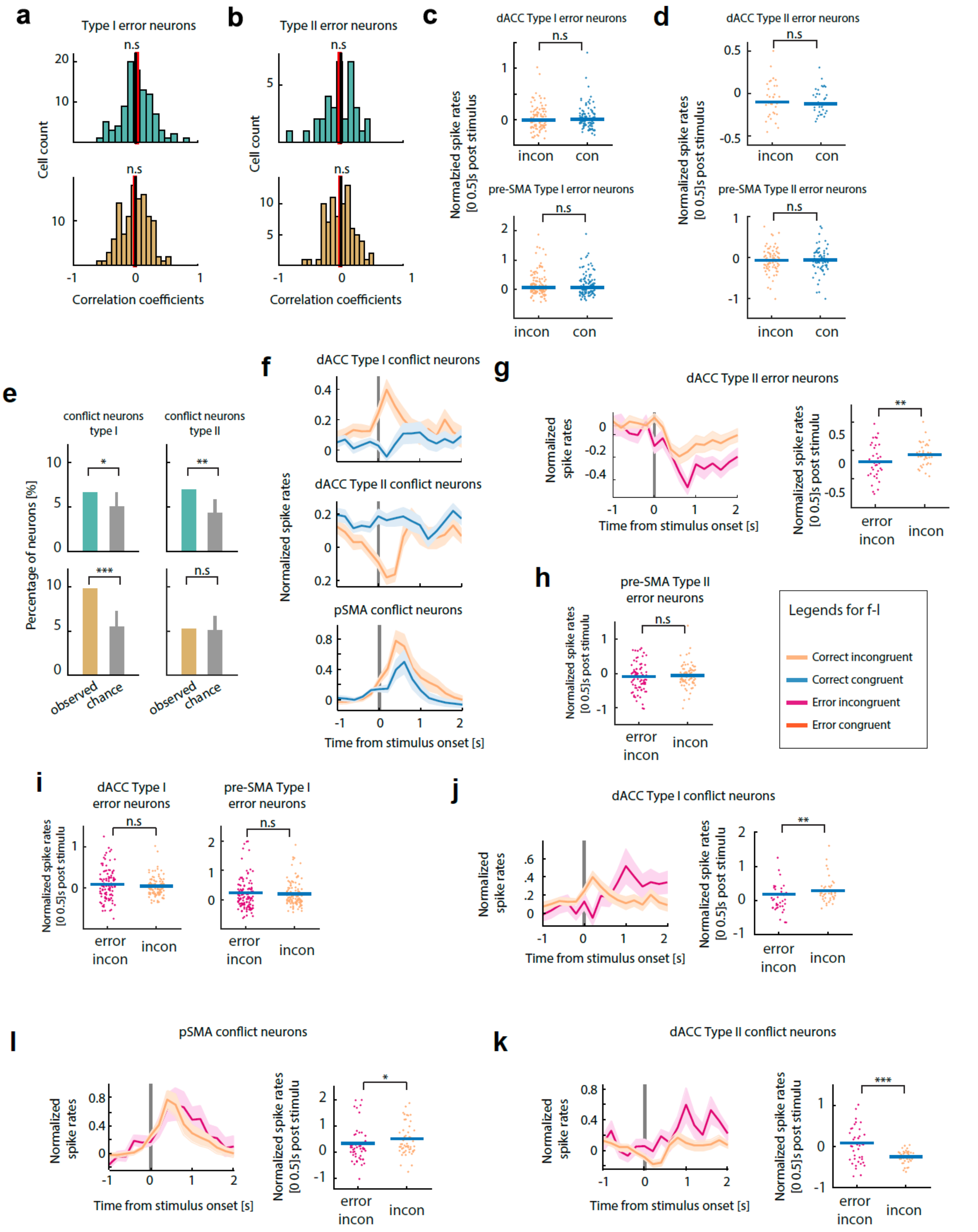

Figure S3. Signatures of conflict and control. Related to Figure 3.

(a) Correlation between RT and the spike rate of Type I error neurons identified in dACC (top) and pre-SMA (bottom) on the same error trials. There was no significant correlation (p > 0.04, t(98) = 0.86, t-test) in either area.

(b) Same as in (a) but for Type II error neurons. There was no significant correlation (p > 0.05, t(117) = -0.41, t-test) in either area.

(c) Spike rate of Type I error neurons did not differentiate significantly between correct congruent and incongruent trials in both dACC (upper; p = 0.92, z value = 0.1) and pre-SMA (lower; p = 0.18, z value = 1.33). Each data point shown is one neuron. Spike rates were quantified within a bin of 500ms size starting at stimulus onset and normalized by the baseline spike rates ([-700ms -200ms] relative to stimulus onset). Blue horizontal bars represent median values of the population.

(d) Same as in (c) but for Type II error neurons. Spike rate of Type II error neurons did not differentiate significantly between correct congruent and incongruent trials in both dACC (upper; p = 0.61, z value = 0.51) and pre-SMA (lower; p = 0.91, z value =0.12).

(e) Number of conflict neurons identified in dACC (green) and pre-SMA (brown). For the definition of Type I and II, see Methods. In dACC, both Type I and Type II conflict neurons have significantly greater number than that is expected by chance (p = 0.03 for Type I, p < 0.001 for Type II conflict neurons, permutation test). Gray bar shows the mean value of the empirical null distribution. Error bar shows the 95th percentile of the empirical null distribution.

(f) Average spike rates as a function of time for Type I conflict neurons (top), Type II conflict neurons (middle) in dACC and conflict neurons in pre-SMA (bottom). The spike rates were normalized by the baseline ([-700ms -200ms] relative to stimulus onset). Gray bar marks the onset of stimulus.

(g) Signature of control. Average spike rates as a function of time (left) and within the post-stimulus epoch ([0 500ms] relative to stimulus onset) of Type II error neurons in dACC for error incongruent vs. correct incongruent trials. The spike rates within the post-stimulus epoch differentiated error incongruent and correct incongruent trials significantly (p = 0.0062, z value = -2.74, Wilcoxon's signed rank test). Spike rates were normalized by the baseline ([-700ms -200ms] relative to stimulus onset). Gray bar marks the onset of stimulus. Blue bars on the scatter represents median of the population.

(h) Same as in (g) but for Type II error neurons in pre-SMA. The spike rates within the post-stimulus epoch did not differentiate error incongruent and correct incongruent trials significantly (p = 0.26, z value = -1.12).

(i) Same as in (g) but for Type I error neurons in dACC (left; p = 0.41, z value = 0.82, Wilcoxon's signed rank test) and pre-SMA (right; p = 0.26, z value = -1.12).

(j) Same as in (g) but for Type I conflict neurons in dACC. The spike rates within the post-stimulus epoch differentiated error incongruent and correct incongruent trials significantly (p = 0.006, z value = -2.75).

(k) Same as in (g) but for Type II conflict neurons in dACC. The spike rates within the post-stimulus epoch differentiated error incongruent and correct incongruent trials significantly (p < 0.001, z value = 3.54).

(l) Same as in (g) but for conflict neurons in pre-SMA. The spike rates within the post-stimulus epoch differentiated error incongruent and correct incongruent trials significantly (p = 0.034, z value = -2.12).

(c,f) Orange represent correct incongruent trials and blue represents correct congruent trials.

(g-l) Orange represents correct incongruent trials and magenta represents error incongruent trials.

'*', '**', '***' mark groups of neurons which were observed in proportions different then in the overall population with p values ≤ 0.05, ≤ 0.01 and ≤ 0.001 respectively (for a-b, t-test; for c,g-l, Wilcoxon's signed rank test). 'n.s' marks not significant (p > 0.05).

Figure S4

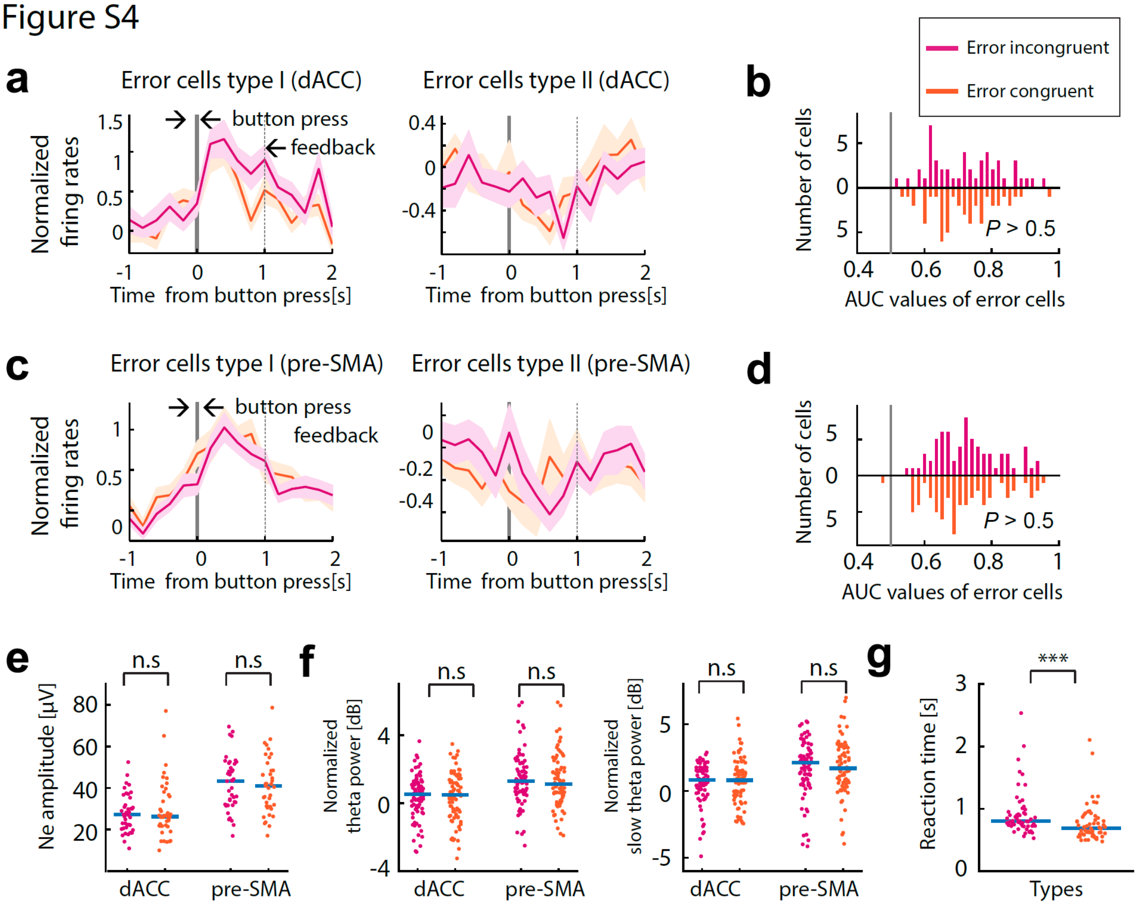

Figure S4. Response of error neurons and the iERN did not differ by error types (congruent error/incongruent error). Related to Figure 3 and 5.

(a) Average spike rates as a function of time for error neurons (Type I and II) in the dACC, normalized by the baseline. Right, Type I error neurons; Left, Type II error neurons. Response is aligned at button press (right). Trials are grouped by congruence (colors; magenta for error incongruent trials and orange for error congruent trials).

(b) Single-neuron ROC analysis of error neurons (both Type I and II) in dACC. The ability of spike rates in the post-button press time window ([0 1]s relative to button press) to differentiate error incongruent and error congruent trials each from correct trials did not differ significantly (AUC values computed from differentiating between correct/congruent error, between correct/incongruent error, 0.59±0.02 vs. 0.58 ±0.02 in dACC, 0.54±0.02 vs. 0.53±0.02 in pre-SMA; p > 0.5 for both areas, Wilcoxon rank sum test).

(c) Same as in (a), but for pre-SMA.

(d) Same as in (b), but for pre-SMA.

(e,f) iERN did not differ between incongruent and congruent errors (iERN amplitude comparisons: p = 0.8, z = 0.25 for dACC, p = 0.93, z =-0.09 for pre-SMA, signed rank test. Theta power comparisons: p = 0.72, z = -0.35 for dACC, p = 0.93, z = -0.09 for pre-SMA, signed rank test; Slow theta power comparisons; p = 0.49, z = -0.68 for dACC, p = 0.19, z = -1.3 for pre-SMA, signed rank test). Shown are the ERN amplitudes (e), theta and slow-theta power (f) for dACC and pre-SMA. Color code, same as in (a-d). Each dot shows one session, horizontal line shows the mean.

(g) The Stroop effect was significant on error trials. Average reaction times in the error incongruent trials were significantly longer than in the error congruent trials (p < 0.001, sign rank test). Errors thus did not occur due to an absence of conflict processing.

'***' mark statistical comparison with p value ≤ 0.001. 'n.s' marks not significant (p > 0.05).

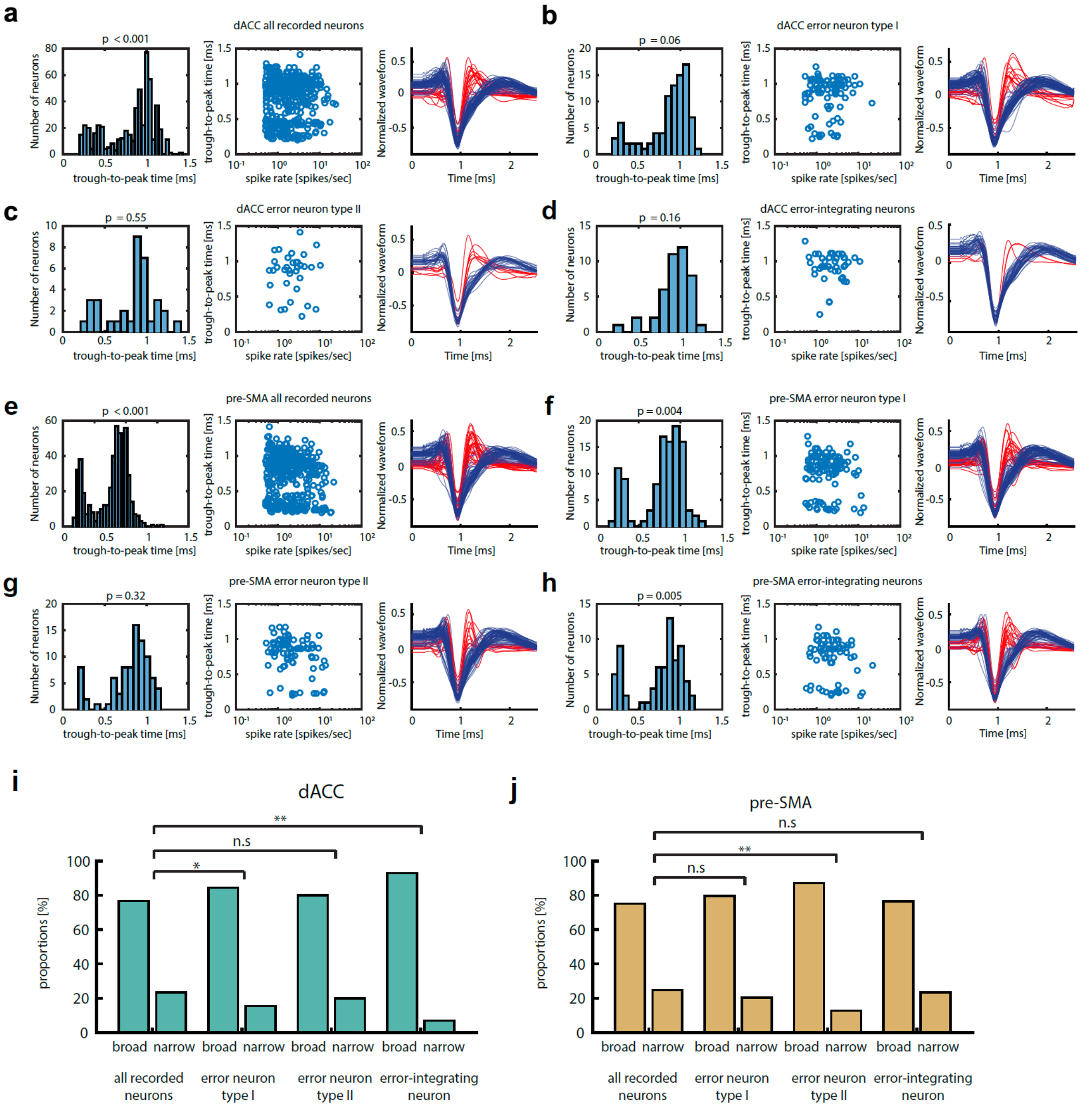

Figure S5. Waveform analyses of error and error-integrating neurons. Related to Figure 2.

(a-h) Distribution of trough-to-peak time (left), trough-to-peak time as a function of spike rates (middle) for each recorded neuron of a given group (described below). The rightmost plot shows the average spike waveforms of all neurons in the group, colored either blue or red depending on whether their trough-to-peak time was longer (blue) or shorter (red) than 0.5ms.

(a) All recorded neurons in dACC. The trough-to-peak distribution is significantly bimodal ($p < 0.001$).

(b) All Type I error neurons in dACC. The distribution of trough-to-peak time is not significantly different from unimodal ($p = 0.05$).

(c) All Type II error neurons in dACC. The distribution of trough-to-peak time is not significantly different from unimodal ($p = 0.55$).

(d) All error-integrating neurons in dACC. The distribution of trough-to-peak time is not significantly different from unimodal ($p = 0.16$).

(e) All recorded neurons in pre-SMA. The distribution of trough-to-peak time is significantly bimodal ($p < 0.001$).

(f) All Type I error neurons in pre-SMA. The distribution of trough-to-peak time is significantly bimodal ($p = 0.004$).

(g) All Type II error neurons in pre-SMA. The distribution of trough-to-peak time is significantly bimodal (p = 0.32).
(h) All error-integrating neurons in pre-SMA. The distribution of trough-to-peak time is significantly bimodal (p = 0.005).
(i) Proportions of putative pyramidal neurons (trough-to-peak time > 0.5ms) and interneurons (trough-to-peak time < 0.5ms) in dACC. Type I error neurons and error-integrating neurons have a significantly lower proportion of putative inhibitory neurons than the rest of the dACC population (15% and 7% vs 25% in the overall population, p = 0.05, odds ratio = 1.81 and p = 0.0074, odds ratio = 4.41, respectively; Fisher's exact test).
(j) Same as in (i) but for pre-SMA. Only the Type II error neurons had a significantly lower proportion of putative inhibitory neurons than the rest of the pre-SMA population (12% vs 26% in the overall population, p = 0.0034, odds ratio = 2.58, Fisher's exact test).

'*', '**' mark groups of neurons which were observed in proportions different then in the overall population with p values ≤ 0.05, ≤ 0.01, respectively (Hartigan's dip test). 'n.s' marks not significant (p > 0.05).

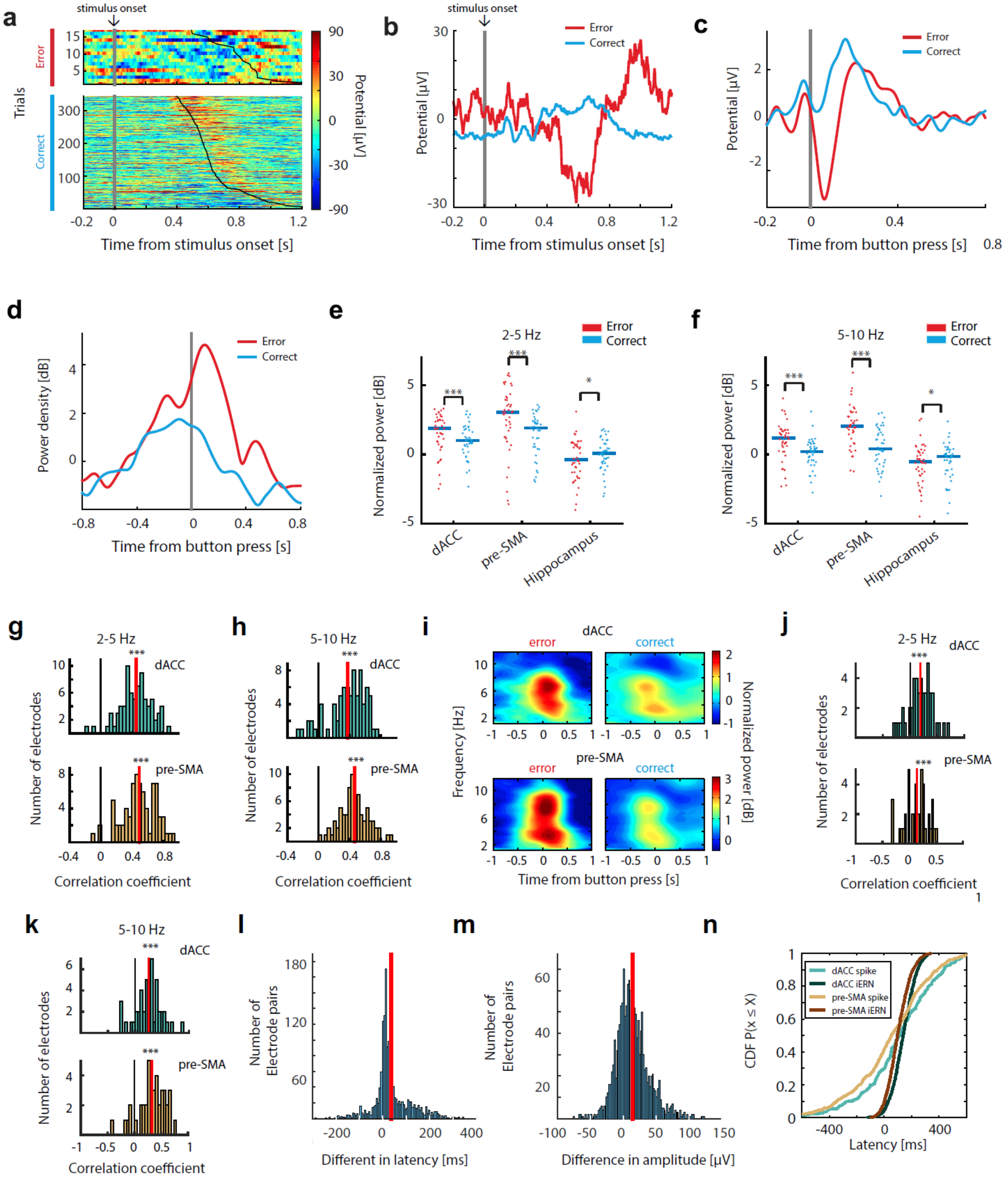

Figure S6. Stimulus-onset aligned intracranial event-related responses, statistics for time-frequency analysis of iERN, and scalp ERN control. Related to Figure 5.
(a) The same example single-trial event-related potential data as shown in Fig. 5a, but aligned at stimulus onset (t=0). The trials were sorted by reaction time (black lines; RT increases from top to bottom) and trial types (upper: error trials, lower: correct trials). Color bar represents ERP amplitude. Note the prominent ERP activities following button press (black line) as well as shortly after stimulus onset (blue). Gray bar represents stimulus onset.

(b) Average of data shown in (a) by trial types (colors; red for error, green for correct), aligned at stimulus onset (t = 0). Note that the sensory-evoked potential did not differ between trial types. Gray bar represents stimulus onset.
(c-d) Scalp-EEG recordings of non-surgical control subjects (N = 12) performing the same task reproduced the classical error-related negativity (ERN) and response-locked theta power (Compare with Fig. 5a-b).
(c) ERN (negative peak following button press, red) is significantly larger in amplitude in error compared to in correct trials (blue, t (11) = 4.53, p < 0.001). Gray bar represents button press.
(d) Theta power as a function of time. Error-related theta power (red) is significantly larger compared to in correct trials (green) after button press (t (11) = 6.47, p < 0.001). Gray bar represents button press.
(e) Power in the 2-5Hz band (0 to 500ms following button press) increased significantly more in error trials than in correct trials in both dACC ($p < 10^{-5}$, z value = 6.17) and pre-SMA ($p < 10^{-5}$, z value = 7.3). By contrast, hippocampal theta power also differed, but these differences were of opposite sign (p = 0.01, z = -2.55 for theta power, p = 0.01, z = -2.56 for slow theta power). Each dot shows one session, horizontal bar shows mean.
(f) Same as in (e) but for power in the 5-10Hz band.
(g) Mean Pearson's correlation coefficients between iERN amplitude and slow theta (2-5Hz) power are significantly larger than zero over all electrodes in dACC (mean correlation= 0.44, $p < 10^{-10}$, t(78) = 19.2) and pre-SMA (mean correlation = 0.48, $p < 10^{-10}$, t(79) = 19.4, t(79) = 11.52, t-test versus 0) Red vertical bars show population means.
(h) Same as in (g) but for correlations between iERN amplitude and theta (5-10Hz) power in dACC (mean correlation = 0.25, $p < 10^{-10}$, t(78) = 9.7) and pre-SMA (mean correlation = 0.33, $p < 10^{-10}$, t(79) = 11.52, t-test versus 0).
(i) Induced theta power, calculated after subtracting the ERP for each condition separately. Spectrograms shown are averaged across all sessions, see panel h for single-session statistics.
(j) Induced power was significantly correlated with the iERN amplitude in the slow theta (2-5Hz) band in both dACC (mean correlation = 0.21, $p < 10^{-10}$, t(78) = 8.41) and pre-SMA (mean correlation = 0.24, $p < 10^{-10}$, t(79) = 9.02, t-test versus 0) in pre-SMA.
(k) Induced power was significantly correlated with the iERN amplitude in the theta (5-10Hz) band in both dACC (mean correlation = 0.25, $p < 10^{-10}$, t(78) = 9.7) and pre-SMA (mean correlation = 0.33, $p < 10^{-10}$, t(79) = 11.52).
(l) Latency difference between pairs of iERNs recorded simultaneously in dACC and pre-SMA. The median latency difference of 18ms is significantly different from zero ($p < 10^{-5}$, z value = 19.27, Wilcoxon's signed rank test).
(m) Amplitude difference between pairs of iERNs recorded simultaneously in dACC and pre-SMA. The median latency difference of 11 μV is significantly different from zero ($p < 10^{-5}$, z value = 20.14, Wilcoxon's signed rank test).
(n) Comparisons of spike latencies and iERN latencies (replotting of data shown in main figure on different scale).
'*', '**', '***' mark statistical significance for $p \leq 0.05$, ≤ 0.01 and ≤ 0.001 respectively.

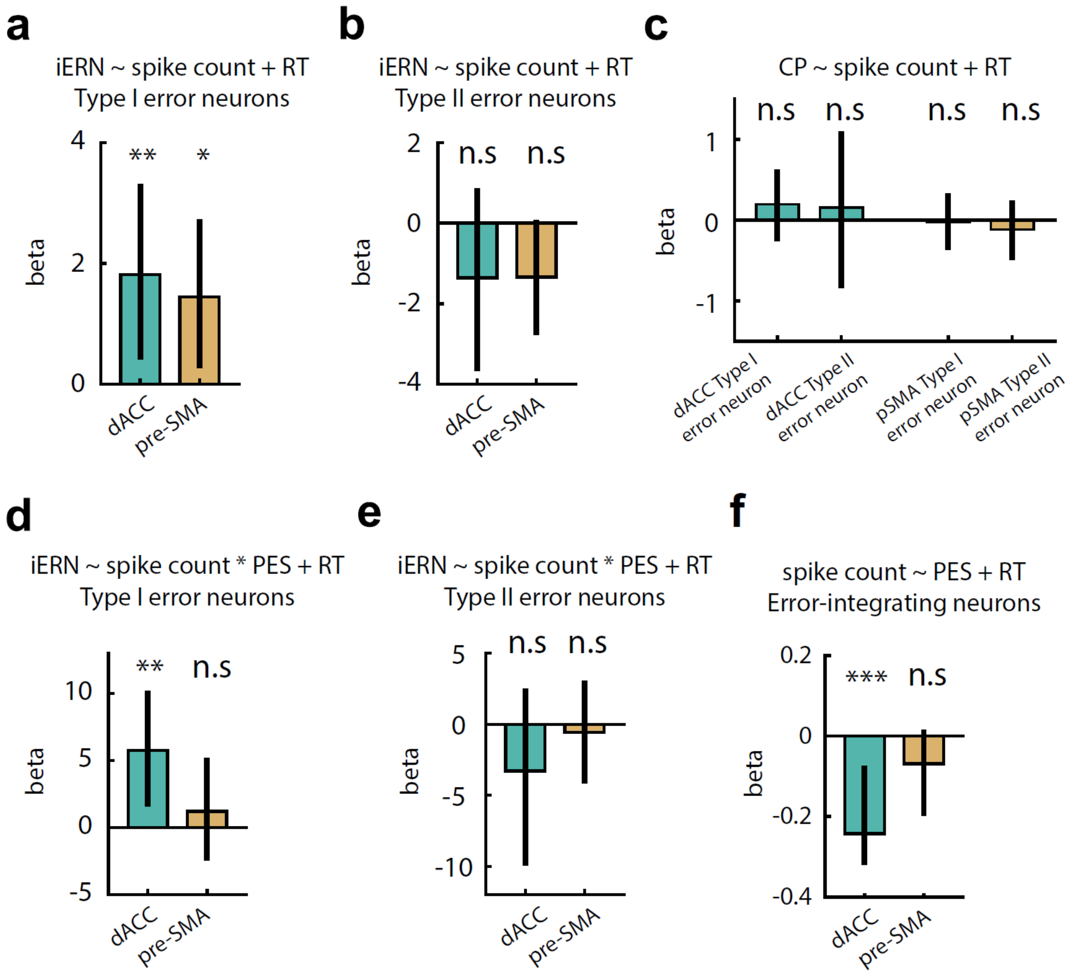

Figure S7. Regression coefficients of GLM models. Related to Figure 6-7.
(a) Regression coefficients for the fixed effect of spike rates (within [0 1s] after button press). The iERN amplitude was significantly correlated with spike of Type I error neurons in both dACC (p = 0.008, likelihood ratio = 6.56, likelihood ratio test) and pre-SMA (p = 0.012, likelihood ratio = 5.81, likelihood ratio test).
(b) Same as in (a) but for Type II error neurons. The iERN amplitude did not correlate significantly with spike rates (within [0 1s] after button press) in either dACC (p = 0.19, likelihood ratio = 1.64) nor pre-SMA (p = 0.07, likelihood ratio = 3.36, likelihood ratio test)
(c) Regression coefficients for the fixed effect of spike rates (within [0 1s] after button press). The CP (‘correct potential’) was not correlated with the spike rates of any types of error neurons in either dACC (for Type I error neurons, p = 0.34, likelihood ratio = 0.92; for Type II error neurons, p = 0.74, likelihood ratio = 0.11, likelihood ratio tests) or pre-SMA (for Type I error neurons, p = 0.88, likelihood ratio = 0.023; for Type II error neurons, p = 0.48, likelihood ratio = 0.49, likelihood ratio test).
(d) Regression coefficients of the interaction term between the spike rate (within [0 1s] after button press) of the Type I error neurons and PES levels. The correlation between iERN amplitude and the spike rates of the Type I error neurons in dACC was stronger when PES was larger (p = 0.009, likelihood ratio = 6.56, likelihood ratio test). The same relationship did not hold significantly for Type I error neurons in pre-SMA (p = 0.5, likelihood ratio = 0.44, likelihood ratio test).
(e) Same as in (d) but for Type II error neurons. The strength of correlation between iERN amplitude and the spike rates of the Type II error neurons did not vary significantly between PES levels in either dACC (p = 0.67, likelihood ratio = 0.18, likelihood ratio test) or pre-SMA (p = 0.48, likelihood ratio = 0.5, likelihood ratio test).

(f) Regression coefficients of the fixed effect of PES levels. The spike rates of the error-integrating neurons were strongly correlated with PES in dACC ($p < 0.001$, likelihood ratio = 15.76, likelihood ratio test), but only marginally so in pre-SMA ($p = 0.07$, likelihood ratio = 3.31, likelihood ratio test).
Error bars represent 95% confidence interval obtained from parametric bootstrapping. '*', '**', '***' mark statistical significance with p values ≤ 0.05, ≤ 0.01 and ≤ 0.001 respectively using the likelihood ratio test. 'n.s' marks not significant ($p > 0.05$).

2.5.2 Supplementary Tables

Table S1. Subjects recorded. Related to Figure 1.

List of all subjects recorded.

ID	Sex	Age	Epi Diagnosis	Macro recording performed	Sessions performed
H11	M	16	right lateral frontal	N	2
H14	M	31	Bilateral indep. temporal	N	2
H16	F	34	right frontal	N	2
H17	M	19	left inferior frontal	N	3
H18	M	40	Right temporal	N	1
H19	M	34	Left frontal	N	1
H21	M	20	Not localized	N	2
H28	M	23	Right mesial temporal	N	1
H31	M	30	Right temporal	N	1
H41	M	19	Right posterior temporal	N	1
H42	M	29	Not localized	N	1
H49	F	54	Right amygdala and hippocampus	N	2
C24	F	47	Not localized	N	2
C25	F	36	Bilateral indep. Temporal	N	2
C26	F	56	Right temporal	N	1
C27	M	45	Left temporal	N	1
C29	M	19	Left temporal neocortical	N	4
C31	M	31	Left temporal neocortical	N	3
C32	M	19	Not localized	N	1
C33	F	44	Right temporal	N	4
C34	M	70	Bilateral temporal	N	5
C35	M	63	Left temporal neocortical	Y	6
C36	M	45	Right Hippocampus	Y	6
C37	F	33	Right Hippocampus	Y	11
C39	M	26	Right insula	Y	6
C40	M	25	Right motor cortex	Y	4
C42	F	25	Not localized	Y	5
C47	M	33	Right mesial temporal	Y	2
C48	F	32	Left medial temporal	Y	1

Table S2. Percentage and average spike rate of neurons. Related to Figures 2-3.
Summary of percentages and average spike rates (±s.d.) of neuronal categories. Neurons of the types marked as "NA" were found not more than expected by chance.

dACC	**Error neurons**		**Error-integrating neurons**	
Types	**I**	**II**	**I**	**II**
Spike rate (Hz, ±s.d)	2.61±2.8	3.24±2.2	2.77±2.1	NA

Percentage (%)	24.8	8.8	11.9	NA

pre-SMA	**Error neurons**		**Error-integrating neurons**	
Types	**I**	**II**	**I**	**II**
Spike rate (Hz, ±s.d)	2.47±2.3	3.6±3.3	NA	3.47±3.6
Percentage (%)	27.4	18.6	NA	13.6

2.6 Experimental Methods and Subject Details

Depth electrode subjects. 29 patients (see Table S1 for age and gender) who were evaluated for possible surgical treatment of epilepsy using implantation of depth electrodes volunteered for the study and gave written informed consent. We only included patients with well-isolated single-neuron activity on at least one electrode in the areas of interest.

Scalp EEG subjects. 12 naïve non-surgical control subjects participated (seven females). All participants gave informed consent, and the protocol was approved by the Caltech Institutional Review Board. A BioSemi Active2 system collected EEG data and laptop event triggers at 1024 Hz. Electrode montages were in Biosemi's standard 64 or 128 channel cap arrays, with additional electrodes for right eye vertical EOG.

Method Details

Task. Subjects performed a speeded version of the classical color-word Stroop task. In each trial, the stimulus was chosen randomly to be one of the three words (red, green and blue) printed in either red, green, or blue color (see Fig. 1a). Subjects were instructed to indicate the color the word was printed in as quickly as possible (ignoring the meaning of the word) by pressing one of the three buttons on an external response box (RB-740, Cedrus Corp., San Pedro, CA). The stimulus was replaced with a blank screen immediately after the button press. One second after button press, subjects were given one of three types of feedback: correct, incorrect, or "too slow". An adaptive staircase procedure was used to establish a reaction time threshold such that 10-15% of trials were rated as "too slow" regardless of the accuracy of the response. Correct trials with 'too slow' feedback were not considered as error trials. This dynamic threshold was implemented to encourage faster responses. The inter-trial interval varied randomly from 1-1.5s. The task was administered in blocks of 90 trials, 30-40% of which were incongruent (randomly intermixed). Patients performed 3 – 6 blocks in a session. Trials with RT larger than three standard deviations above the mean were excluded for all analyses. The task was implemented using the Psychophysics Toolbox (Brainard, 1997). Scalp EEG participants performed the same task as described above (350 trials total).

Electrophysiology. We recorded from up to 4 electrodes in each subject (bilateral dACC and pre-SMA), each with eight high-impedance microwires at the medial end and eight low-impedance macro-contacts along the shaft (Fig. S1a; AdTech Medical Inc.). Here, we used only the most medial macro contact (which is located within the dACC or the pre-SMA) and all microwires. We recorded the broadband 0.1Hz-9kHz continuous extracellular signal with a sampling rate of 32-40kHz from each microwire and with a sampling rate of 2kHz from each macro-contact (ATLAS, Neuralynx Inc., Bozman, MT). One microwire on each electrode served as a local reference (bi-polar recording).

Electrode localization. For each patient, two structural MRI scans were obtained: one before and one after implantation. Electrodes were localized based on these scans in each individual patient. Only electrodes that could be clearly localized to the dACC (cingulate gyrus or cingulate sulcus; for patients with a paracingulate sulcus, electrodes were assigned to the dACC if they were within the paracingulate sulcus or superior cingulate gyrus or the pre-SMA (superior frontal gyrus) were included. We also merged the subject-specific MRI onto an Atlas brain, which was used only for visualization purposes (all localization was based on individual MRIs without using an Atlas). We described the analysis pipeline for transforming the post-implantation MRI into the same space as a MNI152-based atlas as in a previous study(Minxha et al., 2017).

Spike detection and sorting. We filtered the raw signal with a zero-phase lag filter in the 300-3000Hz band. Spikes were detected and sorted using a template-matching algorithm (Rutishauser et al., 2006). We carefully evaluated isolation quality of units and analyzed only well-isolated single units. We used the following criteria (see Fig. S1d-i): i) percentage of ISIs smaller than 3ms, ii) SNR of the waveform, calculated as the ratio of the peak amplitude of the mean waveform of each cluster and the standard deviation of the noise, iii) the pairwise projection distance as provided by the projection test (Pouzat et al., 2002) between all pairs of neurons isolated on the same wire, iv) the modified coefficient of variation of variability in the ISI (CV2), and v) the isolation distance (Schmitzer-Torbert et al., 2005, Harris et al., 2000), which we computed as previously defined (Rutishauser et al., 2006). Channels with inter-ictal epileptic events were excluded. All research protocols were approved by the institutional review boards of Cedars-Sinai Medical Center, Huntington Memorial Hospital and the California Institute of Technology.

Quantification and Statistical Analysis

Behavioral analyses. We constructed a mixed-effect one-way ANOVA model with nested design to test for the Stroop effect. We entered reaction time (RT) as the response variable, the stimulus type ('congruent' or incongruent') as the fixed effect and session numbers nested within subject ID as a random effect. To test for post-error slowing (PES) effects, we used two complementary approaches. First, we constructed a mixed-effect one-way ANOVA model with nested design, with RT as the response variable, the previous outcome and current trial stimulus type ('congruent' or 'incongruent') as the fixed effects and the session numbers nested within subject ID as the random effect. For this model, we also included an interaction term between the two fixed effects. Second, we identified quadruplets of trials that formed a 'CCEC' sequence ('C', correct trial. 'E', error trial) and the stimulus types (congruent or incongruent) were matched for the second and fourth trial within this sequence. This ensured that the PES measure was not confounded by the Stroop effect. For each quadruplet, we then defined the trial-by-trial PES as the difference in RT between the fourth and the second trial in this sequence. We then compared the mean of the trial-by-trial PES extracted this way with zero using a t-test to confirm the statistical significance of PES. This PES measure was used for subsequent iERN amplitude-error neuron spike rates correlation analyses and spike-field coherence analyses. This method restricted the post-correct trials to a subset that directly preceded the post-error trials to avoid confounding factors due to non-specific RT slowing, a caveat previously described (Dutilh et al., 2012).

Selection of neurons. We only considered neurons that had a mean spike rate > 0.5 Hz. We sought neurons whose spike rate differed significantly between trial types of interest in two epochs that were defined with respect to stimulus onset or action onset (button press): (i) neurons signaling errors ('error neurons'), (ii) neurons signaling preceding trial accuracy ('error-integrating neurons'), (iii) neurons signaling conflicts ('conflict neurons'). We fit a generalized linear model (GLM) to each neuron (using matlab function "fitglm.m") and then evaluated whether the model explained significant variance to determine whether a neuron was selective or not for a variable of interest. We entered the spike count in the epoch of interest as the response variable. We entered two predictor variables: i) a dummy variable coding for either trial outcome or previous trial outcome, and, ii) RT (to control for RT effect). A neuron was significantly selective for the outcome predictor variable if the p value for the first predictor was below 0.05 (p value as returned from the fitglm function). The epoch of interest for the error neurons was a 1 sec epoch starting immediately after button press ('post-action epoch' or 'postBP epoch', see Fig. S2a), comparing between error and correct trials. Only sessions with at least 7 error trials were considered for selecting error neurons, a minimum number of errors that has been demonstrated to be sufficient for stable error signals (Olvet and Hajcak, 2009). The epoch of interest for error-integrating neurons was -0.5 to 0.5s (1s length) centered on stimulus onset ('peri-stimulus epoch', see Fig. S2a), comparing between EC and CC trials. The epoch of interest for conflict neurons was 0 to 0.5s after stimulus onset ('post-stimulus epoch', see Fig. S2a), comparing between correct congruent and correct incongruent trials.

Each group of neurons was further divided into two sub-categories according to the sign of the spike rate difference (the sign of the regression coefficient of the outcome variable predictor; Type I and II, respectively; Fig. S2b). To estimate chance levels of this selection procedure, we repeated the selection procedure (two-tailed bootstrap) 1000 times after randomly permuting the labels to estimate a null distribution (see Fig. 3a; for conflict neurons, see Fig. S3e). We only analyzed groups of neurons with a size larger than expected by chance (p < 0.05).

Single-neuron and group-averaged post-stimulus time histograms (PSTHs) were constructed using non-overlapping bins of 200ms width. PSTH plots were not smoothened and data points were plotted with respect to the center of the bin. Before averaging across neurons, spike rates for each neuron were standardized by subtracting the mean and dividing with the standard deviation of the baseline (-0.7 to -0.2s relative to the stimulus onset).

Single-neuron ROC analysis. For each neuron, a receiver-operating characteristic (ROC) curve was constructed based on the spike rate in the time windows of interest. The ROC was parametrized by a threshold that varied from the lowest to the highest spike rates in 25 linearly-spaced steps. For each threshold, trials were classified as 'label 1' or 'label 2' according to whether the spike rate in a given trial was higher or lower than this threshold. True positive rates ('TPR') and false positive rates ('FPR') were then derived by comparing the assigned labels with the true labels for each threshold. The area under the curve (AUC) of the ROC was used as a summary metric. In order to aggregate AUCs from different neurons, we always assigned the trial type with higher spike rates in the ROI to 'label 1'. We estimated the AUC values expected by chance by a permutation test.

For the error neurons (Both Type I and Type II, Fig. 3f), we computed AUC values using error- and correct-trial spike rates in the post-action epoch (0-1s relative to button press). For the error-integrating neurons, we computed AUC values for the spike rates estimated from the following three epochs: (i) 0-1s relative to feedback onset (error vs correct) in the preceding trials, (ii) -0.5-0.5s relative to stimulus onset ('peri-stimulus epoch') in the current trials (EC vs. CC) and (iii) 0-1s after button press in the current trials ('postBP epoch'; error vs correct).

Temporal profile of neuronal response. We used a sliding-window GLM to quantify the temporal profile of information conveyed by neuronal spike rates of a single neuron about trial outcome (error vs. correct; Fig. 4a). We first used a ±200ms bin moved across the spike train on each trial in successive 10ms steps. For each of these bins, we entered the spike count as the response variable and the trial outcome (error or correct) as one predictor variable, and RT as another predictor variable. This is because spike rates of the neurons in both dACC and pre-SMA can carry a component that covariates with reaction time (RT) and the effect of trial outcome on spike rates can be isolated after regressing out the RT effect in this principled way. For each bin-wise GLM model, the effect size of the trial outcome was quantified by a likelihood ratio, derived from a likelihood ratio test comparing the full model with null model (full model minus the trial outcome predictor). We used the time course of the likelihood ratio to estimate for each neuron the point of time at which it first differentiated between trial outcomes (error vs correct; Fig. 4b). These differential latencies were determined as the first point of time at which the effect size was significant by the likelihood ratio test ($p < 0.05$) for a consecutive 15 time steps (i.e 150 ms).

We used a cross-validated partial correlation analyses to determine the time window (post-action vs. post-feedback) in which a population of neurons conveyed the most information about error (Fig. 3g). Here, a Spearman's partial correlation coefficient was computed by correlating the spike rates of error neurons in the postBP epoch, and the trial outcome dummy variable (error coded as 1, correct coded as 0), while controlling for RT on the same trial. Statistical comparisons between group averages of partial correlation coefficients in different time windows were made using Wilcoxon's rank sum test. However, the group averages in the same time window used to previously select neurons is biased towards larger values. Here, we circumvented this problem by using cross validation to assure that the group averages were computed from out-of-sample data not used for selection. For this, we performed 200 runs of cross validation. In each run, we randomly subsampled 80% trials for selecting neurons and used the remaining 20% of trials to compute the partial correlation coefficients between spike rates and the relevant trial variable (levels of stimulus congruence or outcomes).

Single-trial spike train latency. We estimated the onset latency in individual trials using Poisson spike-train analysis (Fig. 4c). This method detects points of time at which the observed inter-spike intervals (ISI) deviate significantly from that assumed by a constant-rate Poisson process. This is achieved by maximizing a Poisson surprise index (Hanes et al., 1995). We used the average spike rate of each neuron as the baseline rate of the underlying Poisson process. Since the error signal is related to action completion, we required that the detected bursts of spikes ended after the action was completed to exclude activation unrelated to button press. We included spikes in a window 300-2000ms after stimulus onset. The statistical threshold for detecting an onset was $p < 0.01$. Repeating the same procedure with a threshold of $p < 0.001$ did not affect our conclusions.

Single-trial iERN amplitude and latency extraction. We determined the amplitude and latency of the iERN on individual trials using the following algorithm. First, for each electrode we determined the peak position of the average iERN waveform within a time window of [-50 200]ms relative to button press. We then defined a time window of 200ms centered on the peak of the average iERN as the region of interest for single-trial estimation. For each trial, we used 'findpeak' (MATLAB) to identify all local negative peaks within this time window and then picked the local peak closest to the peak position of the averaged iERN. This approach determines the contribution of each single trial to the average iERN. Since the timing of the iERN is well understood and known (from the average), the negative peak closest in time has the highest likelihood of being the true single-trial iERN signal. The point of time (relative to button press) and voltage value of this *negative-going* peak was then used as the single-trial iERN latency and amplitude. In Fig. 7a, we assessed whether iERN amplitudes differed between PES levels using a PES modulation index computed from the iERN amplitudes. For this, we first separate the error trials into two groups: one that leads to PES values larger than the median value, and one that leads to PES values smaller than the median value (of this experimental session). We then compute the mean iERN amplitude across these two groups of error trials separately. The PES modulation index is equal to the difference of these two mean values divided by their sum.

Single-trial CP amplitude and latency extraction. We determined the amplitude and latency of the CP on individual trials using the following algorithm. First, for each electrode we determined the peak position of the average iERN waveform within a time window of [-50 200]ms relative to button press. We then defined a time window of 200ms centered on this average CP peak position as the region of interest for single-trial estimation. For each trial, we used 'findpeak' (MATLAB) to identify all local positive peaks within this time window and then picked the local peak closest to the peak position of the averaged CP. The point of time and voltage value of this *positive-going* peak was then used as the single-trial CP latency and amplitude.

ROC analysis of iERN amplitude. For each electrode, a receiver-operating characteristic (ROC) curve was constructed based on the voltage values extracted by the iERN extraction algorithm (see above) on error and correct trials (but not CP values that are extracted by a different algorithm). The ROC was parametrized by a threshold that varied from the lowest to the highest voltage values in 25 linearly-spaced steps. For each threshold, trials were classified as 'label 1' or 'label 2' according to whether the voltage value on a given trial was higher or lower than this threshold. True positive rates ('TPR') and false positive rates ('FPR') were then derived by comparing the assigned labels with the true labels for each threshold. The area-under-the-curve (AUC) of the ROC was used as a summary metric and characterizes how well the iERN amplitude on a given trial is indicative of whether the response was correct or incorrect.

Time-frequency analysis of iEEG signal. We used the Hilbert transform to generate time-frequency representations of the iEEG signal. The continuous raw signal (for the entire task) was first down-sampled from 2kHz to 500Hz and then filtered with fourth-order Butterworth filters centered at 28 linearly-spaced frequencies between 1.2 to 11.7Hz. We used 'filtfilt.m' (MATLAB) to ensure zero-phase distortion and then Hilbert-transformed the filtered data to obtain the corresponding instantaneous amplitude and phase values. Next, we segmented this signal into epochs with respect to time of stimulus onset or button-press separately. Epochs with raw voltage amplitudes larger than 150uV were excluded (<1% of epochs were excluded). Power estimates for each frequency bins were generated by squaring the corresponding instantaneous amplitude, averaged across trials and then combined to form a time-frequency representation. For this, we equalized the trial number and RT across conditions. For normalization, time-frequency spectrograms were divided by the corresponding baseline power for each frequency band and log-transformed into decibels (dB). Baseline power was estimated by averaging across all trials in the pre-stimulus epoch (-0.7s to -0.2s relative to stimulus onset). To test for a correlation between iERN amplitude and theta-band power, we computed the Spearman's rank correlation coefficient for each session and tested the mean of correlation coefficients versus zero. To analyze induced power, we repeated above analyses after subtracting event-related potentials. For this, we first computed the event-related potentials and then subtracted these from each trial for each condition (error and correct trials) separately.

Multi-level models. We constructed linear multi-level models (Aarts et al., 2014, Winter, 2013) to test for relationships between RT, iERN amplitude, and error neuron spike rates. For all of the following analyses, we used only data from error trials. For Fig. 6a, in the bin-wise model we entered iERN amplitude as the response variable, spike counts in each ±300ms bin (the center of the bin moved from -0.5s to 2s relative to button press in steps of 10ms) and RT as the fixed effects, session number as the random intercept and cell number nested within subject ID as the random slope for the effect of spike counts. For Fig. 6c, we entered iERN amplitude as the response variable, RT as the fixed effect, session number as the random intercept and session number nested within subject ID as the random slope for the effect of RT. For Fig. 7b, the model setup is the same as that in Fig. 6a except that we added a dummy variable ('PES levels') indicating whether an error trial corresponds to larger (assigned "1") or smaller (assigned "0") PES than the median PES (of the session) and estimated it as the main effect and its interaction with the spike counts. For Fig. 7c, the bin-wise model has the spike rates of error-integrating neurons within each ±300ms bin as the response variable, the PES level and RT as the fixed effects and session number nested within subject ID as the random slope for the effect of RT. For Fig. S7d, the spike counts of error neurons used in the models were all within the postBP epoch ([0 1s] after button press). The statistical significance of all the models described above was evaluated by a model comparison approach(Winter, 2013). Using the likelihood ratio test, we derived the likelihood ratio by comparing the full model and a null model obtained from the full model by removing the effect of interest, leaving all the other fixed or random effects unchanged. The log likelihood ratio distributes asymptotically as a chi-squared distribution and a theoretical p-value can be computed. For Fig. 6a and 7b-c, we performed cluster-based permutation test to control for multiple comparison (Maris and Oostenveld, 2007). To generate an empirical null distribution (1000 permutations) of likelihood ratio for each bin, we permuted the iERN amplitude data so that each iERN amplitude no longer matched with the spike rate data, while keeping the rest of the model unchanged. We then derived the likelihood ratio using the permuted data by the same model comparison approach. During each iteration, we thresholded the likelihood ratio at the value of 3.84 to identify connected clusters, and then computed the sum of likelihood ratio from each cluster and took the maximum of these sums as the test statistic. The true statistic for the cluster (computed using original un-permuted data) was finally compared with the empirical null distribution to derive a p-value.

Scalp EEG – Analysis. Data were analyzed using Brainstorm 3 (Tadel et al., 2011). Data was re-referenced to average, and then band-pass filtered between 1-16 Hz. Eye-blinks were automatically marked and artifacts removed via peak detection in the VEOG and signal space projection algorithms. Button-press events were added to the EEG record based on the stimulus onset triggers and precise reaction times recorded by the response box (RB-740, Cedrus Inc.). Trial epochs were baseline corrected by the mean potential from -0.7s to -0.2s relative to button-press. To balance correct and error trials in number and reaction time, each subject's correct trials were subsampled by selecting the trials with the RTs most closely matching each error trials' RTs. ERPs were calculated for each subjects' error trials (ERN) and correct trials (CRN). ERN statistics were calculated by taking each subjects' ERP peak negativity between -50ms to 200ms relative to the button press. ERN and CRN peaks were compared across subjects by paired t-test. The control subjects demonstrated a robust Stroop effect (65.2 ± 0.9ms, mean ± s.e.m. across sessions, F(1,11) = 54.07, $p <10^{-10}$, mixed-effect one-way ANOVA with random effect) and post-error slowing (69.0 ± 22.3ms, mean ± s.e.m. across sessions, F (1,32) = 7.3, p = 0.01) and made errors in 14.8 ± 1.3% of trials. During error, but not correct, trials the scalp EEG site Cz revealed an evoked potential analogous to the classical signature of error monitoring expected in this task: the error-related negativity (ERN) (Fig. S6c; mean peak amplitude -50−200 ms relative to button press, paired t-test t (11) = 4.53, p < 0.001). The theta power in error trials is significantly stronger than in correct trials (Fig. S6d; [0 500]ms relative to button press, 2-10 Hz in frequency, paired t-test t(11) = 6.47, p < 0.001).

Waveform analyses. For each neuron, we extracted the trough-to-peak time as the duration between the first negative peak of the mean waveform ('trough') and the first positive peak after the trough (Rutishauser et al., 2015). The mean waveform is obtained by averaging all the waveforms assigned to a particular cluster. We normalized the mean waveforms by its maximal amplitude and inverted the waveforms that have the opposite polarity. We considered neurons with a trough-to-peak time < 0.5ms as 'narrow-spiking' neurons and those >0.5,s as 'broad-spiking' neurons.

References

Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V. & van der Sluis, S. 2014. A solution to dependency: using multilevel analysis to accommodate nested data. *Nat Neurosci,* 17**,** 491-6.

Alexander, W. H. & Brown, J. W. 2011. Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience,* 14**,** 1338-U163.

Amiez, C., Joseph, J. P. & Procyk, E. 2006. Reward encoding in the monkey anterior cingulate cortex. *Cerebral Cortex,* 16**,** 1040-1055.

Aron, A. R., Behrens, T. E., Smith, S., Frank, M. J. & Poldrack, R. A. 2007. Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (MRI) and functional MRI. *J Neurosci,* 27**,** 3743-52.

Aron, A. R. & Poldrack, R. A. 2006. Cortical and subcortical contributions to Stop signal response inhibition: role of the subthalamic nucleus. *J Neurosci,* 26**,** 2424-33.

Bartho, P., Hirase, H., Monconduit, L., Zugaro, M., Harris, K. D. & Buzsaki, G. 2004. Characterization of neocortical principal cells and interneurons by network interactions and extracellular features. *J Neurophysiol,* 92**,** 600-8.

Behrens, T. E., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. 2007. Learning the value of information in an uncertain world. *Nat Neurosci,* 10**,** 1214-21.

Bonini, F., Burle, B., Liegeois-Chauvel, C., Regis, J., Chauvel, P. & Vidal, F. 2014. Action monitoring and medial frontal cortex: leading role of supplementary motor area. *Science,* 343**,** 888-91.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S. & Cohen, J. D. 2001. Conflict monitoring and cognitive control. *Psychol Rev,* 108**,** 624-52.

Brainard, D. H. 1997. The Psychophysics Toolbox. *Spatial Vision,* 10**,** 433-436.

Brazdil, M., Roman, R., Daniel, P. & Rektor, I. 2005. Intracerebral error-related negativity in a simple Go/NoGo task. *Journal of Psychophysiology,* 19**,** 244-255.

Brown, J. W. & Braver, T. S. 2005. Learned predictions of error likelihood in the anterior cingulate cortex. *Science,* 307**,** 1118-1121.

Burle, B., Roger, C., Allain, S., Vidal, F. & Hasbroucq, T. 2008. Error negativity does not reflect conflict: a reappraisal of conflict monitoring and anterior cingulate cortex activity. *J Cogn Neurosci,* 20**,** 1637-55.

Danielmeier, C., Eichele, T., Forstmann, B. U., Tittgemeyer, M. & Ullsperger, M. 2011. Posterior medial frontal cortex activity predicts post-error adaptations in task-related visual and motor areas. *J Neurosci,* 31**,** 1780-9.

Debener, S., Ullsperger, M., Siegel, M., Fiehler, K., von Cramon, D. Y. & Engel, A. K. 2005. Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. *J Neurosci,* 25**,** 11730-7.

Dehaene, S., Posner, M. I. & Tucker, D. M. 1994. Localization of a Neural System for Error-Detection and Compensation. *Psychological Science,* 5**,** 303-305.

Dutilh, G., van Ravenzwaaij, D., Nieuwenhuis, S., van der Maas, H. L. J., Forstmann, B. U. & Wagenmakers, E. J. 2012. How to measure post-error slowing: A confound and a simple solution. *Journal of Mathematical Psychology,* 56**,** 208-216.

Ebitz, R. B. & Platt, M. L. 2015. Neuronal activity in primate dorsal anterior cingulate cortex signals task conflict and predicts adjustments in pupil-linked arousal. *Neuron,* 85**,** 628-40.

Ekstrom, A. 2010. How and when the fMRI BOLD signal relates to underlying neural activity: The danger in dissociation. *Brain Research Reviews,* 62**,** 233-244.

Ekstrom, A., Viskontas, I., Kahana, M., Jacobs, J., Upchurch, K., Bookheimer, S. & Fried, I. 2007. Contrasting roles of neural firing rate and local field potentials in human memory. *Hippocampus,* 17**,** 606-17.

Emeric, E. E., Brown, J. W., Leslie, M., Pouget, P., Stuphorn, V. & Schall, J. D. 2008. Performance monitoring local field potentials in the medial frontal cortex of primates: anterior cingulate cortex. *J Neurophysiol,* 99**,** 759-72.

Emeric, E. E., Leslie, M., Pouget, P. & Schall, J. D. 2010. Performance monitoring local field potentials in the medial frontal cortex of primates: supplementary eye field. *J Neurophysiol,* 104**,** 1523-37.

Falkenstein, M., Hohnsbein, J., Hoormann, J. & Blanke, L. 1991. Effects of Crossmodal Divided Attention on Late Erp Components .2. Error Processing in Choice Reaction Tasks. *Electroencephalography and Clinical Neurophysiology,* 78**,** 447-455.

Frank, M. J., Woroch, B. S. & Curran, T. 2005. Error-related negativity predicts reinforcement learning and conflict biases. *Neuron,* 47**,** 495-501.

Gehring, W. J. & Fencsik, D. E. 2001. Functions of the medial frontal cortex in the processing of conflict and errors. *J Neurosci,* 21**,** 9430-7.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E. & Donchin, E. 1993. A Neural System for Error-Detection and Compensation. *Psychological Science,* 4**,** 385-390.

Gerbrandt, L. K., Lawrence, J. C., Eckardt, M. J. & Lloyd, R. L. 1978. Origin of the neocortically monitored theta rhythm in the curarized rat. *Electroencephalogr Clin Neurophysiol,* 45**,** 454-67.

Godlove, D. C., Emeric, E. E., Segovis, C. M., Young, M. S., Schall, J. D. & Woodman, G. F. 2011. Event-related potentials elicited by errors during the stop-signal task. I. Macaque monkeys. *J Neurosci,* 31**,** 15640-9.

Hajcak, G., McDonald, N. & Simons, R. F. 2003. To err is autonomic: Error-related brain potentials, ANS activity, and post-error compensatory behavior. *Psychophysiology,* 40**,** 895-903.

Hanes, D. P., Thompson, K. G. & Schall, J. D. 1995. Relationship of presaccadic activity in frontal eye field and supplementary eye field to saccade initiation in macaque: Poisson spike train analysis. *Exp Brain Res,* 103**,** 85-96.

Harris, K. D., Henze, D. A., Csicsvari, J., Hirase, H. & Buzsaki, G. 2000. Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. *Journal of Neurophysiology,* 84**,** 401-414.

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M. & Platt, M. L. 2011. Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *J Neurosci,* 31**,** 4178-87.

Holroyd, C. B. & Coles, M. G. H. 2002. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychol Rev,* 109**,** 679-709.

Isoda, M. & Hikosaka, O. 2007. Switching from automatic to controlled action by monkey medial frontal cortex. *Nat Neurosci,* 10**,** 240-8.

Ito, S., Stuphorn, V., Brown, J. W. & Schall, J. D. 2003. Performance monitoring by the anterior cingulate cortex during saccade countermanding. *Science,* 302**,** 120-2.

Jahanshahi, M., Obeso, I., Rothwell, J. C. & Obeso, J. A. 2015. A fronto-striato-subthalamic-pallidal network for goal-directed and habitual inhibition. *Nat Rev Neurosci,* 16**,** 719-32.

Kayser, C., Kim, M., Ugurbil, K., Kim, D. S. & Konig, P. 2004. A comparison of hemodynamic and neural responses in cat visual cortex using complex stimuli. *Cerebral Cortex,* 14**,** 881-891.

Kennerley, S. W., Walton, M. E., Behrens, T. E., Buckley, M. J. & Rushworth, M. F. 2006. Optimal decision making and the anterior cingulate cortex. *Nat Neurosci,* 9**,** 940-7.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., 3rd, Cho, R. Y., Stenger, V. A. & Carter, C. S. 2004. Anterior cingulate conflict monitoring and adjustments in control. *Science,* 303**,** 1023-6.

King, J. A., Korb, F. M., von Cramon, D. Y. & Ullsperger, M. 2010. Post-error behavioral adjustments are facilitated by activation and suppression of task-relevant and task-irrelevant information processing. *J Neurosci,* 30**,** 12759-69.

Koechlin, E. & Hyafil, A. 2007. Anterior prefrontal function and the limits of human decision-making. *Science,* 318**,** 594-8.

Kolling, N., Wittmann, M. K., Behrens, T. E., Boorman, E. D., Mars, R. B. & Rushworth, M. F. 2016. Value, search, persistence and model updating in anterior cingulate cortex. *Nat Neurosci,* 19**,** 1280-5.

Kreiman, G., Hung, C. P., Kraskov, A., Quiroga, R. Q., Poggio, T. & DiCarlo, J. J. 2006. Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. *Neuron,* 49**,** 433-445.

Laming, D. 1979. Choice Reaction Performance Following an Error. *Acta Psychologica,* 43**,** 199-224.

Logothetis, N. K., Kayser, C. & Oeltermann, A. 2007. In vivo measurement of cortical impedance spectrum in monkeys: implications for signal propagation. *Neuron,* 55**,** 809-23.

Luck, S. J. 2014. A closer look at ERPs and ERP components. *In:* Luck, S. J. (ed.) *An introduction to the event-related potential technique.* 2 ed. Cambridge, Massachusetts: The MIT Press.

Luu, P., Tucker, D. M. & Makeig, S. 2004. Frontal midline theta and the error-related negativity: neurophysiological mechanisms of action regulation. *Clin Neurophysiol,* 115**,** 1821-35.

Mansouri, F. A., Koechlin, E., Rosa, M. G. P. & Buckley, M. J. 2017. Managing competing goals - a key role for the frontopolar cortex. *Nat Rev Neurosci,* 18**,** 645-657.

Maris, E. & Oostenveld, R. 2007. Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods,* 164**,** 177-90.

Matsumoto, K., Suzuki, W. & Tanaka, K. 2003. Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science,* 301**,** 229-32.

Matsumoto, M., Matsumoto, K., Abe, H. & Tanaka, K. 2007. Medial prefrontal cell activity signaling prediction errors of action values. *Nat Neurosci,* 10**,** 647-56.

Minxha, J., Mosher, C., Morrow, J. K., Mamelak, A. N., Adolphs, R., Gothard, K. M. & Rutishauser, U. 2017. Fixations Gate Species-Specific Responses to Free Viewing of Faces in the Human and Macaque Amygdala. *Cell Rep,* 18**,** 878-891.

Mitchell, J. F., Sundberg, K. A. & Reynolds, J. H. 2007. Differential attention-dependent response modulation across cell classes in macaque visual area V4. *Neuron,* 55**,** 131-41.

Nachev, P., Kennard, C. & Husain, M. 2008. Functional role of the supplementary and pre-supplementary motor areas. *Nat Rev Neurosci,* 9**,** 856-69.

Narayanan, N. S., Cavanagh, J. F., Frank, M. J. & Laubach, M. 2013. Common medial frontal mechanisms of adaptive control in humans and rodents. *Nat Neurosci,* 16**,** 1888-1895.

Niessing, J., Ebisch, B., Schmidt, K. E., Niessing, M., Singer, W. & Galuske, R. A. 2005. Hemodynamic signals correlate tightly with synchronized gamma oscillations. *Science,* 309**,** 948-51.

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P. & Kok, A. 2001. Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. *Psychophysiology,* 38**,** 752-60.

Nir, Y., Fisch, L., Mukamel, R., Gelbard-Sagiv, H., Arieli, A., Fried, I. & Malach, R. 2007. Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to interneuronal correlations. *Current Biology,* 17**,** 1275-1285.

Olvet, D. M. & Hajcak, G. 2008. The error-related negativity (ERN) and psychopathology: toward an endophenotype. *Clin Psychol Rev,* 28**,** 1343-54.

Olvet, D. M. & Hajcak, G. 2009. The stability of error-related brain activity with increasing trials. *Psychophysiology,* 46**,** 957-61.

Pesaran, B., Vinck, M., Einevoll, G. T., Sirota, A., Fries, P., Siegel, M., Truccolo, W., Schroeder, C. E. & Srinivasan, R. 2018. Investigating large-scale brain dynamics using field potential recordings: analysis and interpretation. *Nature Neuroscience,* 21**,** 903-919.

Pouzat, C., Mazor, O. & Laurent, G. 2002. Using noise signature to optimize spike-sorting and to assess neuronal classification quality. *Journal of Neuroscience Methods,* 122**,** 43-57.

Purcell, B. A. & Kiani, R. 2016. Neural Mechanisms of Post-error Adjustments of Decision Policy in Parietal Cortex. *Neuron,* 89**,** 658-71.

Quilodran, R., Rothe, M. & Procyk, E. 2008. Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron,* 57**,** 314-325.

Rabbitt, P. M. A. 1966. Error Correction Time without External Error Signals. *Nature,* 212**,** 438-&.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A. & Nieuwenhuis, S. 2004. The role of the medial frontal cortex in cognitive control. *Science,* 306**,** 443-7.

Rushworth, M. F. & Behrens, T. E. 2008. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nat Neurosci,* 11**,** 389-97.

Rutishauser, U., Ross, I. B., Mamelak, A. N. & Schuman, E. M. 2010. Human memory strength is predicted by theta-frequency phase-locking of single neurons. *Nature,* 464**,** 903-7.

Rutishauser, U., Schuman, E. M. & Mamelak, A. N. 2006. Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. *J Neurosci Methods,* 154**,** 204-24.

Rutishauser, U., Ye, S. X., Koroma, M., Tudusciuc, O., Ross, I. B., Chung, J. M. & Mamelak, A. N. 2015. Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nature Neuroscience,* 18**,** 1041-+.

Scangos, K. W., Aronberg, R. & Stuphorn, V. 2013. Performance monitoring by presupplementary and supplementary motor area during an arm movement countermanding task. *Journal of Neurophysiology,* 109**,** 1928-1939.

Schmitzer-Torbert, N., Jackson, J., Henze, D., Harris, K. & Redish, A. D. 2005. Quantitative measures of cluster quality for use in extracellular recordings. *Neuroscience,* 131**,** 1-11.

Shenhav, A., Botvinick, M. M. & Cohen, J. D. 2013. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron,* 79**,** 217-40.

Sheth, S. A., Mian, M. K., Patel, S. R., Asaad, W. F., Williams, Z. M., Dougherty, D. D., Bush, G. & Eskandar, E. N. 2012. Human dorsal anterior cingulate cortex neurons mediate ongoing behavioural adaptation. *Nature,* 488**,** 218-21.

Shima, K. & Tanji, J. 1998. Role for cingulate motor area cells in voluntary movement selection based on reward. *Science,* 282**,** 1335-8.

Siegel, M., Donner, T. H. & Engel, A. K. 2012. Spectral fingerprints of large-scale neuronal interactions. *Nat Rev Neurosci,* 13**,** 121-34.

Sirota, A., Montgomery, S., Fujisawa, S., Isomura, Y., Zugaro, M. & Buzsaki, G. 2008. Entrainment of neocortical neurons and gamma oscillations by the hippocampal theta rhythm. *Neuron,* 60**,** 683-97.

Stuphorn, V. & Schall, J. D. 2006. Executive control of countermanding saccades by the supplementary eye field. *Nat Neurosci,* 9**,** 925-31.

Stuphorn, V., Taylor, T. L. & Schall, J. D. 2000. Performance monitoring by the supplementary eye field. *Nature,* 408**,** 857-60.

Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D. & Leahy, R. M. 2011. Brainstorm: a user-friendly application for MEG/EEG analysis. *Comput Intell Neurosci,* 2011**,** 879716.

Trujillo, L. T. & Allen, J. J. 2007. Theta EEG dynamics of the error-related negativity. *Clin Neurophysiol,* 118**,** 645-68.

Tsujimoto, S., Genovesio, A. & Wise, S. P. 2010. Evaluating self-generated decisions in frontal pole cortex of monkeys. *Nat Neurosci,* 13**,** 120-6.

Ullsperger, M. & Danielmeier, C. 2016. Reducing Speed and Sight: How Adaptive Is Post-Error Slowing? *Neuron,* 89**,** 430-2.

Ullsperger, M., Danielmeier, C. & Jocham, G. 2014. Neurophysiology of performance monitoring and adaptive behavior. *Physiol Rev,* 94**,** 35-79.

Vogt, B. A., Berger, G. R. & Derbyshire, S. W. 2003. Structural and functional dichotomy of human midcingulate cortex. *Eur J Neurosci,* 18**,** 3134-44.

Voytek, B., Kayser, A. S., Badre, D., Fegen, D., Chang, E. F., Crone, N. E., Parvizi, J., Knight, R. T. & D'Esposito, M. 2015. Oscillatory dynamics coordinating human frontal networks in support of goal maintenance. *Nat Neurosci,* 18**,** 1318-24.

Wang, C. M., Ulbert, I., Schomer, D. L., Marinkovic, K. & Halgren, E. 2005. Responses of human anterior cingulate cortex microdomains to error detection, conflict monitoring, stimulus-response mapping, familiarity, and orienting. *Journal of Neuroscience,* 25**,** 604-613.

Williams, Z. M., Bush, G., Rauch, S. L., Cosgrove, G. R. & Eskandar, E. N. 2004. Human anterior cingulate neurons and the integration of monetary reward with motor responses. *Nat Neurosci,* 7**,** 1370-5.

Winter, B. 2013. Linear models and linear mixed effects models in R with linguistic applications. *arXiv:1308.5499*.

Wong, Y. T., Fabiszak, M. M., Novikov, Y., Daw, N. D. & Pesaran, B. 2016. Coherent neuronal ensembles are rapidly recruited when making a look-reach decision. *Nat Neurosci,* 19**,** 327-34.

Yeung, N., Bogacz, R., Holroyd, C. B., Nieuwenhuis, S. & Cohen, J. D. 2007. Theta phase resetting and the error-related negativity. *Psychophysiology,* 44**,** 39-49.

Chapter 3. Spike-field coherence during performance monitoring and cognitive control

3.1 Introduction

Cognitive control is flexible. Studies have found that it can be recruited rapidly to resolve conflicts, selectively direct attention to task-relevant stimuli, and inhibit a prepared response. Underlying these modulations of multiple cognitive process are mechanisms by which control can exert influence through functional connectivity between different brain regions, for instance, a strengthening of connectivity between prefrontal cortex and the dorsal anterior cingulate cortex (dACC) during conflict resolution. In hierarchical models of the Flanker task and the Stroop task, cognitive control is modelled as a selective strengthening of functional connectivity between the task-relevant processing units and output units (Cohen et al., 1990). These models are successful in reproducing the behavioral signatures of human performance well. These findings all highlight a central feature about cognitive control: flexible reconfiguration of information representation. What are the possible biophysical mechanisms that underlie such reconfiguration? This question is important in light of my thesis, because psychological models based on behavioral data alone are non-unique: only actual data from the brain can really tell us how the brain solves the problem of communicating between multiple processes to implement cognitive control.

As proposed by Fries et al. (Fries et al., 2001, Fries, 2005), neuronal coherence is a major mechanism by which flexible reconfiguration of functional connectivity can occur. Neuronal groups have been shown to oscillate. In fact, a large part of the EEG literature is concerned with the analysis of oscillations in the EEG data during cognitive functions, and these oscillations are generated by coherent activity in the underlying neuronal populations. Oscillatory neuronal activity poses constraints on the likelihood of spike generation, as well as receptivity of synaptic inputs. These mechanisms have two implications. First, the oscillatory activity in the local neuronal populations can recruit and entrain a single neuron within the same area, forming activity ensembles. As a result, information originally represented by just one neuron is amplified and now is represented by an ensemble of neurons. In addition, this functional coupling is flexible in the sense that a neuron can still decouple from such an ensemble given sufficient external inputs. A physical analogy would be that a child gets on a swing and becomes ‘coupled’ with the swing, but he or she can at any time jump off the swing and becomes decoupled. Second, this oscillatory time window for receiving inputs essentially restrict neurons located remotely that can affect the activity of a given neuron to the subset that is coherently active with this particular neuron. It is possible that the two scenarios work in coordination: synaptic inputs from a certain brain region first act on local neuronal populations through the mechanism of coherence, and the local population also becomes more coherent through interactions among its constituent neurons during this process, amplifying the information carried by the external inputs.

Local field potentials, which are a summation of broad range of synaptic events and subthreshold neuronal activity (including possibly afterpotentials from spiking events), provide us with a measure of average activity in the local neuronal population. The oscillatory components of the LFP serve as a measure of the neuronal oscillation in a specific frequency band. Different frequency bands have been found to correlate with different cognitive processes – and also differentially correlate with other common brain measures, such as BOLD fMRI (see General Discussion). To measure the extent to which a given neuron participates in activity ensembles that involve many neurons in the population, one can measure coherence between spikes of this particular neuron and the LFP generated within the local population, using a metric called the spike-field coherence (SFC). This metric is also useful in quantifying long-range communication using spikes from a given area and the LFP from another area.

We argue that SFC can serve as a mechanism whereby information crucial for cognitive control can be amplified. We specifically tested this hypothesis in the case of error monitoring: the more coherent a local population is within dACC and pre-SMA, the more control is achieved in the form of post-error slowing. In error monitoring, a well-known signal is the error-related negativity (ERN), which is thought to represent inputs to both dACC and pre-SMA that originates from other brain regions (see Chapter 2). This makes error monitoring a very suitable setting to test the hypothesis of coherence as amplification mechanism by analyzing how SFC changes within the local population following the ERN. However, one caveat about studying coherence in the presence of event-related potential such as the ERN is that the ERN itself represents to a large extent phase-resetting of the LFP. Since computing SFC would require both LFP and spikes, a strong phase-resetting in the LFP would bias SFC values, reflecting not the coherence between spikes and LFP, but just LFP synchronization. There are multiple ways to tackle this confound. In the following presentation of some of the preliminary analyses I have done, I attempted to remove this phase-resetting confound by subtracting it from the LFP before SFC computation. However, this could in turn lead to a spurious ERP-like artifact. A better way to control for the confound would be to regress out ERN amplitude from SFC computed with the raw LFP, and analyze the residual. By analyzing how SFC changes over time, we aim to obtain a signature of how a local population of MFC neurons code error information as provided by the ERN (inputs), and then specify control based on this representation.

3.2 Results

3.2.1 Spike-field coherence signature of action outcome monitoring

To gain further insight into the mechanisms that generate the error signals encoded by error neurons, we analyzed local relationships between the LFP and spikes in MFC. Neurons can organize dynamically into assemblies to increase the information saliency, in order to transmit information efficiently to downstream targets (Riehle et al., 1997, Wong et al., 2016, Salinas and Sejnowski, 2001). Given the saliency of errors, we reasoned that neurons in the MFC might form such dynamic assemblies to effectively represent and transmit error information. We measured the extent of each neuron's participation in such an assembly using spike-field coherence (SFC), which quantifies the precision of spike timing of a given neuron relative to the phase of ongoing oscillatory activity in the LFP (Fries et al., 2001, Wong et al., 2016, Rutishauser et al., 2010). The following analysis is based on LFP recorded on the same microwire as each neuron.

SFC between spikes of all recorded MFC neurons and the low frequency (< 4Hz) component of the LFP increased significantly following errors (Fig. 3.1a-c; p < 0.01 for all clusters; contours derived from the cluster-based permutation tests). This increase was accompanied by a simultaneous decrease in SFC to higher frequency 5-10 Hz LFP components (Fig. 3.1b,c; comparisons based on the 'SFC modulation index'; see Eq. 1 in Methods; p < 0.01 for all clusters; contours derived from the cluster-based permutation tests). These patterns of changes in SFC were not seen during correct trials. The low-frequency error-related SFC modulation emerged first in pre-SMA (Fig. 3.1b,c), consistent with a leading role of this brain region. This modulation of SFC was prominent at both the single neuron level (Fig. 3.1a shows examples) as well as at the population level in both brain regions for all recorded neurons (Fig. 3.1b,c) and error neurons alone (Fig. 3.2). Similar patterns of SFC modulation were also seen when considering error neurons alone (Fig. 3.2). In summary, neurons phase-locked to low frequency components of the LFP (< 4Hz) only after errors, suggesting a mechanism whereby error neuron responses are generated through transient functional ensembles that are formed depending on action outcomes.

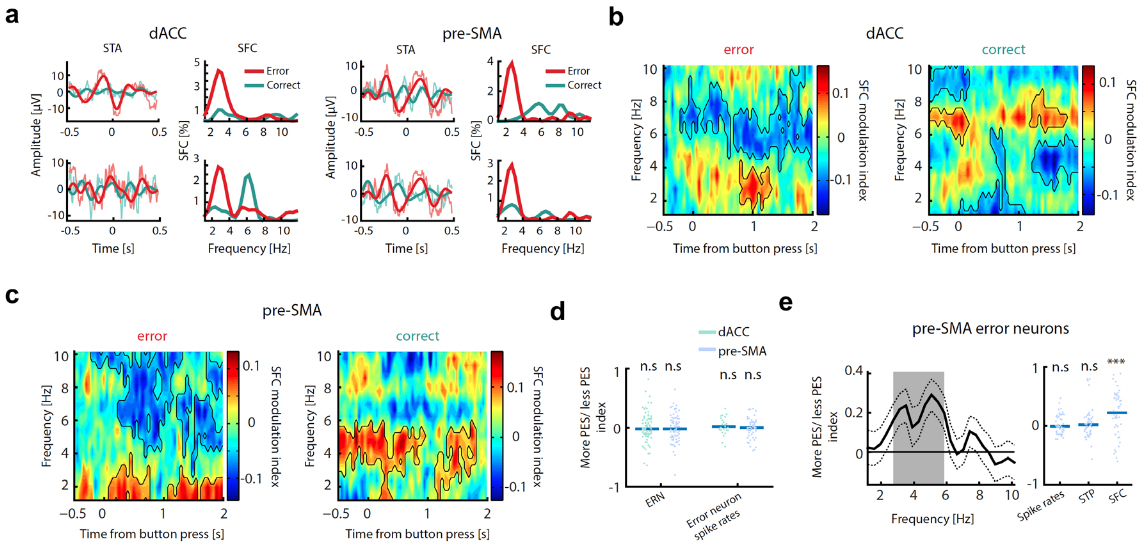

Figure 3.1 Spike-field coherence during errors predicts engagement of control
(a) Spike-trigger average (STA) and spike-field coherence (SFC) for four example neurons (red for error, green for correct). Thin lines, raw STA; thick lines, STA filtered with 2-5Hz fourth-ordered Butterworth band-pass filter. Note the prominent 2-5Hz oscillations in the error STA (red) and that the SFC captures this feature.
(b) SFC modulation index (see Eq. 1 in methods) as a function of time and frequency averaged across all recorded dACC neurons (n= 256). Contour lines delineate significant clusters (p < 0.01) as determined by a cluster-based permutation test. During errors, there was an increase in SFC in the 2-5Hz frequency range with a simultaneous reduction in SFC in the higher (5-10Hz) frequency range. During correct trials, these same neurons increased SFC only in the higher 6-8Hz frequency band.
(c) Same as (b), but for pre-SMA neurons (n = 392). During errors, there was an increase in SFC in the 2-4Hz frequency range with a simultaneous reduction in SFC in the higher (5-10Hz) frequency range. During correct trials, these same neurons increased SFC only in the higher 3-6Hz frequency band.
(d) Error signals alone did not predict the strength of PES. The PES modulation index is defined as the difference between errors that lead to more PES (upper 50% of PES) and those that lead to less PES (lower

50% of PES) divided by their sum (see Eq. 2 in methods). Here, the index was computed using iERN amplitude and spike rates of error neurons (Type I). All comparisons versus zero were not significant ($p > 0.1$; t-test).

(e) SFC predicts strength of behavioral control following errors as measured by post-error slowing (PES), for error neurons in pre-SMA. (left) Shown is the PES modulation index computed using SFC as a function of frequency (left; Type I and II error neurons pooled; 1 to 2s post-action). Grey shading delineates frequencies with a significant difference as determined by a cluster-based permutation test ($p = 0.003$). (right) Firing rates and power (as assessed by spike-triggered power, see methods) in the same time window and frequency range (1 to 2s post-action, 3-6 Hz, as determined from left side) did not predict the extent of PES ($p > 0.1$ for both comparisons, t-test versus 0). By contrast, the SFC was predictive ($p < 0.001$, t-test versus 0; see also left side).

'*', '**', and '***' mark statistical comparisons with p value ≤ 0.05, ≤ 0.01, or ≤ 0.001, respectively. Error bars represent ± s.e.m across cells. All data used in this figure were recorded using micro-electrodes.

3.2.2 Spike-field coherence during errors predicts the extent of post-error slowing

Given the strong error-related modulation of the SFC, we next investigated whether the SFC might serve as a mechanism for engaging behavioral control processes. Specifically, we tested whether the strength of SFC on an error trial predicts the extent of slowing in the next trial (PES). We again partitioned error trials based on a median-split of the PES magnitude. We then compared whether neural signals differentiated between these two groups using the "more/less PES modulation index" (see Eq. 2 in Methods). We found that the error signals analyzed above (spike rates of error neurons in the 0-1s after the erroneous actions and iERN amplitude) did not predict the extent of post-error slowing (Fig. 3.2d; $p > 0.1$ for all comparisons versus zero using t-test). By contrast, the SFC was predictive: the strength of SFC computed using spikes emitted by pre-SMA error neurons during later part of error trials (1-2s after the erroneous actions) predicted the extent of reaction time slowing on the next trial (Fig. 3.2e; more vs. less PES, $p = 0.003$, significant frequency range was obtained by cluster-based permutation tests; significant after Bonferroni's correction at the level of $q = 0.0125$). This effect was only significant for error neurons in pre-SMA (both Type I and Type II), but not for those in dACC or non-error neurons in either brain region ($p > 0.05$, cluster-based permutation tests). In addition, the SFC computed with spikes emitted by pre-SMA error neuron in the early part of error trials (0-1s after button press) was not predictive. This result suggests that engagement of behavioral control follows error detection. As a comparison, we also tested whether spike rates or LFP power in this later time window (thus the same data used to compute SFC that is predictive of PES) could also predict the extent of PES (see Methods for details). We found that these metrics were not predictive of PES (Fig. 3.2e, bar plots), highlighting the importance of spike timing of error neurons relative to ongoing oscillations in engaging behavioral control.

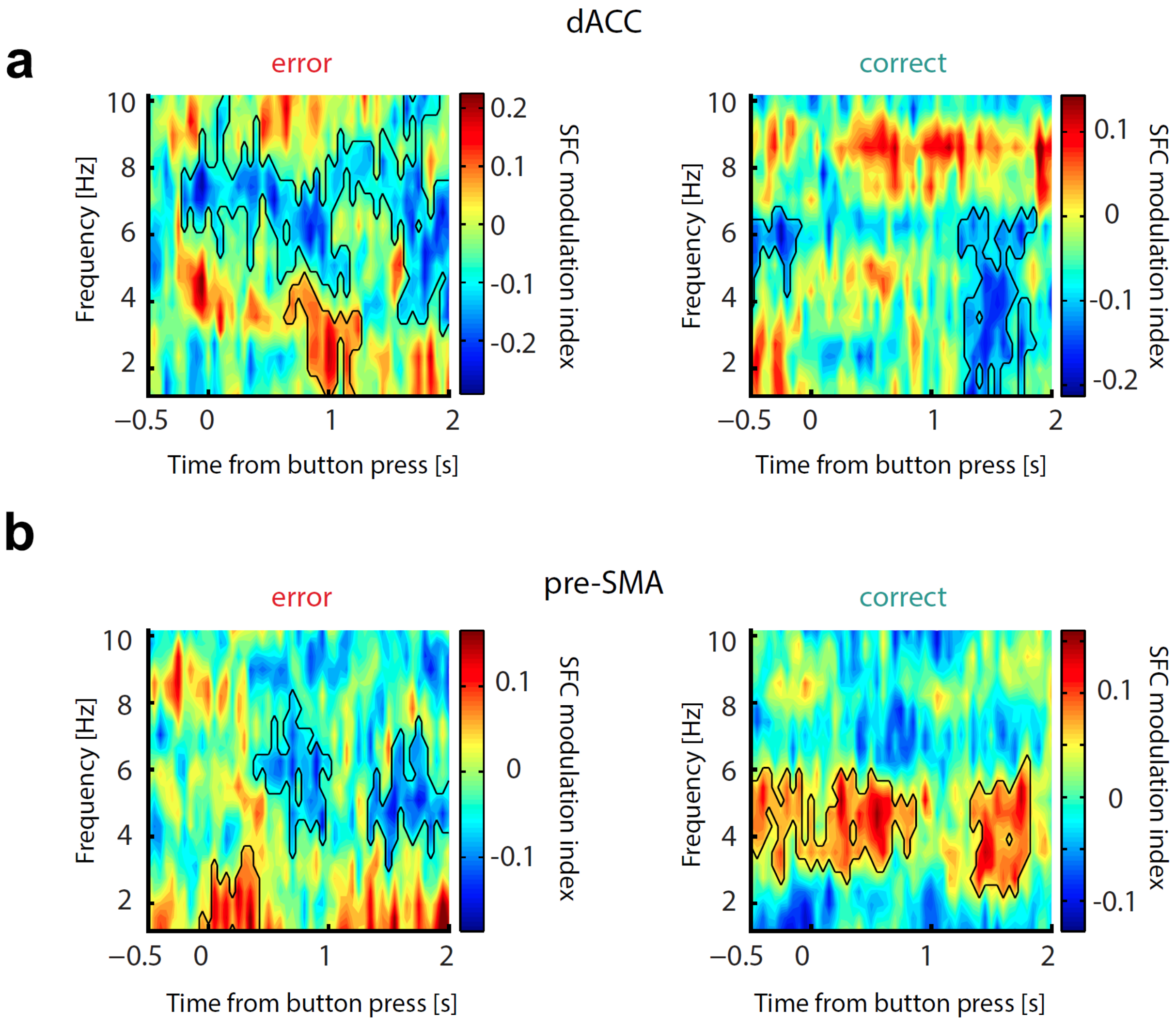

Figure 3.2 Spike-field coherence of error neurons
(a) SFC modulation index (see Eq 1 methods) as a function of time and frequency for error neurons in the dACC (Type I and type II pooled; n= 74). Contour lines delineate significant clusters ($p < 0.05$) as determined by a cluster-based permutation test. During errors (left), there was an increase in SFC in the 2-5Hz frequency range with a simultaneous reduction in SFC in the higher (5-10Hz) frequency range. This pattern was not seen during correct trials (right).
(c) Same as (b), but for pre-SMA error neurons (Type I and type II pooled; n = 163). Contour lines delineate significant clusters ($p < 0.01$) as determined by a cluster-based permutation test. During errors, there was an increase in SFC in the 2-4Hz frequency range with a simultaneous reduction in SFC in the higher (5-10Hz) frequency range. During correct trials, these same neurons increased SFC only to the higher 3-6Hz frequency band. This pattern is consistent with using all the neurons in pre-SMA (Fig. 5b, right).

3.3 Discussion

Neurons in the MFC coordinated their activity not only transiently during iERN generation, but also in a more prolonged fashion as reflected by their phase-locking properties relative to the local LFP. Although errors were accompanied by a power increase in both the 2-5Hz and 5-10Hz bands (Fig. 2.5e in Chapter 2), MFC neurons simultaneously increased their phase-locking to the former and reduce it to the latter band. A possible interpretation is that the MFC neurons disengage from functional ensembles that are active when there is no error, and transiently form different ensembles that operate at a low frequency mode to amplify and saliently represent information about errors. This low frequency band could thus serve as a channel for effectively broadcasting error information to distant regions, such as the prefrontal cortex (Zhou et al., 2016), sensory cortices, motor cortex (Danielmeier et al., 2011), response-inhibition network(Aron et al., 2007, Aron and Poldrack, 2006), parietal cortex (Purcell and Kiani, 2016, Zhou et al., 2016) and hippocampus (Ullsperger et al., 2014), to facilitate task-relevant information processing, motor control and learning (Fujisawa and Buzsaki, 2011).

Our results highlight the importance of spike timing relative to neural oscillations in performance monitoring and suggest that SFC can serve as a substrate for information routing as the performance-monitoring system communicates with other brain regions that maintain flexible goal information. One candidate region for such communication is the lateral prefrontal cortex (LPFC) and/or the frontal polar cortex, which may be involved in error awareness and, together with MFC, jointly evaluate outcomes and determine future actions based on maintained goals (Kouneiher et al., 2009, Kerns et al., 2004, Voytek et al., 2015, Tang et al., 2016, Koechlin and Hyafil, 2007, Tsujimoto et al., 2010, Mansouri et al., 2017). An important avenue for future studies will be to examine the association between neural activity in the MFC and spikes in such distal regions that may receive information from the MFC. The present study offers new insights into the mechanisms of ERN generation and provides potential neural targets for developing closed-loop intervention strategies for psychiatric diseases with self-monitoring dysfunctions.

3.3 Methods

The data used in Chapter 2 were analyzed here with spike-field coherence. We used spike-field coherence to quantify the strength of phase-coupling between spikes of individual neurons and the local field potential (LFP) recorded from the same wire. We preprocessed the LFP data as follows. Since the LFP was recorded from the same microwires as the spiking activity, spurious correlations between spikes and the LFP could confound the results of phase coupling (Zanos et al., 2011). We therefore removed spikes using a Bayesian based method that replaces each spike with short snippets of data (1ms before and 2ms after the spike peak) with statistics similar to that of the LFP (Zanos et al., 2011). This method successfully removes the effects of spiking activities on subsequent analyses performed in the lower frequency ranges (Zanos et al., 2011). The LFP was then band-pass filtered between 1 to 100Hz and down-sampled to 250 Hz for further analyses. Because coincidence between evoked potentials and spiking activity alone can potentially lead to a spurious increase in spike-field coherence, we computed the trial-averaged LFP from error and correct trials separately and then subtracted this average from each error and correct trial, respectively. The methods we used for computing the SFC has been described extensively in previous studies (Fries et al., 2001, Rutishauser et al., 2010). In brief, the SFC is computed as a ratio between the spectrum of the spike-triggered average (STA), normalized by the average power spectrum of the LFP segments used to compute the STA (‘spike-triggered power’ or ‘STP’), as a function of frequency. The STA was constructed by extracting LFP segment of ±480ms centered on each spike, followed by averaging. The minimum number of spikes used was 15 for each bin/ROI. This data length provides a frequency resolution of approximately 1Hz before tapering. The power spectra were computed using the multi-taper method as implemented in the Chronux Toolbox (Bokil et al., 2010). The multi-taper method provides a particularly powerful method to estimate single-trial power spectra with a trade-off between variance and frequency resolution that can be easily controlled (Mitra and Pesaran, 1999). We used a time-bandwidth product of TW = 4, resulting in a half-width of 4.2Hz as previously documented (Rutishauser et al., 2010). We equalized the number of spikes used across conditions for all SFC comparisons for each neuron. We computed the SFC modulation index for each condition (error and correct trials, respectively) (Fig. 3.1b,c,e) (Fries et al., 2001), which was defined as:

$$\mathrm{M}_{\mathrm{condition}_k,\mathrm{cell}_i}(\mathrm{t},\mathrm{f}) = \frac{\mathrm{SFC}_{\mathrm{condition}_k,\mathrm{cell}_i}(\mathrm{t},\mathrm{f})\text{-}\mathrm{SFC}_{\mathrm{baseline},\mathrm{cell}_i}(\mathrm{t},\mathrm{f})}{\mathrm{SFC}_{\mathrm{condition}_k,\mathrm{cell}_i}(\mathrm{t},\mathrm{f}) + \mathrm{SFC}_{\mathrm{baseline},\mathrm{cell}_i}(\mathrm{t},\mathrm{f})} \qquad (1)$$

The SFC modulation index normalizes for numerical differences across cells that are due to different numbers of spikes used, because the same numerical difference affects both the nominator and the denominator equally and was thus divided out. The SFC was computed for center frequencies varying from 1.17Hz to 11.7Hz in steps of 0.4Hz. The baseline SFC was computed using spikes within -0.7s to -0.2s relative the stimulus onset and LFP data after removing the stimulus-locked evoked potential. To plot a time course of the SFC modulation index, we estimated its value using spikes in a moving window of ±250ms, advancing from -500ms to 1.5s relative to button press in successive steps of 50ms.

We used a two-sided cluster-based permutation test (Maris and Oostenveld, 2007) to estimate the statistical significance of the SFC modulation index at a given point of time and frequency (Fig. 3.1b,c), or frequency alone when the region-of-interest in time was fixed (Fig. 7e). To estimate the threshold for statistical significance, we first performed the following procedure. We first randomly permuted the trial labels, partitioned the data into two groups and computed SFC as a function of time and frequency separately for each group. A t-value was then computed across cells, for each bin in time and frequency. This process was repeated 1000 times, generating 1000 t-values for every point of time and frequency. We then thresholded the t-values at the value of ±1.66 (two-sided) and clustered the connected sets of significant bins. This procedure was done separately for bins with negative and positive significant t-values. The sum of the t-values was then computed for each of the identified clusters. For each cell, the maximum across the absolute values of all the computed sums was taken as the test statistic. Finally, the same computation of t-values and clustering was performed using data with the original trial labels. The statistical significance of each cluster in the original data was determined by comparing the absolute values of the sum of t-values with the empirically estimated null distribution as described above.

For Fig. 3.1d,e, the error trials were split into two groups: ‘more PES’ represents error trials that lead to PES (for definition, see *Behavioral analyses*) that are greater than the median size of all trial-by-trial PES in a particular session, whereas ‘less PES’ represents error trials that lead to PES that are less than the median. The “more/less modulation index” for SFC (Fig. 7e, left) is computed as follows:

$$\mathrm{M}_{\mathrm{more/less\ PES,cell_i}}(\mathrm{t,f}) = \frac{\mathrm{SFC}_{\mathrm{morePES,cell_i}}(\mathrm{t,f})\text{-}\mathrm{SFC}_{\mathrm{lessPES,cell_i}}(\mathrm{t,f})}{\mathrm{SFC}_{\mathrm{morePES,cell_i}}(\mathrm{t,f}) + \mathrm{SFC}_{\mathrm{lessPES,cell_i}}(\mathrm{t,f})} \qquad (2)$$

To compute modulation indices in Fig. 3.1d,e, we used the same equation (Eq. 2) but replaced the SFC with either iERN amplitude or error neuron spike rates (0-1s relative to button press, the epoch we used to select error neurons as in Fig. 3.1d, or 1-2s relative to button press as in Fig. 3.1e), or spike-triggered spectrum (Fig. 3.1e) to compute the modulation index for these metrics.

References

Aron, A. R., Behrens, T. E., Smith, S., Frank, M. J. & Poldrack, R. A. 2007. Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (MRI) and functional MRI. *J Neurosci,* 27**,** 3743-52.

Aron, A. R. & Poldrack, R. A. 2006. Cortical and subcortical contributions to Stop signal response inhibition: role of the subthalamic nucleus. *J Neurosci,* 26**,** 2424-33.

Bokil, H., Andrews, P., Kulkarni, J. E., Mehta, S. & Mitra, P. P. 2010. Chronux: A platform for analyzing neural signals. *Journal of Neuroscience Methods,* 192**,** 146-151.

Cohen, J. D., Dunbar, K. & McClelland, J. L. 1990. On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychol Rev,* 97**,** 332-61.

Danielmeier, C., Eichele, T., Forstmann, B. U., Tittgemeyer, M. & Ullsperger, M. 2011. Posterior medial frontal cortex activity predicts post-error adaptations in task-related visual and motor areas. *J Neurosci,* 31**,** 1780-9.

Fries, P. 2005. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in Cognitive Sciences,* 9**,** 474-480.

Fries, P., Reynolds, J. H., Rorie, A. E. & Desimone, R. 2001. Modulation of oscillatory neuronal synchronization by selective visual attention. *Science,* 291**,** 1560-3.

Fujisawa, S. & Buzsaki, G. 2011. A 4 Hz oscillation adaptively synchronizes prefrontal, VTA, and hippocampal activities. *Neuron,* 72**,** 153-65.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., 3rd, Cho, R. Y., Stenger, V. A. & Carter, C. S. 2004. Anterior cingulate conflict monitoring and adjustments in control. *Science,* 303**,** 1023-6.

Koechlin, E. & Hyafil, A. 2007. Anterior prefrontal function and the limits of human decision-making. *Science,* 318**,** 594-8.

Kouneiher, F., Charron, S. & Koechlin, E. 2009. Motivation and cognitive control in the human prefrontal cortex. *Nat Neurosci,* 12**,** 939-45.

Mansouri, F. A., Koechlin, E., Rosa, M. G. P. & Buckley, M. J. 2017. Managing competing goals - a key role for the frontopolar cortex. *Nat Rev Neurosci,* 18**,** 645-657.

Maris, E. & Oostenveld, R. 2007. Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods,* 164**,** 177-90.

Mitra, P. P. & Pesaran, B. 1999. Analysis of dynamic brain imaging data. *Biophysical Journal,* 76**,** 691-708.

Purcell, B. A. & Kiani, R. 2016. Neural Mechanisms of Post-error Adjustments of Decision Policy in Parietal Cortex. *Neuron,* 89**,** 658-71.

Riehle, A., Grun, S., Diesmann, M. & Aertsen, A. 1997. Spike synchronization and rate modulation differentially involved in motor cortical function. *Science,* 278**,** 1950-3.

Rutishauser, U., Ross, I. B., Mamelak, A. N. & Schuman, E. M. 2010. Human memory strength is predicted by theta-frequency phase-locking of single neurons. *Nature,* 464**,** 903-7.

Salinas, E. & Sejnowski, T. J. 2001. Correlated neuronal activity and the flow of neural information. *Nat Rev Neurosci,* 2**,** 539-50.

Tang, H., Yu, H. Y., Chou, C. C., Crone, N. E., Madsen, J. R., Anderson, W. S. & Kreiman, G. 2016. Cascade of neural processing orchestrates cognitive control in human frontal cortex. *Elife,* 5.

Tsujimoto, S., Genovesio, A. & Wise, S. P. 2010. Evaluating self-generated decisions in frontal pole cortex of monkeys. *Nat Neurosci,* 13**,** 120-6.

Ullsperger, M., Danielmeier, C. & Jocham, G. 2014. Neurophysiology of performance monitoring and adaptive behavior. *Physiol Rev,* 94**,** 35-79.

Voytek, B., Kayser, A. S., Badre, D., Fegen, D., Chang, E. F., Crone, N. E., Parvizi, J., Knight, R. T. & D'Esposito, M. 2015. Oscillatory dynamics coordinating human frontal networks in support of goal maintenance. *Nat Neurosci,* 18**,** 1318-24.

Wong, Y. T., Fabiszak, M. M., Novikov, Y., Daw, N. D. & Pesaran, B. 2016. Coherent neuronal ensembles are rapidly recruited when making a look-reach decision. *Nat Neurosci,* 19**,** 327-34.

Zanos, T. P., Mineault, P. J. & Pack, C. C. 2011. Removal of spurious correlations between spikes and local field potentials. *J Neurophysiol,* 105**,** 474-86.

Zhou, X., Qi, X. L. & Constantinidis, C. 2016. Distinct Roles of the Prefrontal and Posterior Parietal Cortices in Response Inhibition. *Cell Rep,* 14**,** 2765-73.

Chapter 4 Ex-post Conflict Signals and Mixed Representations

4.1 Introduction

Performance monitoring and cognitive control is indispensable for successful goal-directed behaviors. These functions can be conceptualized as forming a closed-loop feedback control system, enabling the animal to improve upon performance based on external or internal feedback. Specifically, goal-directed action control can be recruited based on monitoring of either ex-ante or ex-post information. Examples of ex-ante information are action outcome prediction, interference between goal-relevant and goal-irrelevant representations, which proactively influences an ongoing action and is thus available only before an action is committed. Ex-post information, on the other hand, is reactive in nature and can be signaled either by external sensory feedback or inferred from learned internal forward models. The ex-post monitoring signals discovered so far are all concerned with action outcomes, with signals representing gaining or losing of rewards and error detection (discussed in Chapter 2, (Fu et al., 2019)). Due to their timing, ex-post monitoring signals influence future performance of similar actions. It is well established that the medial frontal cortex subserves such monitoring functions and is crucially involved in specifying the identity and intensity of control signal under the constraints of control costs and risks. However, the neuronal mechanisms of how these functions are achieved remains unknown. Existing evidence comes largely from the field of functional imaging with known constrains on resolution. This presents a key challenge in resolving the ex-ante and ex-post signals: decisions normally occurs on a fast time scale and signals from before and after the motor output/decision could be intermingled when measured with insufficient temporal resolution. We sought to investigate this problem by recording directly single neuron activity from human MFC, which has the necessary high temporal resolution.

The ex-ante conflict-monitoring signal is supported by evidence from both neuroimaging studies (Botvinick et al., 2001) as well as from human single unit studies (Sheth et al., 2012) . This signal is thought to reflect competition between concurrently active representations of conflicting response options, and is thus greatly attenuated after successful execution of the goal-relevant action because in such case the competition is necessarily resolved. This signal is also hypothesized to influence next trial behavior as well. However, given the timing of this signal is strictly prior to the action, it is rather unclear how it can exert influence on action on the next trial that is separated in time by well over several hundred of milliseconds (a long period on the scale of neuronal spiking activity). Here, we show that a conflict signal, different from the ex-ante conflict signal, emerged after the action is concluded. An ex-post conflict signal offers a key advantage: it collapses the moment-to-moment conflict signal to a ‘summary metric’ of conflict experienced, and thus potentially allows the representation to be more efficient.

Another important aspect of the executed action is reaction time. Internal representation of reaction time could provide crucial input for positioning performance along the speed-accuracy trade-off axis. Since information about reaction time is only available after the action has concluded, any signals monitoring reaction time is necessarily ex post. We show that ex-post monitoring signals related to reaction time indeed exists in the human MFC at the single neuron level as well as population level, and they are carried by a population of neurons separated from the ex-post conflict-signaling neurons. We additionally show that on a trial-by-trial basis, the MFC contained neuronal signals predictive of reaction time adjustment. We consider such signals as related to control mechanisms that help position the system on the speed-accuracy axis compatible with task demand.

Neurons in the prefrontal cortex have been shown to code information in a multiplexing manner (Rigotti et al., 2013). The spike rates of these neurons depend on the specific combination of task variables such that a putative decoder downstream to these neurons, even if they operate linearly, can still decode the complete task configuration efficiently. Given these robust findings, we hypothesize that such a coding scheme could also be implemented in the performance monitoring system. When interference of various sources co-occur, such a high-dimensional representation is capable of conveying the whole profile of interference to other brain areas for control. Even when a downstream neuronal decoder operates in a linear fashion, it can still read out what types of interference are experienced during action performance and implement the specific control mechanism as needed.

Cognitive conflict can occur in the competition between information from a variety of sources, which is evident from the variety of psychological tasks designed to isolate them. For example, in the Simon task interference is induced when a pre-potent propensity to match the spatial location of the key press response with the spatial location of stimulus and spatial needs to be suppressed. This is different from the type of interference induced by competition between the visually salient flanking arrows competing for attention in the Eriksen flanker task. In the Stroop task (Chapter 2), however, the interference by the word-reading response on color naming could represent yet another type of interference. Are there overlap between MFC neurons signaling different types of interference? Are the population codes for conflict shared between different types of interference?

In this study, we investigated this hypothesis using the multi-source interference task (MSIT), which was designed to integrate two sources of interference that are akin to the Simon and Flanker effect, respectively. MSIT has been used in many studies and has robustly yielded BOLD activation in the frontal-parietal attention network (Bush and Shin, 2006), as well as single neuron signaling ex-ante conflict (Sheth et al., 2012). We recorded single neuron activity from the dorsal anterior cingulate cortex (dACC) and pre-supplementary motor area (pre-SMA) from patients with drug-intractable epilepsy, while they performed the word-color Stroop task and the MSIT task. We found neurons that carry an ex-post conflict signal. Importantly, these neurons multiplexed the two types of interference in the MSIT task, reflected as a significant interaction effect between the interference types in a generalized linear model (GLM).

4.2 Results

4.2.1 Task and behavior

We performed re-analyses of the Stroop data (see Chapter 2 for details of patients and setup). Here we focused on seeking a single-trial behavioral signature of cognitive control. The classical congruence sequence effect states that the mean RT of iI is significantly shorter than that of cI, and the mean RT of cC is significantly shorter than that of iC. The interpretation is that the when the preceding trial was an incongruent trial, cognitive control was engaged and so the conflict induced by the incongruent stimulus on the current trial was reduced, resulting in faster RT as the conflict was resolved more quickly than the case where the preceding trial was a congruent trial. Similarly, when the preceding trial was an incongruent trial, cognitive control was engaged so the even though the current trial is congruent, it was re-positioned along the speed-accuracy tradeoff axis so its RT is slower. However, this behavioral effect does not offer a single-trial measure. One caveat is that in the Stroop task there are only three colors and thus altogether 9 unique stimuli, thus repetitions of stimuli could occur and reduced RT for a pair of trials can simply be due to familiarity. One additional caveat is that there can be general changes in attention across the block that complicate analyses of RT sequences. We thus sought a single-trial measure and made sure that trial sequences with repeated were excluded. We extracted two types of quadruplets of trials: ciiI and iccC trials. In this notation, "c" means congruent trial and "i" means incongruent trial. The capitalized letter denotes the current trial stimulus congruence and the rest of the letters denote stimulus congruence on the three preceding trials. The single-trial signature of control was taken as the difference between the fourth and the second trial in this quadruplet. We extracted these RT differences and plotted them in histograms. The mean of RT difference was -16ms ($p = 0.017$, one-tailed t-test given prior hypothesis), showing that the current trial generally has faster RT than the preceding trial with the same stimulus congruence. Thus, we were able to identify the single-trial signature of congruence sequence effect.

Eleven subjects performed 41 sessions of the MSIT task (Fig. 4.1a). As in the previous three Chapters of this thesis, all of the subjects were patients with epilepsy, who had depth electrodes implanted in the brain for monitoring seizures prior to neurosurgery. In this task, three numbers in the set of [0,1,2,3] were printed on the screen; two were the same and the task was to detect the target number that was different (unique) from the other two. Subjects were instructed to indicate when they had found the target number by pressing one of the three keys on the response pad that were labelled with '1', '2' and '3'. '0' thus did not map to any key press. Two types of manipulation of interference were integrated in the task stimuli: visual distraction and spatial interference. The former was akin to the Flanker effect whereas the latter to the Simon effect. Visual distraction (henceforth abbreviated as 'VD') refers to the distracting effect of two task-irrelevant numbers. Since only non-zero numbers mapped to keys on the response pad, when they served as the distractors they created more distraction than when '0' was used as the distractor. Spatial interference (henceforth abbreviated as 'SI') refers to the mismatch between serial key position and the identity of the target number. There were thus four types of trials, categorized according to whether there was visual distraction or spatial interference manipulation in the stimuli. We coded VD and SI with 0 (without the interference) and 1(with the interference), and referred to these four trial types as "VD0SI0" (e.g. "1 0 0"), "VD0SI1" (e.g. "0 2 0"), "VD1SI0" (e.g. "1 2 2") and "VD1SI1" (e.g. "2 1 2").

Using a GLM in which log-transformed RT data were entered as single trials, we found that the main effects of VD and SI are both highly significant (p<10-10, t(7107) = 17.2 for VD and p < 10-10, t(7107) = 9.8 for SI), but the interaction effect between the two predictors was not (p = 0.73, t(7107) = 0.35). Although Figure 4.1 appears to show a non-linear interaction effect (yellow slope greater than blue slope), the slopes between two conditions did not differ significantly when tested again using a two-way ANOVA test (p

= 0.27, F(1) = 1.23). This suggests that the effects of two types of interference on RT was linearly additive (Fig. 4.1a).

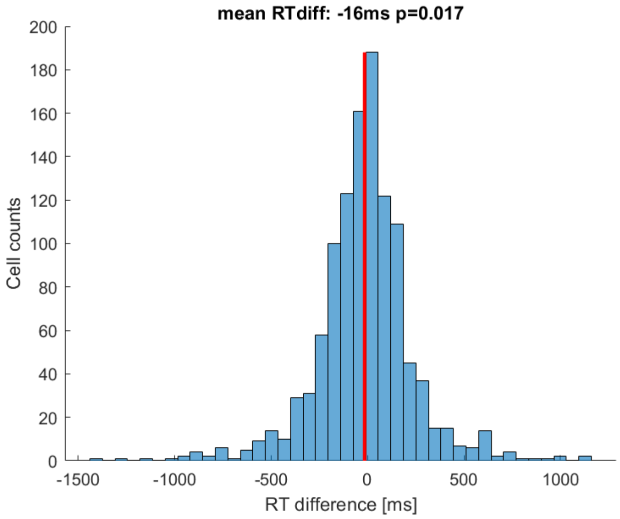

Figure 4.1 (a) Histogram of RT difference. Trial sequences extracted were “ciiI” and “iccC” trials. RT difference was computed by taking the difference between the fourth and the second trial in the trial sequences. We pooled the RT difference from these two trial types. Red vertical bar represents the mean of the distribution.

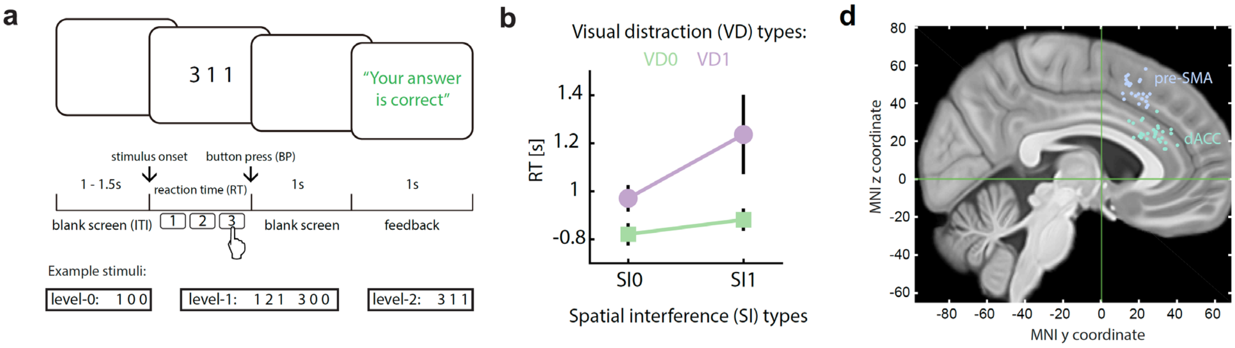

Figure 4.2 (a) Task layout. Feedback screen is only displayed 1s after the response. (b) Average reaction time split by interference types. Each data points represent the mean RT of an interference condition. Blue represents trials with no visual distraction (distractors are ‘0’); yellow represents trials with visual distraction (distractors are non-zero numbers). The effects of two types of interference are additive. Error bars, s.e.m across sessions. (c) Electrode locations.

4.2.2 Ex-post conflict signaling and RT signaling neurons in the Stroop task

Previous studies in the macaque and human literature has suggested that the MFC contained signals related to conflict monitoring, reinforcement history and cognitive control. However, most of the human studies use indirect measures of neural activity (EEG and fMRI) and no studies so far validates the homology of cognitive control between human and non-human primate despite the substantial behavioral task/anatomical differences between the species. It thus remains unknown how these signals are related to neuronal dynamics in the human MFC.

In the Stroop data set, we isolated 433 neurons from dACC and 425 neurons from pre-SMA. We selected neurons using an ANOVA model with interaction. We entered spike rates in the post-action epoch (1s) as the response variable, and congruence level ('0' for congruent trials and '1' for incongruent trials) and RT (log-transformed) as predictor variables. Neurons with p values less than 0.05 were selected. We found that in dACC, 54 neurons (12.5%) were selected by a significant congruence effect, 87 neurons (20%) were selected by a significant RT effect and 38 neurons (8.8%) were selected by a significant congruence × RT interaction. In pre-SMA, 75 neurons (17.7%) were selected by a significant congruence effect, 103 neurons (24.2%) were selected by a significant RT effect and 39 neurons (9.2%) were selected by a significant congruence × RT interaction. To determine whether these number of neurons selected were significantly above chance, a permutation test was run to generate a null distribution of the number of neurons selected (Fig. 4.2 left: dACC, right: pre-SMA). The p values generated by the permutation test were all below 0.05. As is evidence from the single cell examples, the spike rate modulation were temporally restricted to after action execution, demonstrating the ex-post nature of the signals. There were neurons that signaled stimulus congruence exclusively, but their spike rates did not correlate with RT (Fig. 4.4). There were also neurons that signaled RT exclusively, but the correlation between spike rates and RT did not differ between congruent and incongruent trials (Fig. 4.5). Interestingly, a separate class of neurons exist: they signaled RT in a congruence-dependent way, as captured by the interaction effect between RT and stimulus congruence in the ANOVA model used to select them (Fig. 4.6).

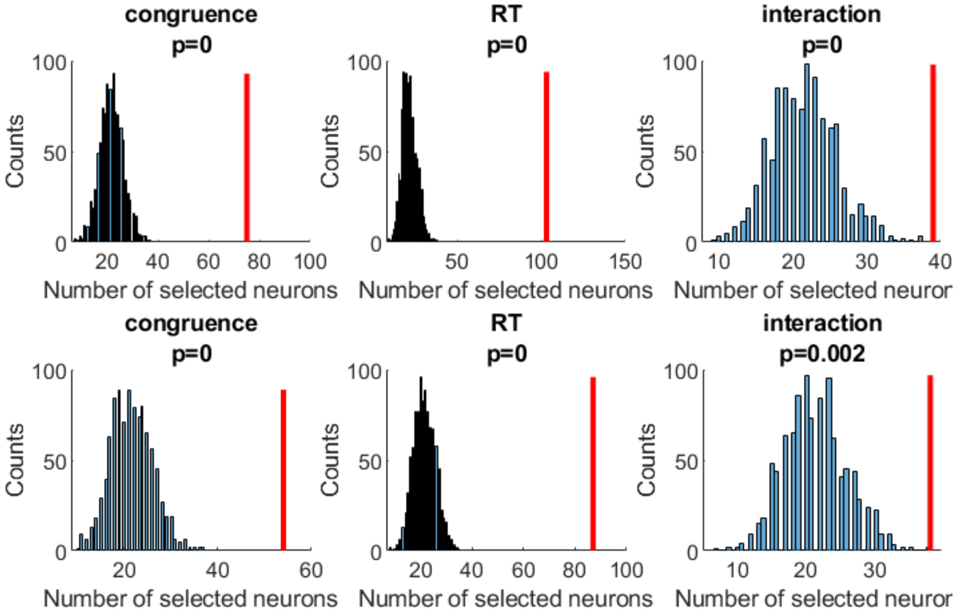

Figure 4.3 Permutation tests for univariate neuron selection

Top row: dACC neurons; lower row: pre-SMA neurons. Here the red vertical bars represent the number of selected neurons by p-values from each predictors in the ANOVA. The histograms show the null distribution of the number of selected neurons using a permutation test.

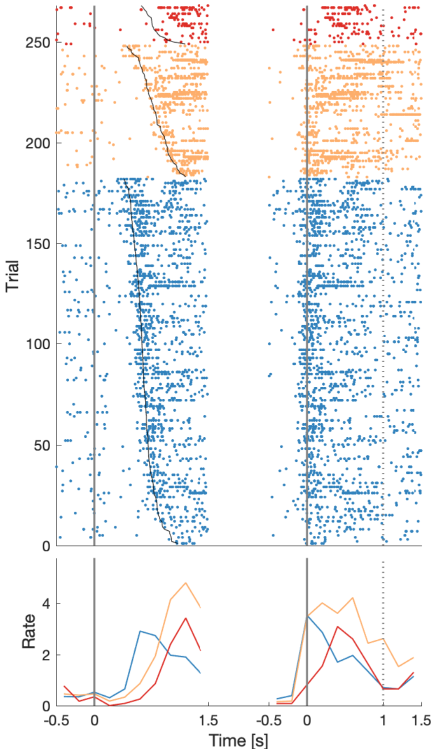

Figure. 4.4 Example ex-post congruence signaling neurons. Raster plots showed data were aligned with stimulus onset (t = 0, left) and time of button presses (t = 0, right) and were sorted by RT (black lines). Here, the neuron on the left signaled stimulus congruence by having a higher spike rates on correct incongruent trials (yellow) than on correct congruent trials (blue). The neuron on the right had the reverse spike rate patterns. Red represents error trials. The spike rate modulation occurred after the button presses.

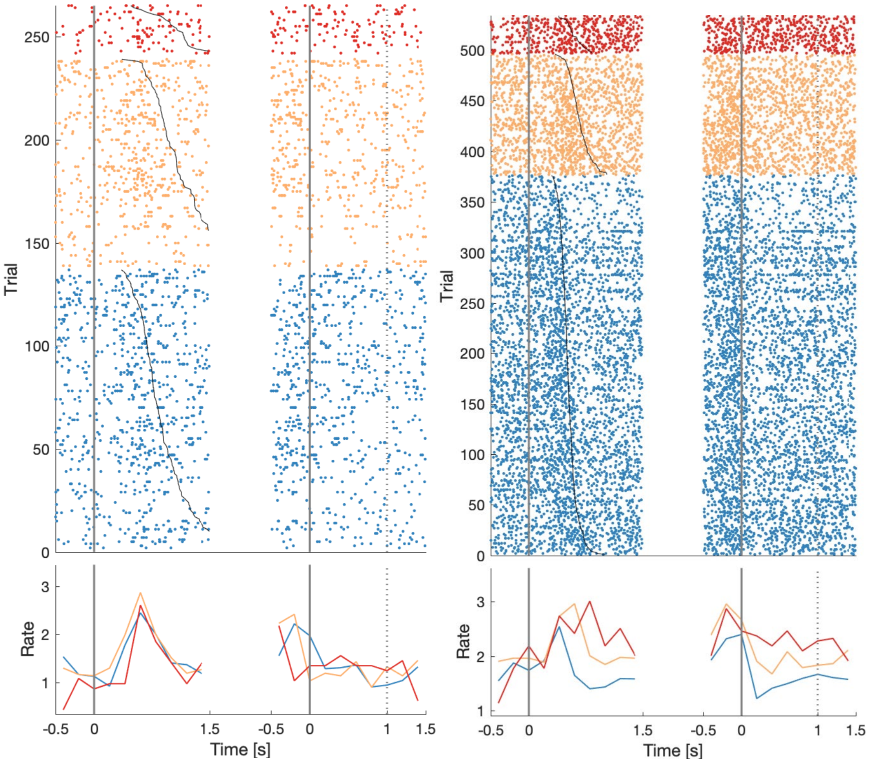

Figure 4.5 Example ex-post RT signaling neurons. Raster plots showed data were aligned with stimulus onset (t = 0, left) and time of button presses (t = 0, right) and were sorted by RT (black lines). Here, the neuron on the left signaled RT on a trial-by-trial basis: spike rates on both correct congruent (blue) and correct incongruent (yellow) trials decreased as RT increased. The neuron on the right had the reverse spike rate patterns. Red represents error trials. The spike rate modulation occurred after the button presses.

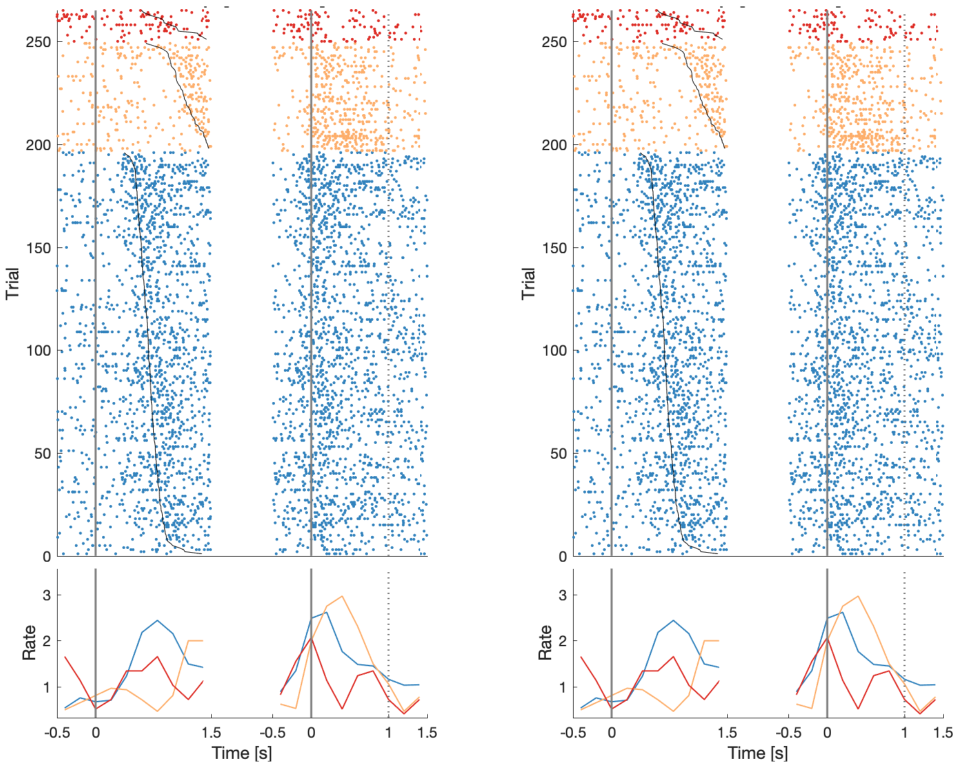

Figure 4.6 Example ex-post congruence-RT interaction neurons. Raster plots showed data were aligned with stimulus onset (t = 0, left) and time of button presses (t = 0, right) and were sorted by RT (black lines). Here, the neuron on the left signaled RT but this depended on stimulus congruence. The spike rates increased on correct incongruent trials (yellow) but decreased as RT increased. The neuron on the right had the reverse spike rate patterns. Red represents error trials. The spike rate modulation occurred after the button presses.

4.2.2 Ex-post conflict signaling neurons in the MSIT task

In the MSIT data, we isolated in total 455 single units from dACC (n = 296) and pre-SMA (n= 343). Neurons that signaled conflicts ex-post were selected using a Poisson GLM. In this model, spike counts in the one-second epoch after button press were entered as the response variable. The main effects and the interaction effect of the VD and SI variable (‘0’ represents no interference, ‘1’ represents with interference), as well as the main effect of RT were included. In both dACC and pre-SMA, we found that significant proportion of neurons whose spike rates correlated with a non-linear interaction between VD and SI (n = 49 in dACC, n = 45 in pre-SMA), henceforth referred to as the VD-SI multiplexing neurons. These multiplexing neurons did not represent levels of interference by changing their spike rates monotonically. Instead, there was a diversity of coding patterns. Some neurons only coded VD in the presence of SC, whereas some did so in the absence of SI (Fig. 4.7 a,c; p = 0.001 for the interaction effect). Some neurons coded SI only in the presence of VD, where as some did so in the absence of VD (Fig 4.7 b,d; p = 0.001 for the interaction effect). Apart from these multiplexing neurons, we also found some neurons that coded only VD (n = 17 in dACC and n = 29 in pre-SMA) or only SI (n = 22 in dACC, n = in 35 in pre-SMA); they spike rates showed main effects of VD or SI, but no interaction effect between the two). By contrast, few neurons have spike rates that reflect VD and SI additively (n = 2 in dACC, n= 5 in pre-SMA). A third large category of neurons are the RT-signaling neurons, whose spike rates in the one second window after button press were correlated with RT on the same trial significantly (n = 55 in dACC, n=100 in pre-SMA). Note that while levels of conflict affect RT (see Fig 4.2), the response of the conflict cells discussed in this section cannot be explained by differences in RT because the regression model takes this variable in account as a nuisance variable. To summarize the coding behaviors of all recorded neurons, we computed modulation indices for VD and SI (see methods), which provided a continuous measure of how well spike rates indicated the presence of a certain type of interference (VD or SI). The VD-SI interaction effect, which referred to the situation where whether spike rates could report the presence of a certain interference type depended on the presence of the other type, was captured by the product of modulation indices (e.g. product of modulation index of VD when SI = 0 and that of VD when SI = 1). For example, if for a certain neuron its spike rate difference between VD0 and VD1 switched sign between SI0 and SI1, namely, spike rates was greater on VD1SI0 than VD0SI0, but spike rates were greater on VD0SI1 than on VD1SI1, then the product of modulation would have a large negative value. This thus provided a way to characterize the interaction effect on a two-dimensional space. As is evident in Figure 4.8, neurons that multiplexed the two interference types fell on the diagonal and were separated from the rest of the population.

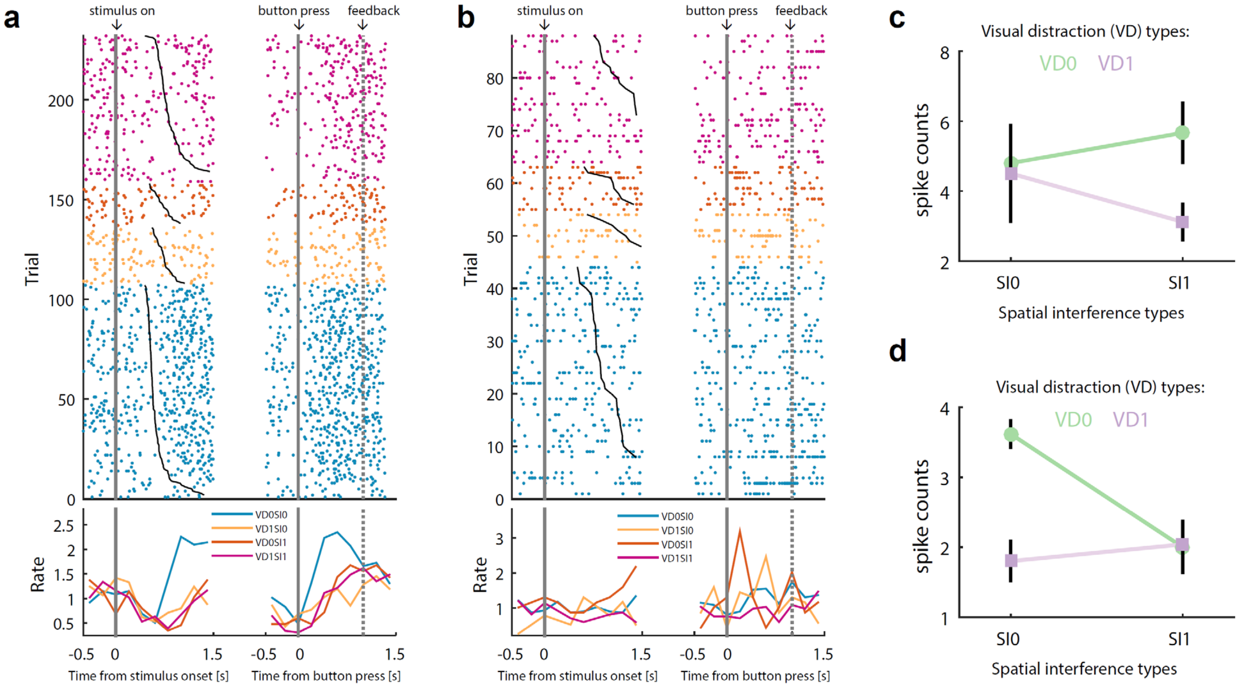

Figure 4.7 Two example neurons that signal conflicts ex post

(a) An example neuron that demonstrates the VD-SI interaction effect. Upper panel shows the raster plot (each dot represents a single spike); lower panels shows average spike rates. Data are aligned to stimulus onset (left) and to button press (right). (b) A different example neuron, displayed in the same way as (a). (c) The spike count as a function of interference types for the neuron in (a). Here, the neuron distinguished VD types but only when SI = 1. (d) Same as in c, but for neuron in (b). The neuron distinguished VD types but only when SI = 0.

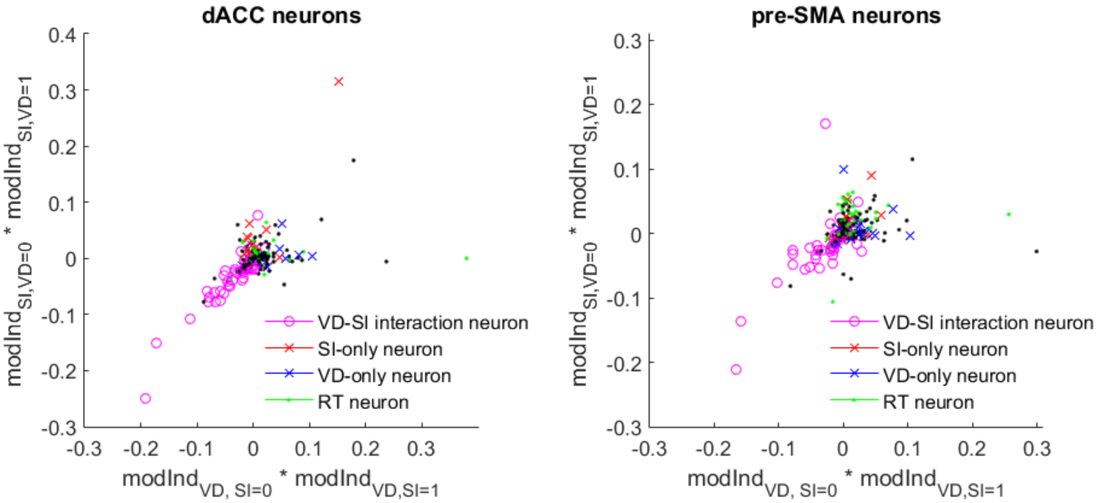

Figure 4.8 Population summary of different neuronal types

This plot shows the distribution of dACC (left) and pre-SMA (right) neurons on the 2-D product space of modulation indices. Neurons that showed significant effect (spike rates in 1s post button press epoch tested using a Poisson GLM, see Methods) were marked with colored markers. Neurons that showed significant non-linear VD-SI interaction effect (magenta) was concentrated on the diagonal.

4.2.3 Population decoding of ex-post action monitoring signals

One key question regarding the ex-post monitoring signals is whether they constitute different signals from the ex-ante monitoring signals reported in previous literature, or just its continuation in time. We investigate this problem with the population decoding approach. We chose this method for two reasons. First, there is extensive literature on how information is encoded in the population dynamics and is not revealed by analyzing single neuron activity. Performance monitoring signals could also be embedded in population activity. Second, determining whether the ex-ante and ex-post monitoring signals are different using a single neuron approach would require analyzing the overlap neurons selected in these two epochs using univariate selection method (e.g. ANOVA). However, there can be overlap between the two subsets just by chance and it is unclear to what extent the overlap can reflect functional overlap in the neural circuitries, and whether these overlap neurons play any role in population representation of monitoring signals. If the ex-ante and ex-post monitoring signals are indeed separate, the population code representing each of these signals would be substantially different despite the fact that some neurons participated in both patterns.

To test whether this is the case, we trained linear decoders using ex-post spike rates from MFC neurons to decode task variables. The task variables of interest are conflict levels on the current trial and previous trial (history effect), RT, and RT signature of cognitive control. If the ex-ante and ex-post signals were different signals, a decoder trained using ex-post spike rates would only generalize weakly across time and classify ex-ante data with significantly reduced accuracy. However, including all neurons as features would often causes the learning algorithm to overfit and thus reduces decoding performance, because many neurons do not contain information about the task variables to be decoded. Such decoders would not be able to generalize well to either testing data from the same epoch or from other epochs. To prove our point, we need to show that even the *best* the decoders constructed using data from one epoch (in this case, post-action epoch) would not generalize well to other epochs (pre-action epochs).

Recursive feature elimination (RFE) algorithm provides a principled way to select the best features for decoding performance. This method uses the weights of support vector machines (SVM) as feature importance and iteratively eliminate features with the least importance, eventually arriving at a subset of features with the highest cross-validated accuracy. Note that this selection method is multivariate in nature and takes into account the mutual information between features, and thus selects neurons excluded by the univariate selection method documented in 4.2.2. In the following RFE analyses, we used the spike rates of neurons in the post-action epoch (1s) as features to build the decoder. The decoder performance was tested on independent test data from the same epoch (1s) or with data collected in a moving window that moved across the trial from -500ms to 2000ms relative to button press in steps of 10ms. This latter analysis allowed a clear visualization of how well this ex-post decoder generalized across various time points.

We first used the RFE procedure to construct decoders that determined whether a given trial was correct congruent or correct incongruent. To control for the effect of RT on ex-post spike rates (described below), we sampled the trials so that the two conditions had comparable RTs. We focused first on the Stroop data. The maximal accuracy reached for determining stimulus congruence was 78.8% for dACC and 84% for pre-SMA. At the maximal accuracy, the dACC decoder incorporated 52 neurons with the highest rankings and the pre-SMA decoder 76 neurons with highest rankings as ranked by the RFE procedure. As is apparent from the temporal generalization plot (Fig. 4.9 middle), the decoding performance peaked only after, but not before, the execution of action (t = 0). In dACC, as more neurons that were ranked below the 52th-ranking neuron were included to build the decoder, the decoding performance considerably decayed. This suggested that the information about congruence was largely restricted to a limited set of 'best' neurons. By contrast, pre-SMA neurons all seem to carry congruence information to some extent, as this was evident from the slow decay after the peak performance was reached. Even if all pre-SMA neurons were included, the decoder still performed at more than 70% accuracy (Fig 4.9 left and middle panels). Since the RFE procedures was repeated for each fold of cross-validation (five folds in total), five lists of selected neurons were generated. To combine these lists, we computed the geometric means of rankings across the five folds for each neuron, and used this as its final ranking. We then selected the same number of neurons as determined earlier by the maximal cross-validated accuracy from this combined ranking list. A decoder was re-trained with these selected neurons and its decoding accuracy was compared against a null distribution generated by the permutation test to obtain a p-value (Fig. 4.9 right panels). The decoding performance of these selected neurons in distinguishing between correct congruent and correct incongruent trials were significantly above what was predicted by chance ($p < 0.002$, permutation tests) for both dACC and pre-SMA.

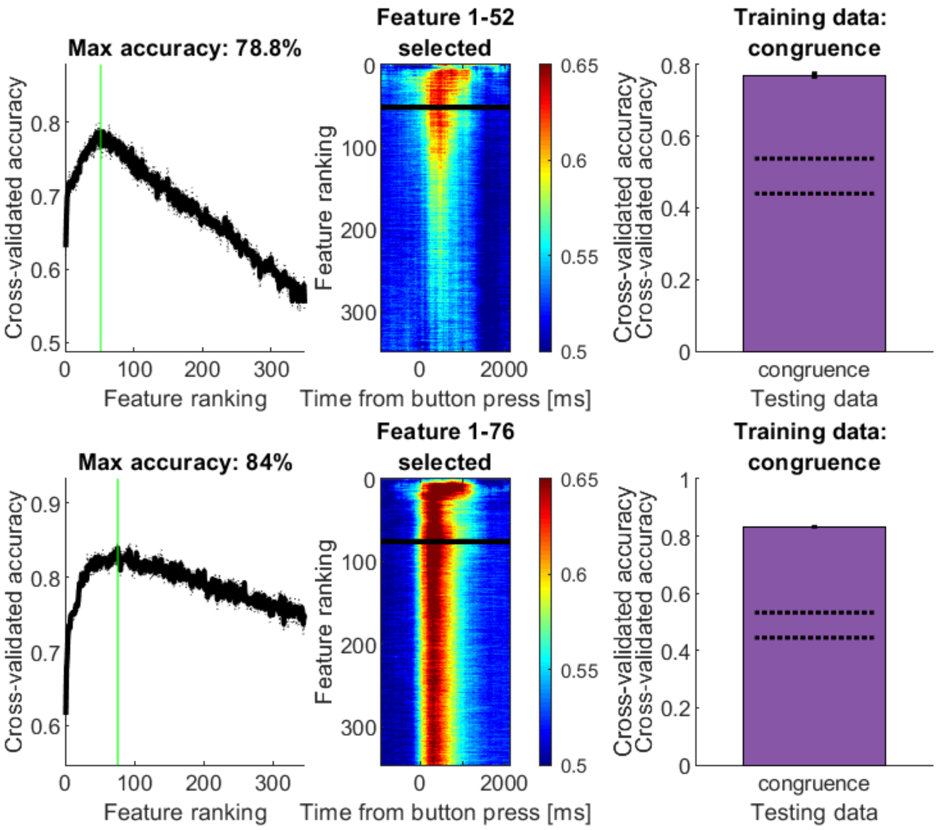

Figure 4.9 RFE-selected congruence decoders using neurons in dACC (top row) and pre-SMA (bottom row). First column of panels show the cross-validated accuracy as a function of neurons incorporated as features. The green vertical bars represent the feature ranking which, together with all features ranked above, constructed the decoder that reached the maximal cross-validated accuracy. Second column of panels show how decoders trained with data from the post-action epoch (1s) during each step of the RFE procedure generalized across time. The third columns show cross-validated accuracy for decoders trained using neurons consistently ranked best across cross-validation folds and the null distribution for decoding accuracy. The dotted lines represent the 2.5th and 97.5th percentiles of the null distribution. Error bars represent the standard error across cross-validation folds.

The neuronal population also contained significant information about reaction time (Fig. 4.10 and Fig. 4.11). To show this, we used RFE to construct decoders that determined median partitions of RTs, which was whether the RT of a particular trial was above or below the median RT (of the session). To control for the effects of previous trial and current trial stimulus congruence on current trial RT and spike rates, we analyzed cI (incongruent trials following a congruent trial) trials and cC (congruent trials following a congruent trial) trials separately. We also equalized the number of trials across cI and cC conditions. We tested the decoders within and across conditions. For example, we tested the RFE-selected decoders trained with cI data with cI hold-out data and also with cC data, and vice versa. In dACC (Fig. 4.10), the maximal accuracy reached for determining RT partitions on cI trials was 84.7% with 43 neurons selected, whereas the maximal accuracy reached for determining RT partitions on cC trials was 81.2% with 56 neurons selected. In pre-SMA (Fig. 4.11), the maximal accuracy reached for determining RT partitions on cI trials was 89% with 32 neurons selected, whereas the maximal accuracy reached for determining RT partitions on cC trials was 84.4% with 102 neurons selected. Interestingly, the decoders trained with cI data still performed above chance level when tested with cC data, and vice versa (Fig. 4.10 and 4.11 right), albeit with significantly reduced accuracy. This suggests that the RT information was largely consistent across congruence types and the population codes for RT information was similar between the two congruence types. This also likely reflected the fact that RT-congruence interaction neurons only took up small proportion of the RT-signaling neurons.

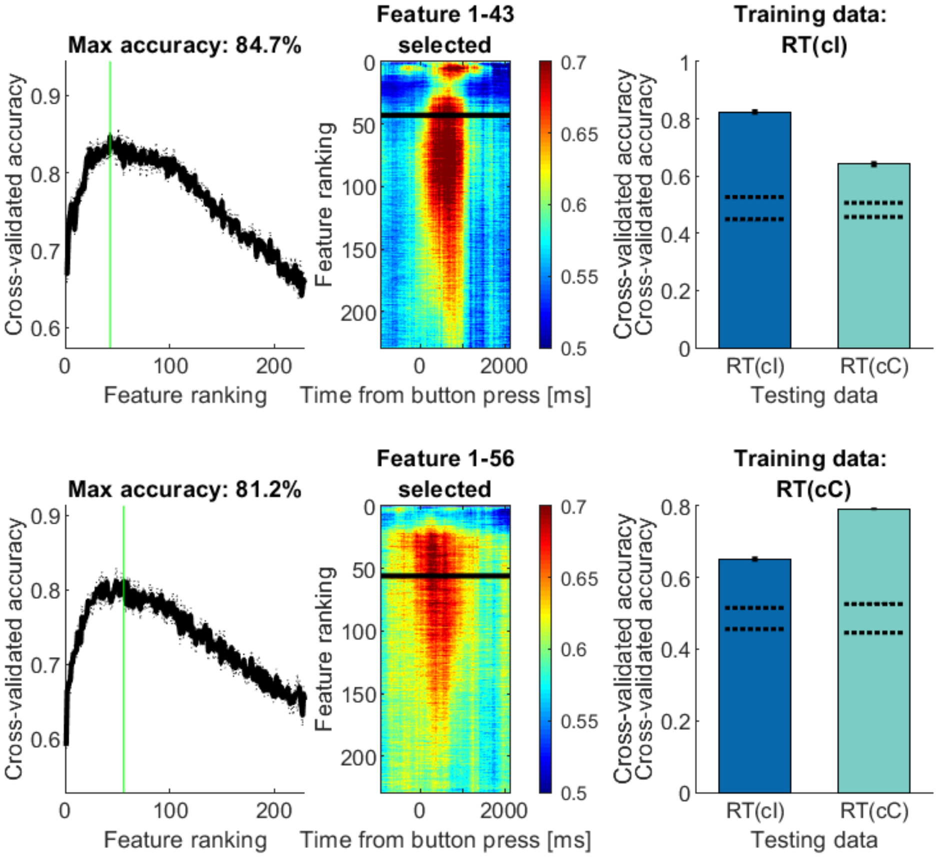

Figure 4.10 RFE-selected s decoders for RT median partition on cI trials (top row) and cC trials (bottom row), using neurons in dACC. First column of panels show the cross-validated accuracy as a function of neurons incorporated as features. The green vertical bars represent the feature ranking which, together with all features ranked above, constructed the decoder that reached the maximal cross-validated accuracy. Second column of panels show how decoders trained with data from the post-action epoch (1s) during each step of the RFE procedure generalized across time. The third columns show cross-validated accuracy for decoders trained using neurons consistently ranked best across cross-validation folds, each tested with both the cI and cC data, and the null distribution for decoding accuracy. The dotted lines represent the 2.5th and 97.5th percentiles of the null distribution. Error bars represent the standard error across cross-validation folds.

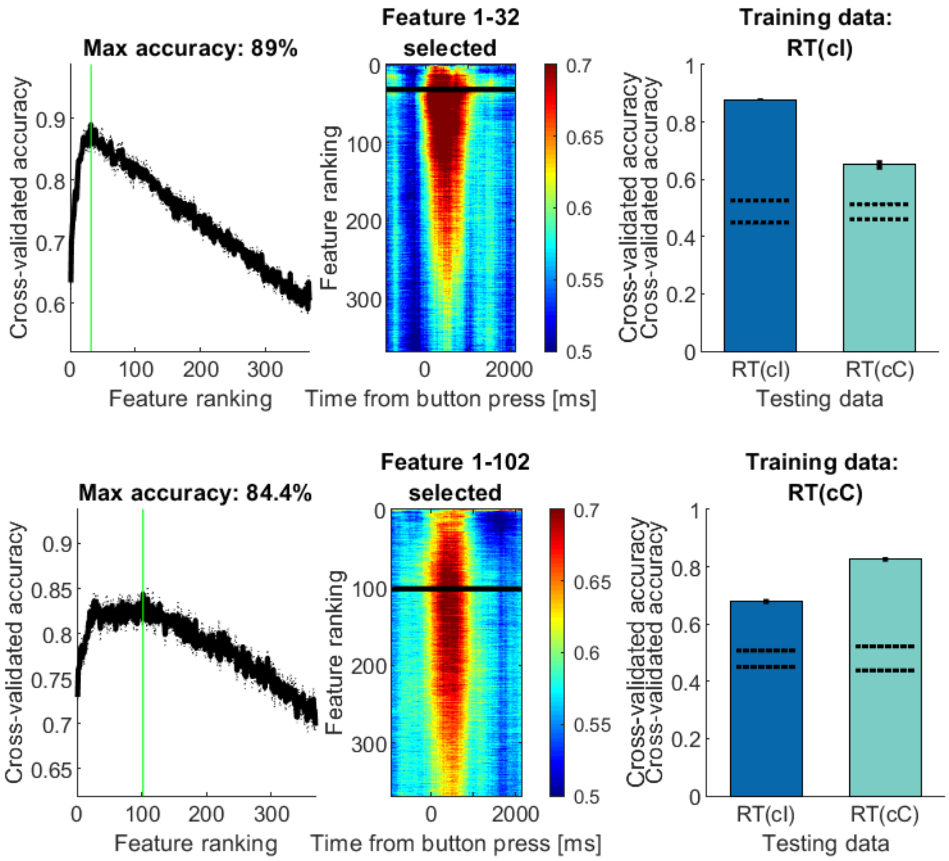

Figure 4.11 RFE-selected decoders for RT median partition on cI trials (top row) and cC trials (bottom row), using neurons in pre-SMA. First column of panels show the cross-validated accuracy as a function of neurons incorporated as features. The green vertical bars represent the feature ranking which, together with all features ranked above, constructed the decoder that reached the maximal cross-validated accuracy. Second column of panels show how decoders trained with data from the post-action epoch (1s) during each step of the RFE procedure generalized across time. The third columns show cross-validated accuracy for decoders trained using neurons consistently ranked best across cross-validation folds, each tested with both the cI and cC data, and the null distribution for decoding accuracy. The dotted lines represent the 2.5th and 97.5th percentiles of the null distribution. Error bars represent the standard error across cross-validation folds.

We also showed that the MFC retained robust information about stimulus congruence on the previous trial (Fig. 4.12). Since spike rates on the current trial were sensitive to RT (as shown above), we equalized RT on the current trial across conditions. In dACC (Fig. 4.12 top row), the maximal accuracy reached for determining stimulus congruence on the previous trial was 76.4% with 51 neurons selected as features, whereas the maximal accuracy reached in pre-SMA (Fig. 4.12 bottom row) was 74.9% with 44 neurons selected as features. It has been proposed that the MFC is responsible for learning the values of actions. The information about past stimulus congruence, which was related closely to the experienced action difficulty on the previous trial, could potentially provide useful input to computing the action values.

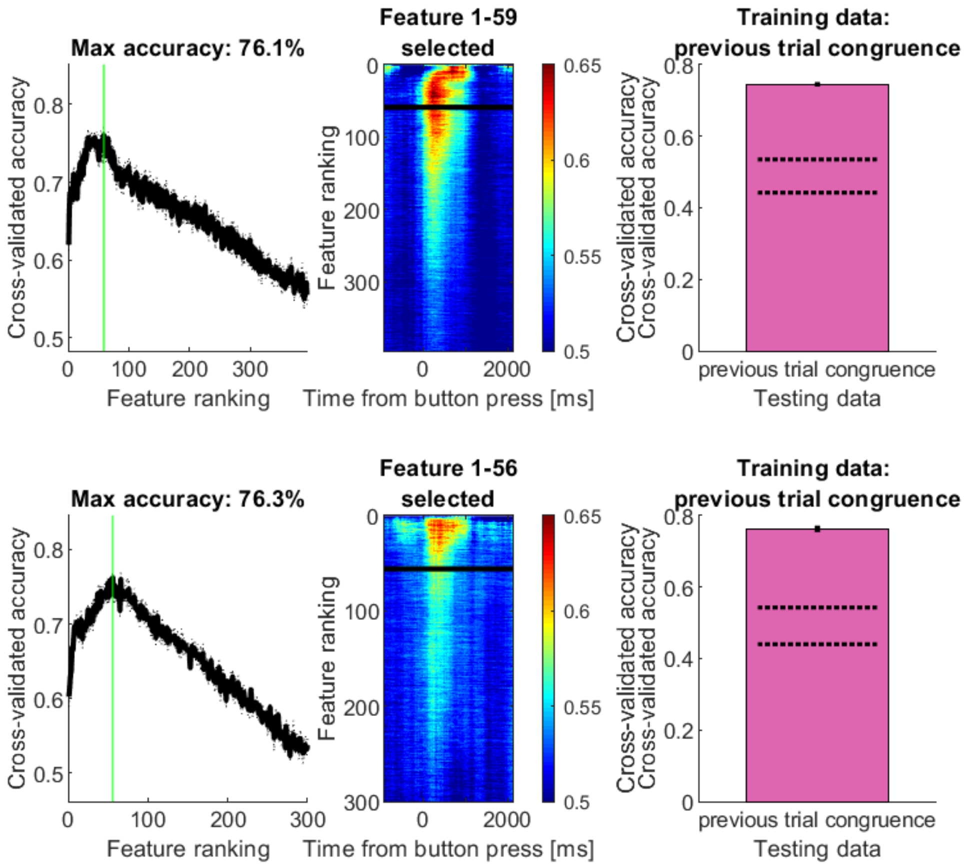

Figure 4.12 RFE-selected decoders for stimulus congruence on the previous trial, using neurons in dACC (top row) and pre-SMA (bottom row). First column of panels show the cross-validated accuracy as a function of neurons incorporated as features. The green vertical bars represent the feature ranking which, together with all features ranked above, constructed the decoder that reached the maximal cross-validated accuracy. Second column of panels show how decoders trained with data from the post-action epoch (1s) during each step of the RFE procedure generalized across time. The third columns show cross-validated accuracy for decoders trained using neurons consistently ranked best across cross-validation folds and the null distribution for decoding accuracy. The dotted lines represent the 2.5th and 97.5th percentiles of the null distribution. Error bars represent the standard error across cross-validation folds.

Given that the aforementioned ex-post performance monitoring signals provided a basis for cognitive control, we investigated whether the MFC carried signatures of control. Specifically, we investigated whether there was any signals in the population activity that could predict the levels of RT adjustment. For this analysis, we partitioned the RT differences between pairs of consecutive incongruent trials or congruent trials by their median values respectively, and constructed decoders to determine which partition a particular trial belong to. We tested the decoders across conditions as well. In dACC, the maximal accuracy reached for determining the partitions of RT difference of iI pairs was 74.1% with 15 neurons selected, whereas the maximal accuracy reached for determining the partitions of RT difference of cC pairs was 62.4% with 22 neurons selected. In pre-SMA, the maximal accuracy reached for determining the partitions of RT difference of iI pairs was 84.9% with 21 neurons selected, whereas the maximal accuracy reached for determining the partitions of RT difference of cC pairs was 30% with 22 neurons selected.

Interestingly, the decoders constructed to decode the partitions of RT difference of cC pairs generalized to data of iI pairs in both dACC and pre-SMA. We further analyzed the decoding performance for a more accurate measure of RT adjustment, which is to compute the difference between trial 2 and 4 within a trial sequence of “iccC” or “ciiI” (see Section 4.2.1). In such cases, the sample size was considerably smaller so that RT difference computed in “iccC” and “ciiI” were pooled to increase statistical power. Indeed, we were still able to select neurons that construct decoders that determined the partitions of RT difference in these trial sequences accurately (Fig. 4.15). The maximal accuracy reached for decoding this partition was 95.2% in dACC and 87.4% in pre-SMA, with 31 and 10 neurons selected, respectively. The decoding accuracy was higher than determining the partitions of iI or cC RT difference, as these RT adjustment measures were less accurate because the stimulus congruence on trials preceding the pairs were not controlled. As is evident in the time course of testing accuracy, the signal was localized after the button press and thus was an ex-post monitoring signal. In summary, both dACC and pre-SMA contained signals that were predictive of the amount of cognitive control engaged after encountering an incongruent trial at the population level.

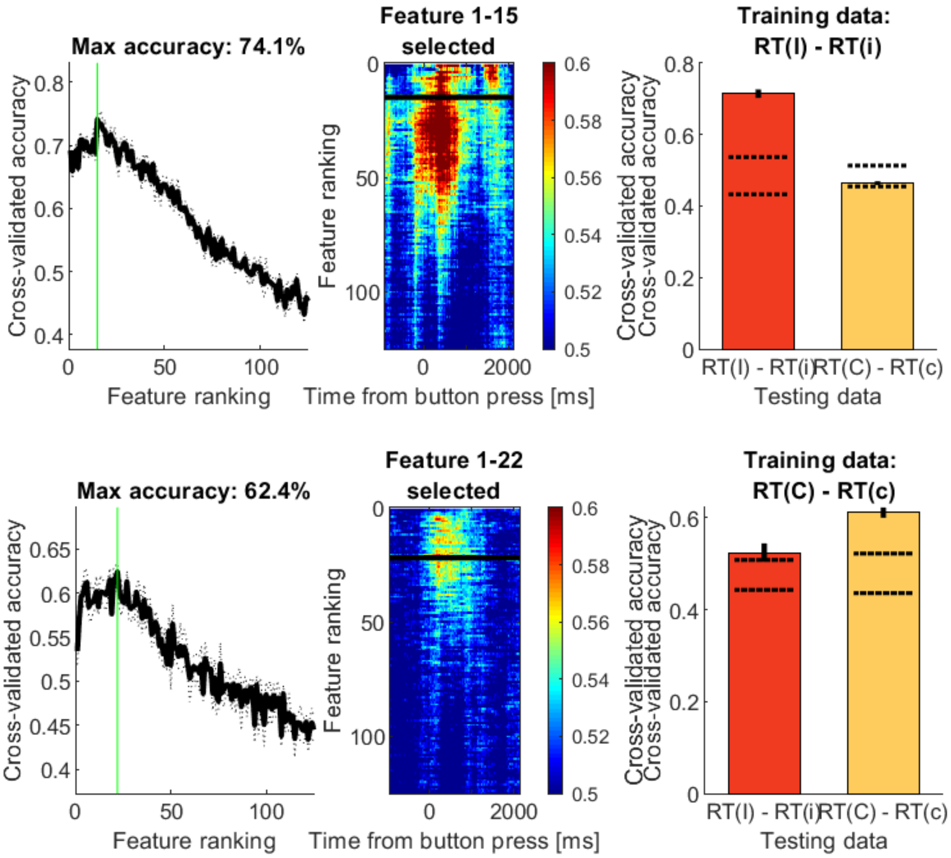

Figure 4.13 RFE-selected decoders for median partitions of RT difference of iI pairs (top row) and cC pairs (bottom row), using neurons in dACC. First column of panels show the cross-validated accuracy as a function of neurons incorporated as features. The green vertical bars represent the feature ranking which,

together with all features ranked above, constructed the decoder that reached the maximal cross-validated accuracy. Second column of panels show how decoders trained with data from the post-action epoch (1s) during each step of the RFE procedure generalized across time. The third columns show cross-validated accuracy for decoders trained using neurons consistently ranked best across cross-validation folds, each tested with both the iI and cC data, and the null distribution for decoding accuracy. The dotted lines represent the 2.5th and 97.5th percentiles of the null distribution. Error bars represent the standard error across cross-validation folds.

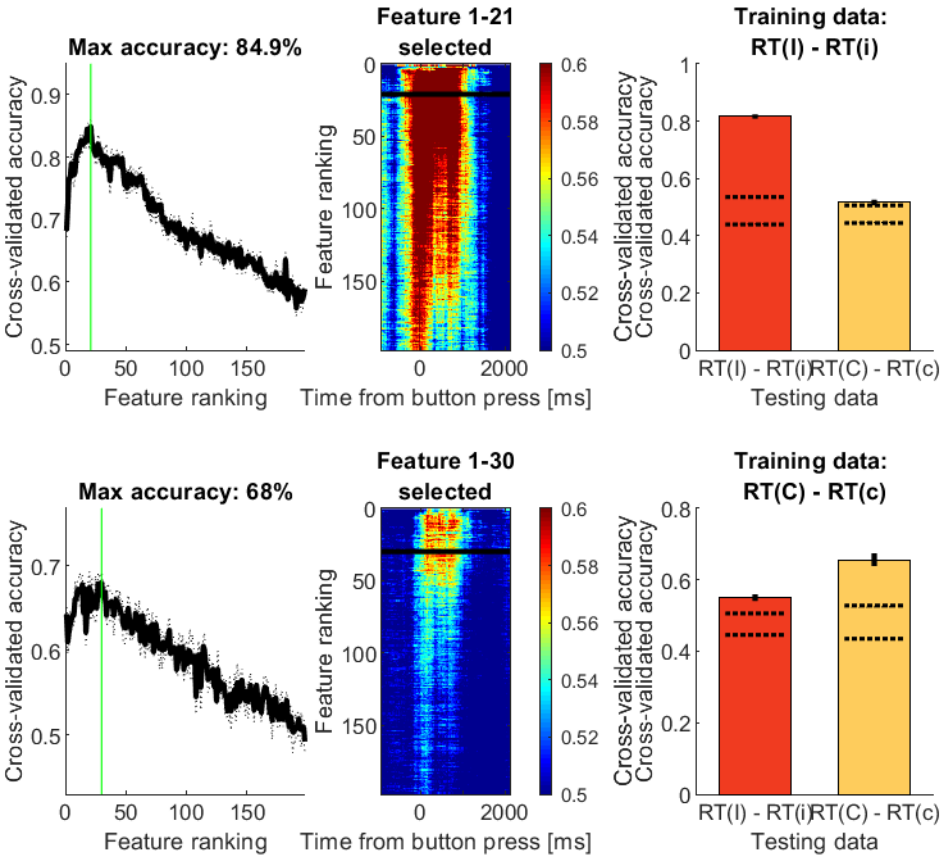

Figure 4.14 RFE-selected decoders for median partitions of RT difference of iI pairs (top row) and cC pairs (bottom row), using neurons in pre-SMA. First column of panels show the cross-validated accuracy as a function of neurons incorporated as features. The green vertical bars represent the feature ranking which, together with all features ranked above, constructed the decoder that reached the maximal cross-validated accuracy. Second column of panels show how decoders trained with data from the post-action epoch (1s) during each step of the RFE procedure generalized across time. The third columns show cross-validated accuracy for decoders trained using neurons consistently ranked best across cross-validation folds, each tested with both the iI and cC data, and the null distribution for decoding accuracy. The dotted lines represent the 2.5th and 97.5th percentiles of the null distribution. Error bars represent the standard error across cross-validation folds.

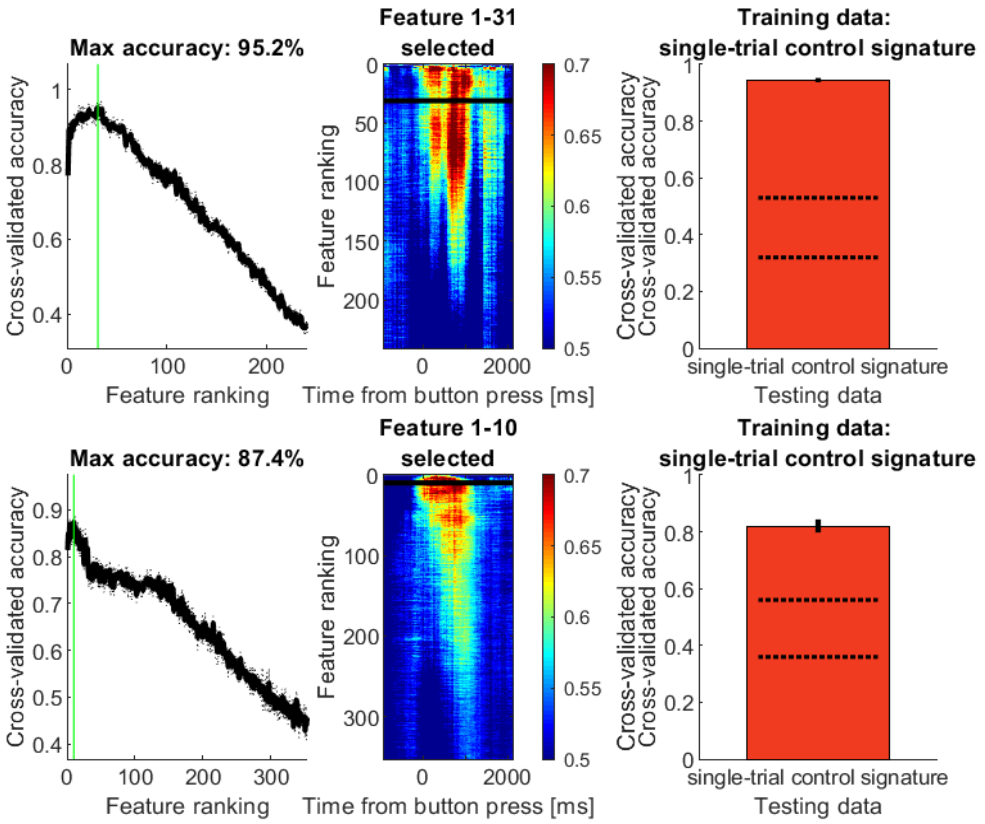

Figure 4.15 RFE-selected decoders for median partitions of RT difference of the fourth and second trials in the iccC and ciiI sequences using neurons in dACC (top row) and pre-SMA (bottom row). First column of panels show the cross-validated accuracy as a function of neurons incorporated as features. The green vertical bars represent the feature ranking which, together with all features ranked above, constructed the decoder that reached the maximal cross-validated accuracy. Second column of panels show how decoders trained with data from the post-action epoch (1s) during each step of the RFE procedure generalized across time. The third columns show cross-validated accuracy for decoders trained using neurons consistently ranked best across cross-validation folds and the null distribution for decoding accuracy. The dotted lines represent the 2.5th and 97.5th percentiles of the null distribution. Error bars represent the standard error across cross-validation folds.

4.3 Discussion

In this chapter, we focused on the performance monitoring signals on the correct trials (errors were discussed in Chapter 2). We confirmed the existence of ex-post conflict signals. Previously, studies using fMRI have reported conflict signals, but from these studies, it remains unknown when exactly these signals are present – i.e. before or after button press (Carter et al., 1998, MacDonald et al., 2000, Kerns et al., 2004). Here, we now show a group of neurons that carry signals after button press. The reported conflict signals in EEG appeared around 350ms after stimulus onset and thus was a correlate of ex-ante conflict monitoring (Yeung et al., 2004, Botvinick et al., 2004, Van Veen and Carter, 2002). Similarly, a previously published single-neuron correlate of conflicts in human dACC was analyzed as a stimulus-locked signal (Sheth et al., 2012) and was replicated in this thesis (Chapter 2). However, in the macaque literature, conflict-sensitive neurons were only found in the supplementary eye field and their timing was late: it occurred after the stop-signal reaction time, which was the inferred time of successful stopping of the prepare action (when the 'stop' and 'go' processes were modelled as a racing). Given these previous findings, our report of ex-post conflict signals identified a missing link between the human and animal literature. Using decoding, we showed that the population code for ex-ante and ex-post conflict was different and non-generalizable, suggesting that the ex-post conflict signal was indeed a separate and independent monitoring signal with a late onset (after the action had been performed). Future work is needed to investigate whether the same or different neurons signal ex-post conflict signals in the Stroop and MSIT tasks.

We found prominent ex-post RT monitoring signal. RT was considered an internal measure of experience conflict, as conflicts had strong effects on RT (Fig. 4.2b). However, RT was more than conflicts: it also contains components of many other cognitive processes, such as variability in muscle movements. We showed that on the population level, RT signals were robust and ex-post and was broadly carried by many neurons (as by the number of neurons selected using RFE). Interestingly, at the single neuron level, some neurons exhibited interaction effect between RT and stimulus congruence: their post-action spike rates correlated with RT but the correlation coefficients differed drastically between congruent and incongruent trials. One interpretation is that RT information is maintained for controlling the RT on the next trial according to specified speed-accuracy tradeoff, and since controlling RTs could involve different mechanisms depending on the conflict levels, these interaction neurons could reflect such difference. Future analyses are needed to reveal roles of these interaction neurons in specifying behavioral control and in population activity.

We additionally found that the neuronal population in both dACC and pre-SMA contained decodable information about stimulus congruence on the preceding trial. This is consistent with the hypothesis that these brain regions could make use of ex-post conflict neurons to update learned models and to specify future control (Shenhav et al., 2013). We also were able to reveal on the population level signatures of cognitive control, illustrating the central role of the MFC in bridging representations of past information with specification of future behavioral outputs.

4.4 Methods

Behavioral analyses. For the Stroop task data, we sought single-trial behavioral signature of cognitive control. We extracted trial sequences of "ciiI" and "iccC". Trials with stimulus repetitions were excluded from this analysis. We took the difference between the fourth and the second trials in the trial sequences, and test the mean of this RT difference population against zero using a one-tailed t-test, as motivated by prior literature on congruence sequence effect.

We constructed a mixed-effect linear regression to test the effects of different interference types on RT. For this model, we entered two fixed effects: visual distraction types (coded 0 if no visual distraction, coded 1 if there is visual distraction) and spatial interference types (coded 0 if no spatial interference, coded 1 if there is spatial interference). We entered reaction time (RT) as the response variable and session numbers nested within subject ID as the random effect. Additionally, we used another method to confirm the additive effects of interference types. We extracted mean RT for each combination of interference types ('VD0SI0','VD1SI0','VD0SI1','VD1SI1') for each subject and each session, and entered these data into a two-way ANOVA and tested for the interaction effect.

Univariate selection of neurons. We only considered neurons that had a mean spike rate > 0.5 Hz. In the Stroop data, we sought neurons whose spike rate differed significantly between trial types of interest in the post-action epoch (1s after button press): (i) neurons whose spike rates differed between correct congruent and correct incongruent trials (ii) neurons whose spike rates correlated with RT (iii) neurons whose spike rates demonstrated interaction effect between RT and stimulus congruence. We analyzed the spike rates of each neuron using an ANOVA model. The spike rates were entered as the response variable, and RTs and a dummy variable coding for stimulus congruence ("1" for correct incongruent and "0" for correct congruent) as predictors. We also included the interaction term for the two predictors. Neurons were selected by the p-values from the F tests conducted for each predictors and the interaction effects.

In the MSIT data, we sought neurons whose spike rate differed significantly between trial types of interest in the post-action epoch (1s after button press): (i) neurons that multiplexed two types of interference (ii) neurons that only coded for VD (iii) neurons that only coded for SI (iv) neurons that coded RT. We fit a generalized linear model (GLM) to each neuron (using matlab function "fitglm.m") and then evaluated whether the model explained significant variance to determine whether a neuron was selective or not for a variable of interest. We entered the spike count in the epoch of interest as the response variable. We entered three predictor variables: (i) a dummy variable coding for VD (ii) a dummy variable coding for SI (iii) RT.

Modulation index. We used the modulation index as a continuous measure of the effect of interference on the spike rate. The modulation index is defined as (here 'int type' stands for 'interference type'):

$$M_{int\ type_1\ ,cell_i}(int\ type_2) = \frac{Spike\ rate_{int\ type_1=1,cell_i}\ (int\ type_2) - Spike\ rate_{int\ type_1=0,cell_i\ (int\ type_2)}}{Spike\ rate_{int\ type_1=1,cell_i}\ (int\ type_2) + Spike\ rate_{int\ type_1=0,cell_i\ (int\ type_2)}} \quad (1)$$

Since there are two types of interference in the MSIT task, the modulation index for one type of interference is a function of the other type of interference, reflected in the equation. The product of modulation index is then defined as:

$$M_{product} = M_{int\ type_1\ ,cell_i}(int\ type_2 = 0) * M_{int\ type_1\ ,cell_i}(int\ type_2 = 1)$$

Decoding analyses. We used a sliding-window decoding approach to analyze the population coding of previous trial and current trial stimulus congruence, current trial RT partitions and behavioral signatures of cognitive control. We used a 500ms bin moved across the spike train on each trial in successive 25ms steps. We first equalized the number of trials across all sessions by drawing a random subset from the trial types/session that had more trials than the required number. The spike counts were extracted from each time bin across all recorded neurons from the equalized trial set and concatenated into a feature matrix (trial x neuron number). We then used this matrix as the feature matrix and trained a SVM decoder using LIBSVM (Chang and Lin, 2011). To construct a train-test generalization time series, we constructed the decoders using data from one time point (post-action epoch), and tested the decoder's performance across all the other time points over a trial to see if it generalized to these time points. For all of the decoding procedures described above, we resampled the trials 50 times so that all data collected were represented. To assess the significance of decoder performance, we permuted the trial labels 500 times. For each permutation, we ran the decoding procedure with resampling 50 times and averaged the resulting decoding accuracy for the particular label permutation run. After this, we obtained 500 mean decoding accuracy values as the empirical null distribution. A p-value was obtained by comparing the true accuracy with this empirical distribution. The procedure for Recursive Feature Elimination (RFE) can be found in (Guyon et al., 2002). In short, on each iteration, a SVM decoder was trained, and the feature with smallest value of squared weights were eliminated. The features left were used to train a new decoder. The sequence with which features were eliminated served as the ranking of the feature impact for the decoding performance. To avoid overfitting, we performed fivefold cross-validation using RFE, where features (neurons) were selected using four of the five folds, and the decoding accuracy of the selected neurons were tested with the data in the left-out fold.

References

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S. & Cohen, J. D. 2001. Conflict monitoring and cognitive control. *Psychol Rev,* 108**,** 624-52.

Botvinick, M. M., Cohen, J. D. & Carter, C. S. 2004. Conflict monitoring and anterior cingulate cortex: an update. *Trends Cogn Sci,* 8**,** 539-46.

Bush, G. & Shin, L. M. 2006. The Multi-Source Interference Task: an fMRI task that reliably activates the cingulo-frontal-parietal cognitive/attention network. *Nat Protoc,* 1**,** 308-13.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D. & Cohen, J. D. 1998. Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science,* 280**,** 747-9.

Chang, C. C. & Lin, C. J. 2011. LIBSVM: A Library for Support Vector Machines. *Acm Transactions on Intelligent Systems and Technology,* 2.

Fu, Z., Wu, D. J., Ross, I., Chung, J. M., Mamelak, A. N., Adolphs, R. & Rutishauser, U. 2019. Single-Neuron Correlates of Error Monitoring and Post-Error Adjustments in Human Medial Frontal Cortex. *Neuron,* 101**,** 165-177 e5.

Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning,* 46**,** 389-422.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., 3rd, Cho, R. Y., Stenger, V. A. & Carter, C. S. 2004. Anterior cingulate conflict monitoring and adjustments in control. *Science,* 303**,** 1023-6.

MacDonald, A. W., 3rd, Cohen, J. D., Stenger, V. A. & Carter, C. S. 2000. Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science,* 288**,** 1835-8.

Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K. & Fusi, S. 2013. The importance of mixed selectivity in complex cognitive tasks. *Nature,* 497**,** 585-90.

Shenhav, A., Botvinick, M. M. & Cohen, J. D. 2013. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron,* 79**,** 217-40.

Sheth, S. A., Mian, M. K., Patel, S. R., Asaad, W. F., Williams, Z. M., Dougherty, D. D., Bush, G. & Eskandar, E. N. 2012. Human dorsal anterior cingulate cortex neurons mediate ongoing behavioural adaptation. *Nature,* 488**,** 218-21.

Van Veen, V. & Carter, C. S. 2002. The timing of action-monitoring processes in the anterior cingulate cortex. *J Cogn Neurosci,* 14**,** 593-602.

Yeung, N., Botvinick, M. M. & Cohen, J. D. 2004. The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychol Rev,* 111**,** 931-59.

Chapter 5. Discussion and future directions

5.1 Summary

In this thesis, I have reported results from several related studies on human performance monitoring and cognitive control. I will first summarize the main findings. Chapter 2 of my thesis was concerned with error monitoring and the putative mechanisms for error-triggered cognitive control. Here I used one of the most common tasks, the Stroop task, a classic reaction-time interference task. We found neurons that signaled self-monitored errors in the absence of external feedback, in both dACC and pre-SMA, two key regions in the frontal cortex. These neurons were by no means rare, constituting 30-40% or all recorded neurons in these brain regions. Several types of errors were distinguished in terms of the information that they represented – some increased firing rates when errors were detected ("Type I"), some decreased firing rates ("Type II"), and some signaled information about whether an error had occurred on the preceding trial ("Error integrating neurons"). Although there were also neurons signaling conflicts, in general error neurons did not also signal response conflict (that is, they failed to distinguish difficult, incongruent, trials from easier, congruent, trials). This is an important result, because a key theory of cognitive control predicts that the neurons signaling errors and conflict should be the same (Yeung et al., 2004). This, we found here, is not the case. Rather, these two representations were largely separate, an insight that could not have been derived by non-invasive means. Taken together, these results characterized several common populations of neurons that carry information about errors.

Similarly, we found intracranial ERN responses (iERN) in both of these brain regions as well. These signals were recorded in the same patients, and on the same electrodes, but using low impedance contacts rather than the high impedance contacts from which we were able to isolate single units. The two error signals, error neuron spiking and iERN, were correlated in magnitude. This correlation was specific to the signaling of errors: on correct trials, the same neuron's spikes did not correlate with the amplitude of evoked potentials. These findings support a model whereby the iERN is generated locally by synaptic inputs to error neurons in both dACC and pre-SMA. Furthermore, this iERN amplitude – spike rate correlation was directly related to behavior: it predicted post-error slowing. The greater the correlation between error neuron spikes and iERN, the slower the subject was to respond on the next trial after an error. This phenomena is commonly known as post-error slowing and is a key signature of cognitive control. This finding suggests a mechanistic interpretation, according to which coherence between spikes and iERN is implementing greater cognitive control. This was a highly novel finding, and could not have been obtained with other measures like fMRI. A final further mechanistic piece of evidence was that, following an error, both error neuron spikes, and the iERN, occurred with earlier latencies in the pre-SMA than in the dACC. While considerably more work will be needed (see below), these unique findings from invasive human recordings for the first time suggest the beginning of a specific circuit, comprising pre-SMA, dACC and other regions yet to be determined, within which coordinated neural signals about errors engage control processes that influence subsequent behavioral adjustments.

In Chapter 3, I probed these mechanisms further, using spike-field coherence (SFC) analyses. These provided additional information about this neuronal synchrony. I found that neurons formed dynamic activity ensembles that communicated via the theta frequency band (5-10Hz) during baseline, but switched to the delta frequency band (2-5Hz) after an error. In pre-SMA this delta-range SFC was predictive of post-error slowing. These findings from Chapter 3 further support the interpretations from Chapter 2 that local coordination between field potentials and spikes were required to influence behavior. The additional insight provided in Chapter 3 is that there appear to be specific frequency bands that broadcast specific information. A major open question that I have not yet addressed here is SFC between different brain regions, which could eventually inform models of how information in one brain region is communicated to another.

In Chapter 4, I turned to another commonly used cognitive control task, the multi-source interference task (MSIT). This has some advantages relative to the Stroop task, in particular it provides two different sources of conflict that can cause errors. This allowed me to investigate conflicts and conflict-induced cognitive control in more detail. We found neurons in both dACC and pre-SMA that reported conflicts during the MSIT task ex post, that is after the button press. This is thus a monitoring signal. On a single cell level, these neurons showed interesting non-linear coding patterns that were manifested as an interaction effect between the two types of conflicts (visual distraction, and spatial interference). Interestingly, all four combinations of conflicts (with or without visual distraction and spatial interference) could be decoded from the neuronal population in both dACC and pre-SMA. Interestingly, I found that the latency of responses in dACC was considerably shorter than in pre-SMA, contrary to the case of error monitoring. It is interesting to note that both for error neurons on the Stroop task and for conflict neurons in the MSIT task, the latency differences between pre-SMA and dACC were substantially longer than a simple monosynaptic relay (about 50ms in the Stroop task, and over 100ms in the MSIT). This suggests once again that spikes in one brain region are not simply relayed to the next, but that substantial local processing first needs to occur before information is broadcast to other regions by a population of neurons.

Our findings are largely consistent with existing models of behavioral control but for the first time provide mechanistic detail in the human brain that had been unavailable before. One popular model (Shenhav et al., 2013) suggests that dACC engages in learning the which processes to control (control identity) and how much control is to be specified (control intensity), by updating internal models with outcome information. Our findings that iERN – error neuron synchrony is predictive of the magnitude of cognitive control in dACC provides evidence the brain region indeed contained information for about specifying the intensity of cognitive control. In addition, the timing of such synchrony (immediately after the erroneous response was made) suggests that it might reflect the updating process when dACC error neurons received inputs from other brain regions reflected in iERN. These findings suggest that neuronal coherence provides a candidate mechanism for how the computation that transforms error signals into control signals can be achieved in local neuronal populations, and then communicated to distal brain regions for actual control implementation. These findings add substantial mechanistic details to the normative model of behavioral control, which has classically been based on logic and behavioral data without insight into the neuronal implementations. The ex-post conflict signals found with the MSIT in both dACC and pre-SMA suggest that this framework can be extended also to conflict signals, which are again utilized to update an internal model for control specification.

5.2 Limitations

There are a number of limitations to the results presented here. While there are obvious strengths in the temporal and anatomical precision of recording single neurons, such recordings in humans face major restrictions imposed by the requirement to obtain them in a clinical setting. The patients all have epilepsy, raising questions about the extent to which results obtained from their brains would generalize to healthy brains. This question has been addressed in many studies, and the general consensus is that valid generalizations can be drawn, provided one takes care not to record from a seizure focus. In our analyses, we always checked whether recordings were from a region that was subsequently determined to be a seizure focus. Also, the severity and age of onset of the epilepsy, and anatomical location of the seizure focus, vary considerably between patients. Thus, by basing our analyses on data obtained from a number of patients, we further ensure that the results would not be idiosyncratic to any one of them.

Further limitations of collecting data in a clinical setting are the limited time available for recordings, and the lack of flexibility in many of the parameters. Unlike similar recordings in monkeys, we are only able to record for perhaps an hour at most in any one session, and experiments regularly are interrupted if the patient becomes sleepy or has visitors. Also unlike monkey experiments, we cannot move the electrodes, nor implant them where we would like. This latter restriction no doubt introduces some sampling biases that are important to note. Another limitation is simply the number of neurons we recorded. Multielectrode arrays will be needed to record from a greater number of neurons (see below).

Finally, it is important to note that electrophysiological recordings, like fMRI, are correlational in nature. This means two things here. First, it means we cannot conclude strong causal connections between our neuronal signals and behavior. Even though errors, and behavioral slowing, are both correlated with electrophysiology, as I showed in the preceding Chapters, these associations are not necessarily causal. They could reflect other processing that, in turn, has causal effects on both our neuronal responses and the behaviors. Direct interventions would be required to determine that the association is in fact causal, very difficult to do in humans, but achievable with microstimulation or optogenetics in animals.

A second respect in which the electrophysiological results here are correlational is more subtle. What we, as experimenters, can decode from a neuron or a population of neurons need not be what downstream neurons decode. To really determine what information a neuron signals, simply recording from it is insufficient: we would need to simultaneously record from the neuron and its projection targets, and determine the Shannon information communicated between them. This is also next to impossible in the human brain, and difficult even in the most precise animal models.

While these limitations highlight how little we truly know, they also point to some clear paths forward.

5.3 Future Directions

Findings in this thesis also open up the possibility of many exciting new directions of future research. We have found neuronal signals of error monitoring in both dACC and pre-SMA. One key question is what sets them apart. From the standpoint of energy expenditure, it would not make sense for the brain to represent an exact copy of the error signals in two brain regions; they must be doing something different. Notably, some hints of differences between the two brain regions already emerged in the findings reported here. Error signals occurred earlier in pre-SMA than in dACC by ~50ms, which is a huge latency difference considering the fact that the time it takes for synaptic transmission is just 2-3ms. Given that the iERN amplitude and latency in dACC and pre-SMA were correlated (Chapter 2), this pointed to a possible hierarchical organization between dACC and pre-SMA in recruiting control based on errors.

To test this hypothesis, one needs to find statistical evidence that such communication exists between the two regions. During my post-doc, I next plan investigate three things: 1) investigate simultaneously recorded neurons in dACC and pre-SMA to see if their spike trains have any systematic relationships, using for example, cross-correlation; 2) investigate the coherence and Granger causality between LFPs in the two brain regions across multiple frequency bands; and 3) investigate spike field coherence between spikes in one region and LFP in another region. These measures could provide crucial information about whether the two brain regions communicate and the direction and frequency channels of communication if it exists. Given the importance of dACC and pre-SMA in the literature of cognitive control and a lack of a model to make sense of the existing subtle differences, our results can potentially provide an answer or model of how these two regions coordinate in performance monitoring. It could be that if a specific communication channel between the dACC and pre-SMA is blocked, subsequent cognitive control could be affected. This putative hierarchical organization could ultimately be probed with causal manipulations in a systematic manner, although this would likely require experiments in animal models. Nonetheless, it is possible to experimentally disrupt processing through electrical stimulation, passing current through the very same electrodes from which we record. One could thus stimulate one region and record the effect of stimulation in another area, while measuring how cognitive control changes as a result. Achieving a deeper understanding of how performance monitoring and model updating works could help shed light on developing possible treatments for dysfunctions that involve these processes. An important translational goal could be targeted stimulation of the performance monitoring-cognitive control system to alleviate severe psychiatric disorders that have disorders of cognitive control, such as the obsessive-compulsive disorder, attention-deficit/hyperactivity disorder, and schizophrenia. Indeed, for some of these, deep-brain stimulation is already being considered and piloted.

Another most exciting future domain will be to link electrophysiological response properties (like our "Type I", "Error integrating" definitions) to actual cellular properties. Are these excitatory projection neurons? Inhibitory interneurons? Neurons with a specific morphological or gene expression profile that could be manipulated selectively using optogenetics in animal models? This level of cell-type specificity is of course extremely difficult to obtain in the human brain. Nonetheless, we obtained preliminary evidence from the trough-to-peak time that such a mapping might possible in our data.

While optogenetics are unlikely to be applicable to humans anytime soon, electrical microstimulation is feasible, and has been applied with some success (for instance, in the work of Josef Parvizi at Stanford (Parvizi et al., 2013)). In principle, one could examine effects on behavior resulting from focal microstimulation of the pre-SMA or dACC, although more spatially distributed manipulations may be required to produce an effect.

This also brings up the future direction of massively parallel recordings with high-density arrays. Several designs are being considered whereby a much greater number of neurons can be recorded from multielectrode arrays. Also of interest are electrodes that provide lamina-specific recordings, so that one can make inferences about which cortical layer one is recording from. All of these are still very much under development and will require a close interplay between the technology and the justification of additional risks when implanted.

Finally, there are future directions that would investigate cognitive control with truly multimodal methods. It is quite possible to carry out fMRI in the very same patients in whom one obtains single-unit recordings. Typically, the fMRI is done prior to implantation, as this is more feasible, safer, and avoids the artifacts that are produced on the MRI due to the paramagnetic nature of the electrodes. The advantage of fMRI is that it provides a whole-brain field-of-view. Thus, one could identify putative target regions, which could then be subsequently implanted (of course, based on clinical criteria). Such an approach has been used quite successfully in monkeys, for instance in studies of face processing.

I will close with two final broad future directions. One direction is to ask how cognitive control interacts with other cognitive processing. What perceptual processing needs to occur before an error can be detected? Does the patient have to consciously recognize the error they made while holding the stimulus in working memory? To what extent is attention required? How does the error detection happen as a decision process? Is there some kind of accumulation of sensory evidence? These are questions about how the operational concepts of conflict, error detection, and control interact with the rest of cognition to produce behavior. It may well be that conflicts and errors are actually not the primitives that the brain represents, but that they are assembled from other building blocks.

A related, second, question is one about engineering. Can the neurobiological study of cognitive control provide us with general principles for how control happens in complex system, and how it can be engineered? Over the past decade, our understanding of how humans flexibly control their own behaviors has undoubtedly deepened, as summarized in the Introduction of this thesis. On the other hand, we have also witnessed the great success of deep learning. With some particular features, such as nonlinearity in the input-output relationship, hierarchy and proper learning rules, deep neural nets already provide us with artificial systems that outperform humans in specialized domains of tasks. However, in a sense, current deep neural networks (DNN) are still just expert modules, not yet a full behaving system. What happens if one DNN recognizes the face of a friend in a crowd, and another DNN recognizes a bear charging from another direction? How is a decision made about how to behave with multiple inputs? The contrast between the flexible human brain and artificial neural networks that are specialized in one domain highlights one major missing feature: a control mechanism that organizes and arbitrates between the different expert systems in an intelligent way. This is currently still far from being implemented in artificially intelligent systems. The human brain has many specialized modules, such as the visual cortex, that may function in some ways like artificial neural networks: they can extract expert representations of the external world. What makes it a brain is the ability to bind these systems together and make use of these representations to achieve goals and, when different systems are in conflict with each other, to arbitrate and resolve conflicts, while keeping the control mechanisms flexible and generalizable. I am not necessarily suggesting that some kind of "supervisory system" needs to be layered on top of these expert systems, as older schemes of cognitive control sometimes proposed (Norman and Shallice, 1983). It is also possible that the multiple modules just compete with each other and achieve consensus, or winner-take-all, in a self-organizing manner. Studying the principles of how neurons communicate and how cognitive control is implemented and interacts with different expert systems at the neuronal level will help to reveal general principles that are potentially useful for creating the ultimate artificial autonomy.

Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V. & van der Sluis, S. 2014. A solution to dependency: using multilevel analysis to accommodate nested data. *Nat Neurosci,* 17**,** 491-6.

Alexander, W. H. & Brown, J. W. 2011. Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience,* 14**,** 1338-U163.

Amiez, C., Joseph, J. P. & Procyk, E. 2006. Reward encoding in the monkey anterior cingulate cortex. *Cerebral Cortex,* 16**,** 1040-1055.

Anderson, S. W., Damasio, H., Jones, R. D. & Tranel, D. 1991. Wisconsin Card Sorting Test-Performance as a Measure of Frontal-Lobe Damage. *Journal of Clinical and Experimental Neuropsychology,* 13**,** 909-922.

Aron, A. R., Behrens, T. E., Smith, S., Frank, M. J. & Poldrack, R. A. 2007. Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (MRI) and functional MRI. *J Neurosci,* 27**,** 3743-52.

Aron, A. R., Herz, D. M., Brown, P., Forstmann, B. U. & Zaghloul, K. 2016. Frontosubthalamic Circuits for Control of Action and Cognition. *J Neurosci,* 36**,** 11489-11495.

Aron, A. R. & Poldrack, R. A. 2006. Cortical and subcortical contributions to Stop signal response inhibition: role of the subthalamic nucleus. *J Neurosci,* 26**,** 2424-33.

Bartho, P., Hirase, H., Monconduit, L., Zugaro, M., Harris, K. D. & Buzsaki, G. 2004. Characterization of neocortical principal cells and interneurons by network interactions and extracellular features. *J Neurophysiol,* 92**,** 600-8.

Behrens, T. E., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. 2007. Learning the value of information in an uncertain world. *Nat Neurosci,* 10**,** 1214-21.

Behrens, T. E. J., Fox, P., Laird, A. & Smith, S. M. 2013. What is the most interesting part of the brain? *Trends in Cognitive Sciences,* 17**,** 2-4.

Bokil, H., Andrews, P., Kulkarni, J. E., Mehta, S. & Mitra, P. P. 2010. Chronux: A platform for analyzing neural signals. *Journal of Neuroscience Methods,* 192**,** 146-151.

Bonini, F., Burle, B., Liegeois-Chauvel, C., Regis, J., Chauvel, P. & Vidal, F. 2014. Action monitoring and medial frontal cortex: leading role of supplementary motor area. *Science,* 343**,** 888-91.

Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S. & Cohen, J. D. 1999. Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature,* 402**,** 179-81.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S. & Cohen, J. D. 2001. Conflict monitoring and cognitive control. *Psychol Rev,* 108**,** 624-52.

Botvinick, M. M., Cohen, J. D. & Carter, C. S. 2004. Conflict monitoring and anterior cingulate cortex: an update. *Trends Cogn Sci,* 8**,** 539-46.

Brainard, D. H. 1997. The Psychophysics Toolbox. *Spatial Vision,* 10**,** 433-436.

Brazdil, M., Roman, R., Daniel, P. & Rektor, I. 2005. Intracerebral error-related negativity in a simple Go/NoGo task. *Journal of Psychophysiology,* 19**,** 244-255.

Brown, J. W. & Braver, T. S. 2005. Learned predictions of error likelihood in the anterior cingulate cortex. *Science,* 307**,** 1118-1121.

Burle, B., Roger, C., Allain, S., Vidal, F. & Hasbroucq, T. 2008. Error negativity does not reflect conflict: a reappraisal of conflict monitoring and anterior cingulate cortex activity. *J Cogn Neurosci,* 20**,** 1637-55.

Bush, G. & Shin, L. M. 2006. The Multi-Source Interference Task: an fMRI task that reliably activates the cingulo-frontal-parietal cognitive/attention network. *Nat Protoc,* 1**,** 308-13.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D. & Cohen, J. D. 1998. Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science,* 280**,** 747-9.

Chang, C. C. & Lin, C. J. 2011. LIBSVM: A Library for Support Vector Machines. *Acm Transactions on Intelligent Systems and Technology,* 2.

Cohen, J. D., Dunbar, K. & McClelland, J. L. 1990. On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychol Rev,* 97**,** 332-61.

Danielmeier, C., Eichele, T., Forstmann, B. U., Tittgemeyer, M. & Ullsperger, M. 2011. Posterior medial frontal cortex activity predicts post-error adaptations in task-related visual and motor areas. *J Neurosci,* 31**,** 1780-9.

Debener, S., Ullsperger, M., Siegel, M., Fiehler, K., von Cramon, D. Y. & Engel, A. K. 2005. Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. *J Neurosci,* 25**,** 11730-7.

Dehaene, S., Posner, M. I. & Tucker, D. M. 1994. Localization of a Neural System for Error-Detection and Compensation. *Psychological Science,* 5**,** 303-305.

Dutilh, G., van Ravenzwaaij, D., Nieuwenhuis, S., van der Maas, H. L. J., Forstmann, B. U. & Wagenmakers, E. J. 2012. How to measure post-error slowing: A confound and a simple solution. *Journal of Mathematical Psychology,* 56**,** 208-216.

Ebitz, R. B. & Platt, M. L. 2015. Neuronal activity in primate dorsal anterior cingulate cortex signals task conflict and predicts adjustments in pupil-linked arousal. *Neuron,* 85**,** 628-40.

Ekstrom, A. 2010. How and when the fMRI BOLD signal relates to underlying neural activity: The danger in dissociation. *Brain Research Reviews,* 62**,** 233-244.

Ekstrom, A., Viskontas, I., Kahana, M., Jacobs, J., Upchurch, K., Bookheimer, S. & Fried, I. 2007. Contrasting roles of neural firing rate and local field potentials in human memory. *Hippocampus,* 17**,** 606-17.

Emeric, E. E., Brown, J. W., Leslie, M., Pouget, P., Stuphorn, V. & Schall, J. D. 2008. Performance monitoring local field potentials in the medial frontal cortex of primates: anterior cingulate cortex. *J Neurophysiol,* 99**,** 759-72.

Emeric, E. E., Leslie, M., Pouget, P. & Schall, J. D. 2010. Performance monitoring local field potentials in the medial frontal cortex of primates: supplementary eye field. *J Neurophysiol,* 104**,** 1523-37.

Engel, A. K., Moll, C. K., Fried, I. & Ojemann, G. A. 2005. Invasive recordings from the human brain: clinical insights and beyond. *Nat Rev Neurosci,* 6**,** 35-47.

Eriksen, B. A. & Eriksen, C. W. 1974. Effects of Noise Letters Upon Identification of a Target Letter in a Nonsearch Task. *Perception & Psychophysics,* 16**,** 143-149.

Falkenstein, M., Hohnsbein, J., Hoormann, J. & Blanke, L. 1991. Effects of Crossmodal Divided Attention on Late Erp Components .2. Error Processing in Choice Reaction Tasks. *Electroencephalography and Clinical Neurophysiology,* 78**,** 447-455.

Fellows, L. K. & Farah, M. J. 2005. Is anterior cingulate cortex necessary for cognitive control? *Brain,* 128**,** 788-796.

Frank, M. J., Woroch, B. S. & Curran, T. 2005. Error-related negativity predicts reinforcement learning and conflict biases. *Neuron,* 47**,** 495-501.

Fried, I., Rutishauser, U., Cerf, M. & Kreiman, G. 2014. *Single neuron studies of the human brain : probing cognition,* Cambridge, Massachusetts, The MIT Press.

Friederici, A. D. 2017. *Language in our brain : the origins of a uniquely human capacity,* Cambridge, Massachusetts, The MIT Press.

Fries, P. 2005. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in Cognitive Sciences,* 9**,** 474-480.

Fries, P., Reynolds, J. H., Rorie, A. E. & Desimone, R. 2001. Modulation of oscillatory neuronal synchronization by selective visual attention. *Science,* 291**,** 1560-3.

Fu, Z., Wu, D. J., Ross, I., Chung, J. M., Mamelak, A. N., Adolphs, R. & Rutishauser, U. 2019. Single-Neuron Correlates of Error Monitoring and Post-Error Adjustments in Human Medial Frontal Cortex. *Neuron,* 101**,** 165-177 e5.

Fujisawa, S. & Buzsaki, G. 2011. A 4 Hz oscillation adaptively synchronizes prefrontal, VTA, and hippocampal activities. *Neuron,* 72**,** 153-65.

Fuster, J. M. 2015. *The prefrontal cortex,* Amsterdam ; Boston, Elsevier/AP, Academic Press is an imprint of Elsevier.

Gehring, W. J. & Fencsik, D. E. 2001. Functions of the medial frontal cortex in the processing of conflict and errors. *J Neurosci,* 21**,** 9430-7.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E. & Donchin, E. 1993. A Neural System for Error-Detection and Compensation. *Psychological Science,* 4**,** 385-390.

Gehring, W. J., Liu, Y., Orr, J. M. & Carp, J. 2013. The error-related negativity (ERN/Ne). *In:* Luck, S. J. & Kappenman, E. S. (eds.) *The Oxford handbook of event-related potential components.* Oxford: Oxford University Press.

Gehring, W. J. & Willoughby, A. R. 2002. The medial frontal cortex and the rapid processing of monetary gains and losses. *Science,* 295**,** 2279-82.

Gerbrandt, L. K., Lawrence, J. C., Eckardt, M. J. & Lloyd, R. L. 1978. Origin of the neocortically monitored theta rhythm in the curarized rat. *Electroencephalogr Clin Neurophysiol,* 45**,** 454-67.

Glascher, J., Adolphs, R., Damasio, H., Bechara, A., Rudrauf, D., Calamia, M., Paul, L. K. & Tranel, D. 2012. Lesion mapping of cognitive control and value-based decision making in the prefrontal cortex. *Proc Natl Acad Sci U S A,* 109**,** 14681-6.

Godlove, D. C., Emeric, E. E., Segovis, C. M., Young, M. S., Schall, J. D. & Woodman, G. F. 2011. Event-related potentials elicited by errors during the stop-signal task. I. Macaque monkeys. *J Neurosci,* 31**,** 15640-9.

Gold, J. M., Carpenter, C., Randolph, C., Goldberg, T. E. & Weinberger, D. R. 1997. Auditory working memory and Wisconsin Card Sorting Test performance in schizophrenia. *Archives of General Psychiatry,* 54**,** 159-165.

Gratton, G., Coles, M. G. H. & Donchin, E. 1992. Optimizing the Use of Information - Strategic Control of Activation of Responses. *Journal of Experimental Psychology-General,* 121**,** 480-506.

Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning,* 46**,** 389-422.

Hajcak, G., McDonald, N. & Simons, R. F. 2003. To err is autonomic: Error-related brain potentials, ANS activity, and post-error compensatory behavior. *Psychophysiology,* 40**,** 895-903.

Hanes, D. P., Thompson, K. G. & Schall, J. D. 1995. Relationship of presaccadic activity in frontal eye field and supplementary eye field to saccade initiation in macaque: Poisson spike train analysis. *Exp Brain Res,* 103**,** 85-96.

Hare, T. A., Schultz, W., Camerer, C. F., O'Doherty, J. P. & Rangel, A. 2011. Transformation of stimulus value signals into motor commands during simple choice. *Proceedings of the National Academy of Sciences of the United States of America,* 108**,** 18120-18125.

Harris, K. D., Henze, D. A., Csicsvari, J., Hirase, H. & Buzsaki, G. 2000. Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. *Journal of Neurophysiology,* 84**,** 401-414.

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M. & Platt, M. L. 2011. Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *J Neurosci,* 31**,** 4178-87.

Herrmann, M. J., Rommler, J., Ehlis, A. C., Heidrich, A. & Fallgatter, A. J. 2004. Source localization (LORETA) of the error-related-negativity (ERN/Ne) and positivity (Pe). *Brain Res Cogn Brain Res,* 20**,** 294-9.

Holroyd, C. B. & Coles, M. G. H. 2002. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychol Rev,* 109**,** 679-709.

Isoda, M. & Hikosaka, O. 2007. Switching from automatic to controlled action by monkey medial frontal cortex. *Nat Neurosci,* 10**,** 240-8.

Ito, S., Stuphorn, V., Brown, J. W. & Schall, J. D. 2003. Performance monitoring by the anterior cingulate cortex during saccade countermanding. *Science,* 302**,** 120-2.

Jahanshahi, M., Obeso, I., Rothwell, J. C. & Obeso, J. A. 2015. A fronto-striato-subthalamic-pallidal network for goal-directed and habitual inhibition. *Nat Rev Neurosci,* 16**,** 719-32.

Kayser, C., Kim, M., Ugurbil, K., Kim, D. S. & Konig, P. 2004. A comparison of hemodynamic and neural responses in cat visual cortex using complex stimuli. *Cerebral Cortex,* 14**,** 881-891.

Kennerley, S. W., Walton, M. E., Behrens, T. E., Buckley, M. J. & Rushworth, M. F. 2006. Optimal decision making and the anterior cingulate cortex. *Nat Neurosci,* 9**,** 940-7.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., 3rd, Cho, R. Y., Stenger, V. A. & Carter, C. S. 2004. Anterior cingulate conflict monitoring and adjustments in control. *Science,* 303**,** 1023-6.

King, J. A., Korb, F. M., von Cramon, D. Y. & Ullsperger, M. 2010. Post-error behavioral adjustments are facilitated by activation and suppression of task-relevant and task-irrelevant information processing. *J Neurosci,* 30**,** 12759-69.

Koechlin, E. & Hyafil, A. 2007. Anterior prefrontal function and the limits of human decision-making. *Science,* 318**,** 594-8.

Kolling, N., Wittmann, M. K., Behrens, T. E., Boorman, E. D., Mars, R. B. & Rushworth, M. F. 2016. Value, search, persistence and model updating in anterior cingulate cortex. *Nat Neurosci,* 19**,** 1280-5.

Kouneiher, F., Charron, S. & Koechlin, E. 2009. Motivation and cognitive control in the human prefrontal cortex. *Nat Neurosci,* 12**,** 939-45.

Kreiman, G., Hung, C. P., Kraskov, A., Quiroga, R. Q., Poggio, T. & DiCarlo, J. J. 2006. Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. *Neuron,* 49**,** 433-445.

Laming, D. 1979. Choice Reaction Performance Following an Error. *Acta Psychologica,* 43**,** 199-224.

Logothetis, N. K., Kayser, C. & Oeltermann, A. 2007. In vivo measurement of cortical impedance spectrum in monkeys: implications for signal propagation. *Neuron,* 55**,** 809-23.

Logothetis, N. K. & Wandell, B. A. 2004. Interpreting the BOLD signal. *Annu Rev Physiol,* 66**,** 735-69.

Luck, S. J. 2014. A closer look at ERPs and ERP components. *In:* Luck, S. J. (ed.) *An introduction to the event-related potential technique.* 2 ed. Cambridge, Massachusetts: The MIT Press.

Luu, P., Tucker, D. M. & Makeig, S. 2004. Frontal midline theta and the error-related negativity: neurophysiological mechanisms of action regulation. *Clin Neurophysiol,* 115**,** 1821-35.

MacDonald, A. W., 3rd, Cohen, J. D., Stenger, V. A. & Carter, C. S. 2000. Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science,* 288**,** 1835-8.

MacLean, E. L., Hare, B., Nunn, C. L., Addessi, E., Amici, F., Anderson, R. C., Aureli, F., Baker, J. M., Bania, A. E., Barnard, A. M., Boogert, N. J., Brannon, E. M., Bray, E. E., Bray, J., Brent, L. J. N., Burkart, J. M., Call, J., Cantlon, J. F., Cheke, L. G., Clayton, N. S., Delgado, M. M., DiVincenti, L. J., Fujita, K., Herrmann, E., Hiramatsu, C., Jacobs, L. F., Jordan, K. E., Laude, J. R., Leimgruber, K. L., Messer, E. J. E., Moura, A. C. D., Ostojic, L., Picard, A., Platt, M. L., Plotnik, J. M., Range, F., Reader, S. M., Reddy, R. B., Sandel, A. A., Santos, L. R., Schumann, K., Seed, A. M., Sewall, K. B., Shaw, R. C., Slocombe, K. E., Su, Y. J., Takimoto, A., Tan, J. Z., Tao, R., van Schaik, C. P., Viranyi, Z., Visalberghi, E., Wade, J. C., Watanabe, A., Widness, J., Young, J. K., Zentall, T. R. & Zhao, Y. N. 2014. The evolution of self-control. *Proceedings of the National Academy of Sciences of the United States of America,* 111**,** E2140-E2148.

MacLeod, C. M. 1991. Half a century of research on the Stroop effect: an integrative review. *Psychol Bull,* 109**,** 163-203.

Mansouri, F. A., Buckley, M. J. & Tanaka, K. 2007. Mnemonic function of the dorsolateral prefrontal cortex in conflict-induced behavioral adjustment. *Science,* 318**,** 987-90.

Mansouri, F. A., Koechlin, E., Rosa, M. G. P. & Buckley, M. J. 2017. Managing competing goals - a key role for the frontopolar cortex. *Nat Rev Neurosci,* 18**,** 645-657.

Maris, E. & Oostenveld, R. 2007. Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods,* 164**,** 177-90.

Matsumoto, K., Suzuki, W. & Tanaka, K. 2003. Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science,* 301**,** 229-32.

Matsumoto, M., Matsumoto, K., Abe, H. & Tanaka, K. 2007. Medial prefrontal cell activity signaling prediction errors of action values. *Nat Neurosci,* 10**,** 647-56.

Metcalfe, J. 2008. Evolution of metacognition. *Handbook of metamemory and memory.* New York, NY, US: Psychology Press.

Miltner, W. H., Braun, C. H. & Coles, M. G. 1997. Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a "generic" neural system for error detection. *J Cogn Neurosci,* 9**,** 788-98.

Minxha, J., Mosher, C., Morrow, J. K., Mamelak, A. N., Adolphs, R., Gothard, K. M. & Rutishauser, U. 2017. Fixations Gate Species-Specific Responses to Free Viewing of Faces in the Human and Macaque Amygdala. *Cell Rep,* 18**,** 878-891.

Mitchell, J. F., Sundberg, K. A. & Reynolds, J. H. 2007. Differential attention-dependent response modulation across cell classes in macaque visual area V4. *Neuron,* 55**,** 131-41.

Mitra, P. P. & Pesaran, B. 1999. Analysis of dynamic brain imaging data. *Biophysical Journal,* 76**,** 691-708.

Nachev, P., Kennard, C. & Husain, M. 2008. Functional role of the supplementary and pre-supplementary motor areas. *Nat Rev Neurosci,* 9**,** 856-69.

Nakamura, K., Roesch, M. R. & Olson, C. R. 2005. Neuronal activity in macaque SEF and ACC during performance of tasks involving conflict. *J Neurophysiol,* 93**,** 884-908.

Narayanan, N. S., Cavanagh, J. F., Frank, M. J. & Laubach, M. 2013. Common medial frontal mechanisms of adaptive control in humans and rodents. *Nat Neurosci,* 16**,** 1888-1895.

Niessing, J., Ebisch, B., Schmidt, K. E., Niessing, M., Singer, W. & Galuske, R. A. 2005. Hemodynamic signals correlate tightly with synchronized gamma oscillations. *Science,* 309**,** 948-51.

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P. & Kok, A. 2001. Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. *Psychophysiology,* 38**,** 752-60.

Nir, Y., Fisch, L., Mukamel, R., Gelbard-Sagiv, H., Arieli, A., Fried, I. & Malach, R. 2007. Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to interneuronal correlations. *Current Biology,* 17**,** 1275-1285.

Norman, D. A. & Shallice, T. 1983. Attention to Action - Willed and Automatic-Control of Behavior. *Bulletin of the Psychonomic Society,* 21**,** 354-354.

Notebaert, W., Houtman, F., Opstal, F. V., Gevers, W., Fias, W. & Verguts, T. 2009. Post-error slowing: an orienting account. *Cognition,* 111**,** 275-9.

Olvet, D. M. & Hajcak, G. 2008. The error-related negativity (ERN) and psychopathology: toward an endophenotype. *Clin Psychol Rev,* 28**,** 1343-54.

Olvet, D. M. & Hajcak, G. 2009. The stability of error-related brain activity with increasing trials. *Psychophysiology,* 46**,** 957-61.

Parvizi, J., Rangarajan, V., Shirer, W. R., Desai, N. & Greicius, M. D. 2013. The will to persevere induced by electrical stimulation of the human cingulate gyrus. *Neuron,* 80**,** 1359-67.

Passingham, R. E. & Wise, S. P. 2012. *The neurobiology of the prefrontal cortex : anatomy, evolution, and the origin of insight,* Oxford, United Kingdom, Oxford University Press.

Pesaran, B., Vinck, M., Einevoll, G. T., Sirota, A., Fries, P., Siegel, M., Truccolo, W., Schroeder, C. E. & Srinivasan, R. 2018. Investigating large-scale brain dynamics using field potential recordings: analysis and interpretation. *Nature Neuroscience,* 21**,** 903-919.

Potts, G. F., Martin, L. E., Kamp, S. M. & Donchin, E. 2011. Neural response to action and reward prediction errors: Comparing the error-related negativity to behavioral errors and the feedback-related negativity to reward prediction violations. *Psychophysiology,* 48**,** 218-228.

Pouzat, C., Mazor, O. & Laurent, G. 2002. Using noise signature to optimize spike-sorting and to assess neuronal classification quality. *Journal of Neuroscience Methods,* 122**,** 43-57.

Purcell, B. A. & Kiani, R. 2016. Neural Mechanisms of Post-error Adjustments of Decision Policy in Parietal Cortex. *Neuron,* 89**,** 658-71.

Quilodran, R., Rothe, M. & Procyk, E. 2008. Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron,* 57**,** 314-325.

Rabbitt, P. M. A. 1966. Error Correction Time without External Error Signals. *Nature,* 212**,** 438-&.

Rangel, A. & Hare, T. 2010. Neural computations associated with goal-directed choice. *Curr Opin Neurobiol,* 20**,** 262-70.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A. & Nieuwenhuis, S. 2004. The role of the medial frontal cortex in cognitive control. *Science,* 306**,** 443-7.

Riehle, A., Grun, S., Diesmann, M. & Aertsen, A. 1997. Spike synchronization and rate modulation differentially involved in motor cortical function. *Science,* 278**,** 1950-3.

Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K. & Fusi, S. 2013. The importance of mixed selectivity in complex cognitive tasks. *Nature,* 497**,** 585-90.

Rushworth, M. F. & Behrens, T. E. 2008. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nat Neurosci,* 11**,** 389-97.

Rutishauser, U., Ross, I. B., Mamelak, A. N. & Schuman, E. M. 2010. Human memory strength is predicted by theta-frequency phase-locking of single neurons. *Nature,* 464**,** 903-7.

Rutishauser, U., Schuman, E. M. & Mamelak, A. N. 2006. Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. *J Neurosci Methods,* 154**,** 204-24.

Rutishauser, U., Ye, S. X., Koroma, M., Tudusciuc, O., Ross, I. B., Chung, J. M. & Mamelak, A. N. 2015. Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nature Neuroscience,* 18**,** 1041-+.

Salinas, E. & Sejnowski, T. J. 2001. Correlated neuronal activity and the flow of neural information. *Nat Rev Neurosci,* 2**,** 539-50.

Scangos, K. W., Aronberg, R. & Stuphorn, V. 2013. Performance monitoring by presupplementary and supplementary motor area during an arm movement countermanding task. *Journal of Neurophysiology,* 109**,** 1928-1939.

Schmitzer-Torbert, N., Jackson, J., Henze, D., Harris, K. & Redish, A. D. 2005. Quantitative measures of cluster quality for use in extracellular recordings. *Neuroscience,* 131**,** 1-11.

Schneider, W. & Shiffrin, R. M. 1977. Controlled and Automatic Human Information-Processing .1. Detection, Search, and Attention. *Psychological Review,* 84**,** 1-66.

Shenhav, A., Botvinick, M. M. & Cohen, J. D. 2013. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron,* 79**,** 217-40.

Sheth, S. A., Mian, M. K., Patel, S. R., Asaad, W. F., Williams, Z. M., Dougherty, D. D., Bush, G. & Eskandar, E. N. 2012. Human dorsal anterior cingulate cortex neurons mediate ongoing behavioural adaptation. *Nature,* 488**,** 218-21.

Shiffrin, R. M. & Schneider, W. 1977. Controlled and Automatic Human Information-Processing .2. Perceptual Learning, Automatic Attending, and a General Theory. *Psychological Review,* 84**,** 127-190.

Shima, K. & Tanji, J. 1998. Role for cingulate motor area cells in voluntary movement selection based on reward. *Science,* 282**,** 1335-8.

Siegel, M., Donner, T. H. & Engel, A. K. 2012. Spectral fingerprints of large-scale neuronal interactions. *Nat Rev Neurosci,* 13**,** 121-34.

Sirota, A., Montgomery, S., Fujisawa, S., Isomura, Y., Zugaro, M. & Buzsaki, G. 2008. Entrainment of neocortical neurons and gamma oscillations by the hippocampal theta rhythm. *Neuron,* 60**,** 683-97.

Smith, J. D., Shields, W. E. & Washburn, D. A. 2003. The comparative psychology of uncertainty monitoring and metacognition. *Behav Brain Sci,* 26**,** 317-39; discussion 340-73.

Stroop, J. R. 1935. *Studies of interference in serial verbal reactions.* Ph D, George Peabody College for Teachers.

Stuphorn, V. & Schall, J. D. 2006. Executive control of countermanding saccades by the supplementary eye field. *Nat Neurosci,* 9**,** 925-31.

Stuphorn, V., Taylor, T. L. & Schall, J. D. 2000. Performance monitoring by the supplementary eye field. *Nature,* 408**,** 857-60.

Stuss, D. T., Floden, D., Alexander, M. P., Levine, B. & Katz, D. 2001. Stroop performance in focal lesion patients: dissociation of processes and frontal lobe lesion location. *Neuropsychologia,* 39**,** 771-86.

Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D. & Leahy, R. M. 2011. Brainstorm: a user-friendly application for MEG/EEG analysis. *Comput Intell Neurosci,* 2011**,** 879716.

Tang, H., Yu, H. Y., Chou, C. C., Crone, N. E., Madsen, J. R., Anderson, W. S. & Kreiman, G. 2016. Cascade of neural processing orchestrates cognitive control in human frontal cortex. *Elife,* 5.

Trujillo, L. T. & Allen, J. J. 2007. Theta EEG dynamics of the error-related negativity. *Clin Neurophysiol,* 118**,** 645-68.

Tsujimoto, S., Genovesio, A. & Wise, S. P. 2010. Evaluating self-generated decisions in frontal pole cortex of monkeys. *Nat Neurosci,* 13**,** 120-6.

Ullsperger, M. & Danielmeier, C. 2016. Reducing Speed and Sight: How Adaptive Is Post-Error Slowing? *Neuron,* 89**,** 430-2.

Ullsperger, M., Danielmeier, C. & Jocham, G. 2014. Neurophysiology of performance monitoring and adaptive behavior. *Physiol Rev,* 94**,** 35-79.

Ullsperger, M. & von Cramon, D. Y. 2001. Subprocesses of performance monitoring: a dissociation of error processing and response competition revealed by event-related fMRI and ERPs. *Neuroimage,* 14**,** 1387-401.

Van Veen, V. & Carter, C. S. 2002. The timing of action-monitoring processes in the anterior cingulate cortex. *J Cogn Neurosci,* 14**,** 593-602.

Vendrell, P., Junque, C., Pujol, J., Jurado, M. A., Molet, J. & Grafman, J. 1995. The role of prefrontal regions in the Stroop task. *Neuropsychologia,* 33**,** 341-52.

Vogt, B. A., Berger, G. R. & Derbyshire, S. W. 2003. Structural and functional dichotomy of human midcingulate cortex. *Eur J Neurosci,* 18**,** 3134-44.

Voytek, B., Kayser, A. S., Badre, D., Fegen, D., Chang, E. F., Crone, N. E., Parvizi, J., Knight, R. T. & D'Esposito, M. 2015. Oscillatory dynamics coordinating human frontal networks in support of goal maintenance. *Nat Neurosci,* 18**,** 1318-24.

Wang, C. M., Ulbert, I., Schomer, D. L., Marinkovic, K. & Halgren, E. 2005. Responses of human anterior cingulate cortex microdomains to error detection, conflict monitoring, stimulus-response mapping, familiarity, and orienting. *Journal of Neuroscience,* 25**,** 604-613.

Williams, Z. M., Bush, G., Rauch, S. L., Cosgrove, G. R. & Eskandar, E. N. 2004. Human anterior cingulate neurons and the integration of monetary reward with motor responses. *Nat Neurosci,* 7**,** 1370-5.

Winter, B. 2013. Linear models and linear mixed effects models in R with linguistic applications. *arXiv:1308.5499*.

Wong, Y. T., Fabiszak, M. M., Novikov, Y., Daw, N. D. & Pesaran, B. 2016. Coherent neuronal ensembles are rapidly recruited when making a look-reach decision. *Nat Neurosci,* 19**,** 327-34.

Yeung, N., Bogacz, R., Holroyd, C. B., Nieuwenhuis, S. & Cohen, J. D. 2007. Theta phase resetting and the error-related negativity. *Psychophysiology,* 44**,** 39-49.

Yeung, N., Botvinick, M. M. & Cohen, J. D. 2004. The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychol Rev,* 111**,** 931-59.

Yeung, N. & Summerfield, C. 2012. Metacognition in human decision-making: confidence and error monitoring. *Philos Trans R Soc Lond B Biol Sci,* 367**,** 1310-21.

Zanos, T. P., Mineault, P. J. & Pack, C. C. 2011. Removal of spurious correlations between spikes and local field potentials. *J Neurophysiol,* 105**,** 474-86.

Zhou, X., Qi, X. L. & Constantinidis, C. 2016. Distinct Roles of the Prefrontal and Posterior Parietal Cortices in Response Inhibition. *Cell Rep,* 14**,** 2765-73.